Unión Internacional de Telecomunicaciones

UIT-R

Sector de Radiocomunicaciones de la UIT

Recomendación UIT-R BS.1534-2

(06/2014)

Método para la evaluación subjetiva del nivel de calidad intermedia de los sistemas de audio

Serie BS

Servicio de radiodifusión (sonora)

150 ITU 1865-2015

ITU Unión Internacional de Telecomunicaciones

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 el Anexo 1 a 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 http://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	Radio astronomía
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, 2015

© UIT 2015

Reservados todos los derechos. Ninguna parte de esta publicación puede reproducirse por ningún procedimiento sin previa autorización escrita por parte de la UIT.

RECOMENDACIÓN UIT-R BS.1534-2

Método para la evaluación subjetiva del nivel de calidad intermedia de los sistemas de audio

(Cuestión UIT-R 62/6)

(2001-2003-2014)

Cometido

Esta Recomendación describe un método para la evaluación subjetiva de la calidad de audio intermedia. Este método refleja múltiples aspectos de la Recomendación UIT-R BS.1116 y utiliza la misma escala de apreciación utilizada para la evaluación de la calidad de la imagen (es decir, la de la Recomendación UIT-R BT.500).

El método denominado «Ensayo multiestímulo con referencia y patrón ocultos (MUSHRA, MUlti Stimulus test with Hidden Reference Anchor)», se ha ensayado satisfactoriamente. Las pruebas demostraron que el método MUSHRA sirve para la evaluación de la calidad de audio intermedia y arroja resultados precisos y fiables.

Palabras clave

Calidad de audio, calidad de audio intermedia, codificación de audio, efecto perturbador, evaluación subjetiva, prueba de escucha

La Asamblea de Radiocomunicaciones de la UIT,

considerando

a) que en las Recomendaciones UIT-R BS.1116, UIT-R BS.1284, UIT-R BT.500, UIT-R BT.710 y UIT-R BT.811, así como en las Recomendaciones UIT-T P.800, UIT-T P.810 y UIT-T P.830 se han establecido métodos para evaluar la calidad subjetiva de los sistemas de audio, de vídeo y de conversación;

b) que los nuevos tipos de servicios de distribución, tales como los de audio en serie continua por Internet o reproductores de estado sólido, los servicios digitales por satélite, los sistemas de onda corta y media digitales o las aplicaciones móviles multimedio pueden funcionar con una calidad de audio intermedia;

c) que la Recomendación UIT-R BS.1116 sirve para la evaluación de pequeñas degradaciones y no es adecuada para evaluar sistemas con calidad de audio intermedia;

d) que la Recomendación UIT-R BS.1284 no da una valoración absoluta en la evaluación de la calidad de audio intermedia;

e) que la introducción de patrones adecuados y pertinentes en las pruebas permite una utilización estable de la escala de apreciación subjetiva;

f) que las Recomendaciones UIT-T P.800, UIT-T P.810 y UIT-T P.830 se centran en las señales vocales en un entorno telefónico y no resultan suficientes para la evaluación de las señales de audio en un entorno de radiodifusión;

g) que la utilización de métodos subjetivos de ensayo normalizados es importante para el intercambio, la compatibilidad y la evaluación correcta de los datos de prueba;

h) que los nuevos servicios multimedio pueden requerir la evaluación combinada de la calidad de audio y de vídeo;

i) que con frecuencia se utiliza indebidamente el acrónimo MUSHRA para designar pruebas que no utilizan referencias ni patrones;

j) que los patrones pueden afectar a los resultados de la prueba y es conveniente que los patrones se parezcan a los efectos perturbadores de los sistemas que se prueban,

recomienda

1 que deben utilizarse los procedimientos de prueba y evaluación que figuran en el Anexo 1 de esta Recomendación para la evaluación subjetiva de la calidad de audio intermedia,

recomienda además

1 que se prosigan los estudios sobre los patrones que posean las características de degradación que se encuentran en los sistemas de audio más modernos, y que se actualice la presente Recomendación para incluir nuevos patrones, según proceda.

Anexo 1

1 Introducción

Esta Recomendación describe un método para la evaluación subjetiva de la calidad de audio intermedia. Este método refleja múltiples aspectos de la Recomendación UIT-R BS.1116 y utiliza la misma escala de apreciación utilizada para la evaluación de la calidad de la imagen (es decir, la de la Recomendación UIT-R BT.500).

El método denominado «Ensayo multiestímulo con referencia y patrón ocultos (MUSHRA, *MUlti Stimulus test with Hidden Reference Anchor*)», se ha ensayado satisfactoriamente. Las pruebas demostraron que el método MUSHRA sirve para la evaluación de la calidad de audio intermedia y arroja resultados precisos y fiables [2; 4; 3].

Esta Recomendación incluye los siguientes puntos y adjuntos:

Punto 1: Introducción

Punto 2: Alcance, justificación de las pruebas y objetivos del nuevo método

Punto 3: Diseño experimental

Punto 4: Selección de evaluadores

Punto 5: Método de prueba

Punto 6: Atributos

Punto 7: Material de prueba

Punto 8: Condiciones de audición

Punto 9: Análisis estadístico

Punto 10: Informe de prueba y presentación de resultados

Adjunto 1 (normativo): Instrucciones que han de darse a los evaluadores

Adjunto 2 (informativo): Orientaciones sobre el diseño de la interfaz de usuario

Adjunto 3 (normativo): Descripción de la comparación estadística no paramétrica entre dos muestras utilizando técnicas de remuestreo y métodos de simulación Monte-Carlo

Adjunto 4 (informativo): Orientaciones para el análisis estadístico paramétrico

Adjunto 5 (informativo): Requisitos para el comportamiento de patrón óptimo

2 Alcance, justificación de las pruebas y objetivos del nuevo método

Se sabe que las pruebas de audición subjetiva siguen siendo la forma más fiable de medir la calidad de los sistemas de audio. Hay métodos descritos con detalle y de eficacia probada para evaluar la calidad del audio en la parte superior e inferior de la gama de calidades.

La Recomendación UIT-R BS.1116 – Métodos para la evaluación subjetiva de pequeñas degradaciones en los sistemas de audio incluyendo los sistemas de sonido multicanal, se utiliza para la evaluación de los sistemas de audio de gran calidad con pequeñas degradaciones. No obstante, hay aplicaciones en las que es aceptable o inevitable una calidad de audio inferior. La rápida evolución en la utilización de Internet para la distribución y difusión de material de audio, en la que la velocidad de datos está limitada, han llevado a un compromiso en la calidad del audio. Otras aplicaciones que pueden tener una calidad de audio intermedia son las de modulación de amplitud digital (por ejemplo, la Digital Radio Mondiale (DRM)), la radiodifusión digital por satélite, los circuitos de comentarios en la radio y la televisión, los servicios de audio por demanda y los servicios de audio en líneas de marcación. El método de prueba definido en la Recomendación UIT-R BS.1116 no es totalmente adecuado para la evaluación de estos sistemas con calidad de audio inferior [4] porque no llega a discriminar bien entre pequeñas diferencias de calidad en la parte inferior de la escala.

La Recomendación UIT-R BS.1284 ofrece únicamente métodos especializados para la gama de calidad de audio elevada o no ofrece una valoración absoluta de la calidad de audio.

Otras Recomendaciones, como las Recomendaciones UIT-T P.800, UIT-T P.810 o UIT-T P.830 se centran en la evaluación subjetiva de las señales vocales en un entorno telefónico. El Grupo de Proyecto B/AIM de la Unión Europa de Radiodifusión (UER) ha efectuado experimentos con material típico de audio como el que se utiliza en el entorno de la radiodifusión, valiéndose de estos métodos del UIT-T. Ninguno de dichos métodos satisface los requisitos de escala absoluta, comparación con una señal de referencia y pequeños intervalos de confianza con un número razonable de evaluadores al mismo tiempo. Por tanto, la evaluación de las señales de audio en un entorno de radiodifusión no puede efectuarse adecuadamente utilizando uno de estos métodos.

El método revisado que se describe en esta Recomendación trata de dar una medida fiable y repetible de los sistemas cuya calidad de audio encajaría normalmente en la mitad inferior de la escala de degradaciones utilizada por la Recomendación UIT-R BS.1116 [2; 4; 3]. En el método de prueba MUSHRA se utiliza una señal de referencia de gran calidad y se prevé que los sistemas sometidos a ensayo introduzcan degradaciones significativas. MUSHRA está previsto para evaluar los sistemas de audio de calidad intermedia. Si se utiliza MUSHRA con el contenido adecuado, en condiciones ideales la calificación del oyente oscilará entre 20 y 80 puntos MUSHRA. Si la calificación de la mayoría de las condiciones de prueba oscila entre 80 y 100, es posible que los resultados de la prueba no sean válidos.

La horquilla de calificación puede reducirse, entre otros, por los siguientes motivos: inexperiencia de los evaluadores, utilización de contenido no crítico o prueba inadecuada para los algoritmos de codificación que se prueban.

3 Diseño experimental

En un dominio de interés científico se utilizan varias clases de estrategias de investigación para recopilar información fiable. Para la evaluación subjetiva de las degradaciones de los sistemas de audio, deben utilizarse los métodos experimentales más formales. La experimentación subjetiva se caracteriza en primer lugar por el control y la manipulación reales de las condiciones experimentales, y en segundo lugar por la recopilación y el análisis de los datos estadísticos procedentes de los oyentes. Es necesario efectuar un diseño experimental y una planificación minuciosos que aseguren que los factores incontrolados que puedan causar ambigüedades en los resultados de la prueba de audición, sean minimizados. Por ejemplo, si la secuencia real de elementos de audio fuese idéntica para todos los evaluadores en una prueba de audición, no se podría estar seguro de si las evaluaciones efectuadas por los evaluadores se deben a dicha secuencia más que a los distintos niveles de degradaciones presentadas. En consecuencia, las condiciones de prueba deben disponerse de forma que pongan de manifiesto los efectos de los factores independientes, y únicamente los de dichos factores.

En situaciones en las que cabe esperar que las degradaciones potenciales y otras características se distribuyan homogéneamente a lo largo de la prueba de audición, puede aplicarse una aleatorización verdadera a la presentación de las condiciones del ensayo. Cuando se prevea la falta de homogeneidad, debe tenerse ésta en cuenta en la presentación de las condiciones del ensayo. Por ejemplo, cuando el material a evaluar varíe en nivel o dificultad, el orden de presentación de los estímulos debe distribuirse aleatoriamente a lo largo de una sesión y entre sesiones.

Es necesario concebir las pruebas de audición de forma que los evaluadores no estén sobrecargados hasta el punto de disminuir la precisión o la evaluación. Exceptuando los casos en que la relación entre el sonido y la imagen sea importante, es preferible que la evaluación de los sistemas de audio se efectúe sin imágenes asociadas. Una consideración importante es la inclusión de condiciones de control adecuadas. Generalmente, las condiciones de control incluyen la presentación de materiales de audio sin degradaciones, que se introducen en formas impredecibles para los evaluadores. Son las diferencias entre la apreciación de estos estímulos de control y los potencialmente degradados las que permiten concluir que las valoraciones son evaluaciones reales de las degradaciones.

Algunas de estas consideraciones se describirán a continuación. Debe entenderse que los temas de diseño y ejecución experimentales y de análisis estadístico son complejos y no se pueden ofrecer todos los detalles en un documento como la presente Recomendación. Se recomienda consultar a profesionales con conocimientos del diseño experimental y estadísticos o incorporar a dichos profesionales al iniciarse la planificación de las pruebas de audición.

Para que los laboratorios puedan analizar eficazmente y transferirse unos a otros los datos, será necesario dar a conocer el diseño experimental. Se ha de definir detalladamente las variables dependientes e independientes. El número de variables independientes se definirá con sus niveles asociados.

4 Selección de los evaluadores

Los datos de las pruebas de audición en las que se evalúen pequeñas degradaciones de los sistemas de audio, tales como las de la Recomendación UIT-R BS.1116, deben obtenerse de evaluadores que tengan experiencia en la detección de estas pequeñas degradaciones. Cuanto mayor sea la calidad de los sistemas a ensayar, más importante será contar con oyentes experimentados.

4.1 Criterios para la selección de los evaluadores

Aunque el método de prueba MUSHRA no está concebido para evaluar pequeñas degradaciones, sigue siendo recomendable su empleo por oyentes experimentados para garantizar la validez de los datos de prueba obtenidos. Estos oyentes deben contar con experiencia en la audición del sonido en condiciones críticas. Dichos participantes ofrecerán resultados más fiables y de forma más rápida que los no experimentados. También es importante señalar que la mayoría de los oyentes no experimentados tienden a ser más sensibles a los diversos tipos de efectos perturbadores tras una exposición frecuente a ellos. Un evaluador experimentado se escoge por su capacidad para realizar una prueba de escucha. Esta capacidad puede cualificarse y cuantificarse en términos de la fiabilidad y discriminación del evaluador en la prueba. A partir de una serie de evaluaciones, como se define a continuación:

– **Discriminación**: Medida de la capacidad para percibir diferencias entre elementos de prueba.

– **Fiabilidad**: Medida del acercamiento de percepciones idénticas del mismo elemento de prueba.

Sólo los evaluadores considerados *evaluadores experimentados* para cualquiera de las pruebas deberán incluirse en el análisis final de los datos. Existen varias técnicas para realizar este análisis de los evaluadores. Puede encontrarse más información al respecto en el Informe UIT-R BS.2300[1]. Estas se basan en, al menos, dos percepciones idénticas por cada evaluador, lo que permite cualificar y cuantificar la experiencia de un evaluador con un experimento. Los métodos se han de aplicar como parte del proceso de preselección de los evaluadores en un experimento piloto o, preferiblemente, como proceso de preselección y como parte de la prueba principal. Se asocia un experimento piloto a una serie de experimentos y comprende un conjunto representativo de las muestras de prueba que se han de evaluar en el experimento principal. A fin de evaluar la experiencia del oyente, el experimento piloto deberá comprender un subconjunto pertinente de estímulos de prueba, representativos de toda la gama de los estímulos y efectos perturbadores que se evaluarán durante el/los experimento(s) principal(es).

La representación gráfica del análisis debe dar información sobre la fiabilidad de los evaluadores con respecto a su discriminación.

4.1.1 Preselección de los evaluadores

El panel de oyentes debe estar compuesto de participantes experimentados, o dicho de otra manera, personas que entiendan y hayan sido adecuadamente adiestradas en el método descrito de evaluación subjetiva de la calidad. Estos oyentes deben:

– contar con experiencia de la audición del sonido en condiciones críticas;

– contar con una audición normal (debe utilizarse como orientación la Norma 389 de la ISO).

Debe utilizarse el procedimiento de adiestramiento como instrumento de la preselección. Sólo los evaluadores considerados *evaluadores experimentados* en el experimento piloto o el experimento principal deberán incluirse en el análisis final de los datos.

Se utiliza la repetición de estímulos como método para evaluar la fiabilidad de los oyentes.

La razón principal para introducir una técnica de preselección es la de aumentar la eficacia del ensayo de escucha. No obstante, esta técnica debe sopesarse respecto al riesgo de limitar demasiado la relevancia del resultado.

---

[1] El método indicador de experiencia (eGauge), descrito en el Informe UIT-R BS.2300-0, es un ejemplo de la aplicación de esa técnica. Puede encontrarse en http://www.itu.int/oth/R0A07000036.

4.1.2 Postselección de los evaluadores

El método de postselección excluye a los evaluadores que asignan una calificación muy alta a una señal patrón notablemente degradada, y a los que frecuentemente califican la referencia oculta, aunque esté notablemente degradada, de acuerdo con la siguiente escala:

– se debe excluir al evaluador de las respuestas agregadas si califica la referencia oculta para > 15% de los elementos por debajo de una calificación de 90;

– se debe excluir al evaluador de las respuestas agregadas si califica el patrón de gama media para más del 15% de los elementos de prueba por encima de una calificación de 90. Si más del 25% de los evaluadores otorga al patrón de media gama una calificación superior a 90, es posible que el elemento no haya sido suficientemente degradado por el procesamiento patrón. En tal caso, no se excluirá a los evaluadores en función de la calificación de ese elemento.

Esta fase inicial puede realizarse antes de que todos los evaluadores hayan completado la prueba, si es necesario (lo que permitirá al laboratorio de pruebas evaluar si dispone de un número suficiente de evaluadores fiables antes de que se hayan completado las pruebas).

Puede resultar conveniente estudiar los datos para identificar la existencia de datos atípicos erróneos a fin de someterlos a un análisis más detallado. Un método conveniente es la comparación de las calificaciones individuales con el rango intercuartil de todas las calificaciones otorgadas a un elemento de prueba concreto, *j*, y una secuencia de audio, *k*.

La mediana, $\hat{x}$, y los cuartiles, *Q*, se calculan de la siguiente manera:

$$\hat{x} \coloneqq Q_2(x_{jk}) = \text{median}(x) \coloneqq \begin{cases} x_{jk\frac{n+1}{2}}, & n \text{ impar} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1}\right), & n \text{ par} \end{cases}$$

, *x* se ordena por tamaño creciente y

$$Q_1(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ impar} \\ \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n}{2}}\right), & n \text{ par} \end{cases},$$

$$Q_3(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ impar} \\ \text{median}\left(x_{jk\frac{n}{2}+1}, \ldots, x_{jkn}\right), & n \text{ par} \end{cases}$$

El rango intercuartil se calcula como $IQR(x) \coloneqq Q_3(x) - Q_1(x)$.

En este contexto, los valores atípicos pertenecen a *O* (x_{jk}):

$$O(x_{jk}) \coloneqq \{x_{jk} | x_{jk} > Q_3(x_{jk}) + 1{,}5 \cdot IQR(x_{jk})\} \cup \{x_{jk} | x_{jk} < Q_1(x_{jk}) - 1{,}5 \cdot IQR(x_{jk})\}$$

Si un evaluador otorga la calificación x a un estímulo concreto y el sistema que se prueba forma parte de $O(x)$, se ha de examinar la razón que motiva esa calificación. El examen de la grabación de la sesión de prueba puede revelar problemas técnicos de los equipos o errores humanos. Se puede cuestionar al evaluador para saber si la calificación verdaderamente representa su opinión subjetiva. Si se demuestra que la existencia del dato atípico se debe a un error, puede eliminarse del conjunto de datos antes de realizar el análisis final y se anotará en el informe de la prueba el motivo de su eliminación.

La aplicación de un método de postselección puede esclarecer las tendencias en el resultado de una prueba. No obstante, teniendo presente la variabilidad de las sensibilidades de los evaluadores a los distintos efectos perturbadores, debe actuarse con cautela. Aumentando el tamaño del grupo de oyentes, se reducirán los efectos de las valoraciones de todo evaluador individual.

4.2 Tamaño del grupo de participantes

El tamaño adecuado de un grupo de participantes puede determinarse si se puede estimar la variación de las valoraciones de los distintos evaluadores y se conoce la resolución requerida del experimento.

Cuando las condiciones de una prueba de audición se controlan estrictamente en los aspectos técnicos y de comportamiento, la experiencia ha demostrado que los datos procedentes de no más de 20 evaluadores suelen ser suficientes para extraer conclusiones adecuadas de la prueba. Si pueden efectuarse análisis a medida que avanza el ensayo, no es necesario procesar las valoraciones de nuevos evaluadores, cuando puede alcanzarse un nivel adecuado de significación estadística para extraer conclusiones adecuadas de la prueba.

Si por cualquier motivo no puede lograrse un control experimental estricto, puede ser necesario un número mayor de evaluadores para alcanzar la resolución exigida.

El tamaño del grupo de oyentes no es únicamente función de la resolución deseada. El resultado del tipo de experimento del que se ocupa esta Recomendación es, en principio, válido únicamente para dicho grupo preciso de oyentes experimentados que participan realmente en la prueba. Así pues, aumentando el tamaño del grupo de oyentes puede pretenderse que el resultado es válido para un grupo más general de oyentes experimentados y puede por tanto considerarse en ocasiones más convincente. Puede también ser necesario aumentar el tamaño del grupo de participantes para prever la probabilidad de que los evaluadores varíen sus sensibilidades ante los distintos efectos perturbadores.

5 Método de prueba

El método de prueba MUSHRA utiliza los materiales de programa originales no procesados, con toda su anchura de banda como señal de referencia (y se utilizan también como referencia oculta), así como un número de patrones ocultos obligatorios.

Se pueden utilizar otros patrones ocultos, de preferencia los que son conformes con otras Recomendaciones UIT-R pertinentes. Dado que las propiedades de los patrones tienen un efecto significativo en los resultados de una prueba, el diseño de un patrón no normalizado deberá tener en cuenta el comportamiento de patrón óptimo del Adjunto 5. En el informe de prueba deberán describirse detalladamente las características de todo patrón no normalizado que se utilice en la prueba.

5.1 Descripción de las señales de prueba

Se recomienda que la longitud máxima de las secuencias sea de aproximadamente 10 s y que, preferentemente, no supere los 12 s a fin de evitar la fatiga de los oyentes, aumentar la firmeza y estabilidad de las respuestas de los oyentes y reducir la duración total de la prueba de audición. Esta duración también es necesaria para establecer una coherencia de contenido a lo largo de toda la duración de la señal para aumentar la coherencia de las respuestas de los oyentes. Además, una duración más breve también permitirá a los oyentes comparar una mayor proporción continua de señales de prueba.

Si las señales son demasiado largas, las respuestas de los oyentes estarán motivadas por los efectos de primicia y postrimería de las señales de prueba o regiones en bucle aisladas, cuyas características espectrales y temporales pueden variar mucho a lo largo de la duración de la señal de prueba. Para reducir esta variabilidad se acorta la duración de las señales de prueba. Sin embargo, en ciertos casos esta limitación puede no resultar adecuada, por ejemplo cuando se prueba una trayectoria de sonido en movimiento muy lento. En tales condiciones, cuando se determina que debe utilizarse un estímulo más largo, será necesario documentar en el informe final de la prueba por qué es necesario aumentar la duración.

El grupo de señales procesadas consta de todas las señales de prueba y de al menos dos señales patrón adicionales. El patrón normalizado es una versión filtrada en paso bajo de la señal original con una frecuencia de corte de 3,5 kHz. El patrón de calidad media tiene una frecuencia de corte de 7 kHz.

Las anchuras de banda de los patrones corresponden a las de las Recomendaciones para los circuitos de control (3,5 kHz) utilizados con fines de supervisión y coordinación en la radiodifusión, los circuitos de comentarios (7 kHz) y los circuitos ocasionales (10 kHz), conforme a las Recomendaciones UIT-T G.711, UIT-T G.712, UIT-T G.722 y UIT-T J.21, respectivamente.

La característica del filtro paso bajo de 3,5 kHz debe ser la siguiente:

f_c = 3,5 kHz

Rizado máximo en la banda de paso = ±0,1 dB

Atenuación mínima en 4 kHz = 25 dB

Atenuación mínima en 4,5 kHz = 50 dB.

Los patrones adicionales deben dar una indicación de la forma en que los sistemas sometidos a prueba se comparan respecto a niveles de calidad audio bien conocidos y no deben emplearse para ponderar los resultados entre pruebas diferentes.

5.2 Fase de adiestramiento

Para obtener resultados fiables es obligatorio enseñar a los evaluadores en sesiones especiales de adiestramiento con antelación a las pruebas. Se ha visto que este adiestramiento es importante para obtener resultados fiables. En el adiestramiento se debe exponer como mínimo al sujeto a toda la gama y naturaleza de las degradaciones, así como a todas las señales de prueba que recibirá durante el ensayo. Ello puede lograrse utilizando diversos métodos: un simple sistema de reproducción de cinta o un sistema interactivo controlado por computador. Las instrucciones figuran en el Adjunto 1. En esta fase también se familiarizará a los evaluadores con la configuración de la prueba subjetiva (por ejemplo, con el software utilizado para la prueba).

5.3 Presentación de los estímulos

El MUSHRA es un método de prueba doblemente ciega multiestímulo con referencia oculta y patrones ocultos, a diferencia del de la Recomendación UIT-R BS.1116 que utiliza un método de prueba doblemente ciega de triple estímulo con referencia oculta. Se considera que el enfoque MUSHRA es más adecuado para evaluar degradaciones de nivel medio y grande [4].

En una prueba que implique pequeñas degradaciones, la dificultad para el participante consiste en detectar todo efecto perturbador que pueda estar presente en la señal. En esta situación, es necesario incluir en la prueba una señal de referencia oculta, a fin de que el evaluador pueda evaluar la capacidad del participante para detectar satisfactoriamente estos efectos perturbadores. Por el contrario, en una prueba con degradaciones de nivel medio y grande, el evaluador no tiene dificultad para detectar los efectos parásitos, y por tanto no es necesaria a estos fines una referencia oculta. Además, la dificultad surge cuando el participante debe dar una nota de la incomodidad

relativa de los diversos efectos parásitos. En este caso, el sujeto debe valorar su preferencia por un tipo de efecto perturbador respecto a otro.

La utilización de una referencia de gran calidad introduce un problema interesante. Como la nueva metodología ha de utilizarse para evaluar degradaciones de nivel medio y grande, se prevé que la diferencia de percepción entre la señal de referencia y los elementos de prueba sea relativamente grande. A la inversa, las diferencias de percepción entre los elementos de prueba que pertenecen a distintos sistemas pueden ser bastante reducidas. Como resultado de ello, si se utiliza un método de prueba de múltiples tentativas (tal como el de la Recomendación UIT-R BS.1116) puede que los evaluadores tengan grandes dificultades para discriminar de forma precisa entre las diversas señales degradadas. Por ejemplo, en una comparación directa por pares, los evaluadores pueden concordar en que el Sistema A es mejor que el Sistema B. No obstante, en una situación en que cada sistema se compare únicamente con la señal de referencia (es decir, que el Sistema A y el B no se comparan directamente entre sí), pueden perderse las diferencias entre los dos sistemas.

Para superar esta dificultad, en el método de prueba MUSHRA el sujeto puede pasar a voluntad de la señal de referencia a cualesquiera de los sistemas en prueba, utilizando por lo general un sistema de respuesta controlado por computador, aunque puedan utilizarse otros mecanismos que emplean múltiples aparatos de disco compacto o de cinta. Se presenta al participante una secuencia de ensayos. En cada uno de ellos se le presenta la versión de referencia, el patrón bajo y el patrón medio, así como todas las versiones de las señales de prueba procesadas por los sistemas en prueba. Por ejemplo, si una prueba contiene 8 sistemas de audio, se permite al sujeto que escoja casi instantáneamente entre 11 señales de prueba y la referencia abierta (1 referencia + 8 sistemas de prueba + 1 referencia oculta + 1 patrón bajo oculto + 1 patrón medio oculto).

Como el sujeto puede comparar directamente las señales degradadas, este método ofrece la ventaja de una plena comparación por pares en la que el sujeto puede detectar más fácilmente las diferencias entre las señales degradadas y valorarlas en consecuencia. Este aspecto permite obtener un alto grado de resolución en las notas de valoración de los sistemas. No obstante, es importante señalar que los evaluadores obtendrán su valoración de un sistema determinado comparando dicho sistema con la señal de referencia, así como con las otras señales de cada tentativa.

Se recomienda no incluir en cada tentativa más de 12 señales (por ejemplo, 9 sistemas en prueba, 1 patrón bajo oculto, 1 patrón medio oculto y 1 referencia oculta).

En los raros casos en que haya de compararse un gran número de señales, el experimento deberá diseñarse por bloques, de lo cual se rendirá informe detallado.

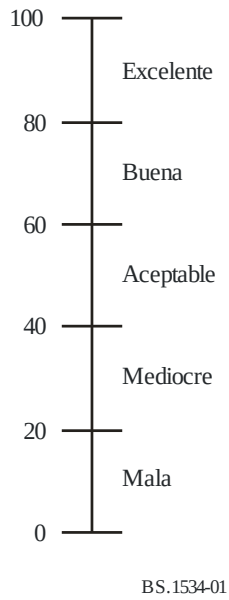
En las pruebas de la Recomendación UIT-R BS.1116, los evaluadores tienden a abordar un ensayo determinado empezando con un proceso de detección, al que sigue un proceso de valoración. La experiencia en la realización de ensayos según el método MUSHRA muestra que los evaluadores tienden a iniciar una sesión con una estimación somera de la calidad. A ello sigue un proceso de clasificación o de ordenación. Después de ello el sujeto efectúa el proceso de valoración. Como la ordenación por grados se efectúa de forma directa, es probable que los resultados de la calidad de audio intermedia sean más congruentes y fiables que los obtenidos si se hubiera utilizado el método de la Recomendación UIT-R BS.1116. Además, la duración mínima del bucle es de 500 ms y todo el contenido del bucle debe ir precedido y seguido de un desvanecimiento global de coseno alzado de 5 ms. Todo cambio de contenido entre sistemas de prueba debe ir precedido y seguido de un desvanecimiento global de coseno alzado de 5 ms. Durante la prueba, en ningún momento se utilizará la transición por desvanecimiento al cambiar de un sistema de prueba a otro. Con estas modificaciones se pretende reducir los cambios de color espectral durante la comparación de transitorios abruptos para identificar y calificar las señales que se prueban.

5.4 Proceso de valoración

Se pide a los evaluadores que den notas a los estímulos según la escala de calidad continua (CQS). La CQS consiste en unas escalas gráficas idénticas (normalmente de 10 cm de longitud o más) que se dividen en 5 intervalos iguales con los adjetivos dados en la Fig. 1.

Esta escala se utiliza también para la valoración de la calidad de la imagen (véase la Recomendación UIT-R BT.500 – Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión).

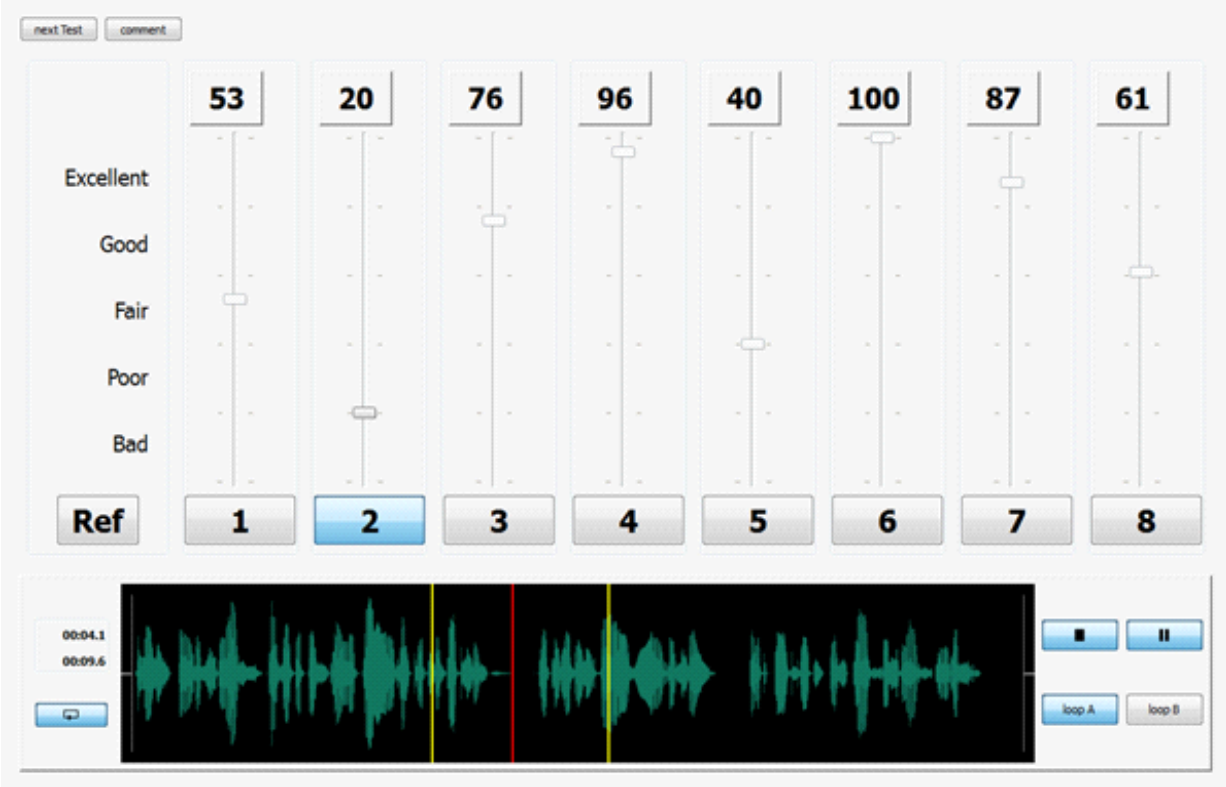
FIGURA 1



El participante registra su evaluación de la calidad en un formulario adecuado, por ejemplo, utilizando los cursores de una visualización electrónica (véase la Fig. 2) o utilizando un lápiz y una escala en papel. Si se utiliza una configuración como la que aparece en la Fig. 2, el participante debe limitarse únicamente a ajustar la puntuación asignada al fragmento que él o ella está escuchando. En el Adjunto 2 figuran algunas orientaciones sobre la creación de interfaces. Se pide al evaluador que califique la calidad de todos los estímulos, conforme a la CQS de cinco intervalos.

FIGURA 2

Ejemplo de visualización en computador para el ensayo MUSHRA



En comparación con el método de la Recomendación UIT-R BS.1116, el método MUSHRA tiene la ventaja de visualizar muchos estímulos al mismo tiempo, de forma que el sujeto puede verificar cualquier comparación entre ellos directamente. El tiempo empleado en realizar el ensayo utilizando el método MUSHRA puede ser significativamente inferior al del método de la Recomendación UIT-R BS.1116.

5.5 Grabación de las sesiones de prueba

Si se observa alguna anomalía en el procesamiento de las puntuaciones asignadas, sería muy útil disponer de una grabación de las circunstancias que las originaron. Una forma relativamente sencilla de lograrlo es realizar una grabación en vídeo o audio de toda la prueba. Si se descubre una nota anómala en una serie de resultados, es posible estudiar la grabación de la cinta para intentar aclarar si se debe a un error humano o a un funcionamiento inadecuado del equipo.

6 Atributos

A continuación se enumeran atributos específicos de evaluaciones monofónicas, estereofónicas y multicanal. Es preferible evaluar en cada caso el atributo «calidad audio básica». Los experimentadores pueden elegir la definición y evaluación de otros atributos.

Sólo tiene que evaluarse un atributo durante un ensayo. Cuando se pide a los evaluadores que evalúen más de un atributo en cada ensayo puede inducírseles a sobrecarga o confusión, si no a ambos, al tratar de responder a múltiples preguntas sobre un estímulo determinado. Esto puede dar lugar a evaluaciones no fiables para todas las cuestiones. Si se han de juzgar por separado múltiples propiedades del audio, se recomienda evaluar en primer lugar la calidad de audio básica.

6.1 Sistema monofónico

Calidad de audio básica: Este atributo simple y general se utiliza para evaluar cualesquiera de las diferencias detectadas entre la referencia y el objeto, y todas ellas.

6.2 Sistema estereofónico

Calidad de audio básica: Este atributo único y general se utiliza para evaluar cualesquiera de las diferencias detectadas entre la referencia y el objeto, y todas ellas. También pueden interesar los atributos siguientes:

Calidad de la imagen estereofónica: Este atributo se refiere a las diferencias entre la referencia y el objeto, en términos de emplazamientos de la imagen sonora y sensaciones de profundidad y de realidad de la presentación de audio. Aunque algunos estudios han demostrado que la calidad de la imagen estereofónica puede degradarse, no se han realizado aún investigaciones suficientes que indiquen la justificación de valorar por separado la calidad de la imagen estereofónica de manera distinta respecto a la calidad de audio básica.

NOTA 1 – Hasta 1993, la mayoría de los estudios de evaluación subjetiva de las pequeñas degradaciones en sistemas estereofónicos utilizaban exclusivamente el atributo de calidad de audio básica. De esta manera el atributo de la calidad de la imagen estereofónica se incluía de forma implícita o explícita en la calidad de audio básica como atributo general en dichos estudios.

6.3 Sistema multicanal

Calidad de audio básica: Este atributo único general se utiliza para evaluar cualesquiera de las diferencias detectadas entre la referencia y el objeto, y todas ellas.

También pueden interesar los atributos siguientes:

Calidad de la imagen frontal: Este atributo está relacionado con la localización de las fuentes sonoras frontales. Incluye la calidad de la imagen estereofónica y las pérdidas de definición.

Impresión de calidad panorámica: Este atributo se refiere a la impresión espacial, el ambiente o los efectos especiales panorámicos direccionales.

7 Material de prueba

Debe utilizarse material crítico que represente el programa de radiodifusión típico para la aplicación deseada, a fin de poner de manifiesto las diferencias entre los sistemas sometidos a prueba. El material crítico es aquel que fuerza los sistemas sometidos a prueba. No hay un material de programa universalmente adecuado que pueda utilizarse para evaluar todos los sistemas en todas las condiciones. En consecuencia, el material crítico debe determinarse explícitamente para cada sistema que haya que probar en cada experimento. La búsqueda del material adecuado suele ser ardua; no obstante, a menos que se utilice un material crítico realmente para cada sistema, los experimentos no conseguirán poner de manifiesto la diferencia entre sistemas y no serán determinantes. Un pequeño grupo de oyentes experimentados seleccionará los elementos de prueba a partir de una amplia selección de elementos posibles. En este proceso de selección, documentado y descrito en el resumen de la prueba, deberán estar incluidos todos los sistemas de prueba.

Debe demostrarse de forma empírica y estadística que la falta de detección de diferencias entre sistemas no es debida a la insensibilidad experimental que pudiera ser producida por la elección inadecuada del material de audio o de cualquier otro aspecto inconveniente del experimento, pues de otra manera esta determinación «nula» no podrá aceptarse como válida.

En la búsqueda del material crítico, debe ensayarse cualquier estímulo que pueda considerarse como material potencial de radiodifusión. No deben incluirse señales sintéticas concebidas deliberadamente para atacar un sistema específico. El contenido artístico o intelectual de una secuencia de programa no debe ser ni tan interesante ni tan desagradable o pesado que distraiga al sujeto de su enfoque de la detección de degradaciones. Debe tenerse en cuenta la frecuencia prevista de aparición de cada tipo de material de programa en las emisiones reales. No obstante, debe entenderse que el carácter del material difundido puede cambiar en el tiempo con los cambios futuros de los estilos y preferencias musicales.

Al seleccionar el material de programa, es importante que los atributos que hayan de evaluarse se definan de forma precisa. La responsabilidad de la selección del material debe delegarse en un grupo de evaluadores experimentados que cuenten con un conocimiento básico de las degradaciones que se prevén. Su punto de partida se basará en una amplia gama de materiales. Dicha gama puede ampliarse mediante grabaciones especializadas.

A los efectos de preparación de la prueba subjetiva formal, el grupo de los evaluadores preparados debe ajustar subjetivamente la sonoridad de cada pasaje, antes de grabarlo en el medio de prueba. Ello permitirá la utilización posterior del medio de prueba con una ganancia fija para todos los programas de todo un grupo de pruebas.

El grupo de evaluadores preparados convendrá para todas las secuencias de prueba unos niveles relativos de sonido de cada muestra ensayada. Además, los expertos deben llegar a un consenso sobre el nivel absoluto de presión acústica reproducida para el conjunto de la secuencia, en relación con el nivel de alineación.

Puede incluirse una ráfaga de tono (por ejemplo de 1 kHz, 300 ms, –18 dBFS) al nivel de alineación de la señal, al principio de cada grabación, a fin de poder ajustar el nivel de alineación de salida con el nivel de alineación de entrada por el canal de reproducción, conforme a la Recomendación R68 de la UER (véase el § 8.4.1 de la Recomendación UIT-R BS.1116). La ráfaga sólo tiene fines de alineación: no debe reproducirse durante la prueba. Debe controlarse la señal del programa sonoro de forma que las amplitudes de las crestas sólo excedan muy raramente la amplitud de cresta de la señal máxima permitida que se define en la Recomendación UIT-R BS.645 (una onda senoidal a 9 dB por encima del nivel de alineación).

El número factible de pasajes de audio que se incluye en una prueba varía: debe ser igual para cada sistema sometido a prueba. Una estimación razonable es de 1,5 veces el número de sistemas sometidos a prueba, con un valor mínimo de 5 pasajes. Dada la complejidad de la tarea, deben ponerse a disposición del probador los sistemas sometidos a prueba. Sólo puede efectuarse una selección adecuada si se define un programa de tiempos apropiado. Además, habida cuenta de que la velocidad de bits utilizada en los códecs de audio es variable en el tiempo, se recomienda codificar secuencias más largas y utilizar una parte de cada secuencia en la prueba de escucha.

El comportamiento de un sistema multicanal en condiciones de reproducción bicanal debe ensayarse utilizando un mezclado descendente de referencia. Aunque puede considerarse que la utilización de un mezclado descendente fijo puede ser restrictiva en algunas circunstancias, es sin duda la opción más sensible que pueden utilizar las autoridades de radiodifusión a largo plazo. Las ecuaciones para el mezclado descendente de referencia son (véase la Recomendación UIT-R BS.775):

$$L_0 = 1{,}00L + 0{,}71C + 0{,}71L_s$$

$$R_0 = 1{,}00R + 0{,}71C + 0{,}71R_s$$

$$R_0 = 1{,}00R + 0{,}71C + 0{,}71R_S$$

La preselección de los pasajes de prueba adecuados para la evaluación crítica del comportamiento de la mezcla descendente bicanal de referencia debe basarse en la reproducción de programas con mezcla descendente bicanal.

8 Condiciones de audición

Los métodos para la evaluación subjetiva de las pequeñas degradaciones de sistemas de audio, incluyendo los sistemas de sonido multicanal se definen en la Recomendación UIT-R BS.1116. Para evaluar sistemas de audio que tengan una calidad intermedia se deben utilizar las condiciones de audición que se describen en los § 7 y 8 de la Recomendación UIT-R BS.1116.

En la prueba pueden utilizarse auriculares o altavoces. No se permite el empleo de ambos en una misma sesión: todos los evaluadores deben emplear el mismo tipo de transductor.

Para medir una señal con una tensión eficaz igual al «nivel de la señal de alineación» (0 dBu0s conforme a la Recomendación UIT-R BS.645; –18 dB por debajo del nivel de recorte de una grabación en cinta digital, según la Recomendación R68 de la UER) aplicada a su vez a la entrada de cada canal de reproducción (es decir, un amplificador de potencia y su altavoz correspondiente) debe ajustarse la ganancia del amplificador para obtener un nivel de presión acústica de referencia (con ponderación CEI/A, lento) de:

$$L_{ref} = 85 - 10 \log n \pm 0{,}25 \text{ dBA}$$

donde n es el número de canales de reproducción del conjunto.

Se admite el ajuste individual del nivel de audición por un sujeto en una sesión, debiéndose limitarse a la gama de ±4 dB respecto al nivel de referencia definido en la Recomendación UIT-R BS.1116. El equilibrio entre las unidades de prueba de un ensayo debe lograrse al nivel del grupo de selección, de forma que los evaluadores no tengan normalmente que realizar ajustes individuales para cada unidad.

No se permitirán los ajustes de nivel en una unidad.

9 Análisis estadístico

Las separaciones de cada condición de prueba se convierten linealmente, pasando de mediciones de longitud en la hoja de valoración a valoraciones normalizadas en la gama de 0 a 100, donde el 0 corresponde al mínimo de la escala (calidad mala). A continuación, se calculan las valoraciones absolutas de la siguiente manera.

Puede realizarse un análisis estadístico paramétrico o no paramétrico a partir del cumplimiento de los supuestos estadísticos (véase el § 9.3.3). Pueden encontrarse orientaciones sobre el análisis estadístico paramétrico en el Adjunto 4.

9.1 Visualización de datos y análisis exploratorio de datos

El primer paso del análisis estadístico debe ser siempre la visualización de los datos brutos. Se pueden utilizar también histogramas con la curva de distribución normal, gráficos de caja o gráficos cuartil-cuartil.

La visualización de los datos en gráficos de caja dará una indicación de la existencia y efecto de valores extremos en los resúmenes descriptivos de los datos. Con la visualización se quiere identificar la dispersión y desviación de las calificaciones individuales con respecto a la media de todos los evaluadores. Se recurrirá a los histogramas para identificar la presencia de una distribución multimodal subyacente. Si en los datos se ve claramente una distribución multimodal, el responsable del experimento deberá analizar la distribución por separado.

Para evaluar el grado de multimodalidad, *b*, se puede utilizar la siguiente fórmula:

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

donde:

n: tamaño de la muestra

g: asimetría de la muestra finita

k: curtosis excesiva de los resultados de la prueba de escucha.

Este coeficiente oscilará entre 0 y 1. Los valores más elevados (> 5/9) pueden interpretarse como una indicación de multimodalidad.

A partir del examen visual de estos gráficos, de *b* y de los supuestos sobre la población subyacente de la muestra observada, se decidirá si se debe suponer o no que se observa una distribución normal. Si el ajuste de curva está claramente sesgada, el histograma contiene muchos valores extremos y el gráfico cuartil-cuartil no es en absoluto una línea recta, no se considerará que la muestra tiene una distribución normal. El cálculo de las medianas de las valoraciones normalizadas de todos los oyentes restantes después de la postselección serán las notas medianas subjetivas.

La mediana debe calcularse de la siguiente manera: $\hat{x} = \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ impar} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ par} \end{cases}$,

x se ordena por tamaño.

El primer paso del análisis es el cálculo de la nota mediana, $\overline{\eta}_{jk}$ para cada una de las presentaciones. Se deduce que η_{ijk} es la nota mediana del observador *i* para una condición de prueba *j* y secuencia de audio *k* determinadas y $\hat{\eta}$ es la mediana de la muestra (todos los observadores, todas las condiciones, todas las secuencias de audio).

De forma similar, pueden calcularse las notas medianas, $\overline{\eta}_j$ y $\overline{\eta}_k$, para cada condición de prueba y cada secuencia de prueba.

Aunque es necesario utilizar los valores medianos para algunos métodos de análisis, como ANOVA (véase el § 9.3), el cálculo de la mediana es una medida alternativa de la tendencia central. La mediana ofrece una medida sólida de la tendencia central, óptima cuando la muestra es pequeña, la distribución es no normal o los datos contienen valores extremos notables. Es posible que haya muchas situaciones de prueba donde estos problemas tengan menos peso. Sin embargo, dado que una de las grandes ventajas de las pruebas normalizadas es la comparación e interpretación de las calificaciones entre usuarios y lugares, conviene identificar los métodos de análisis más robustos y menos sensibles a factores que pueden alterar la validez o reducir la traducción prueba a prueba.

De este modo se pueden aplicar estadísticas no paramétricas. Cuando se recurre al análisis no paramétrico de los datos, las medias y los intervalos de confianza del 95% deben calcularse con los métodos disponibles, como el algoritmo autodocimante común.

Se pueden calcular las medidas de error en torno a la mediana utilizando la desviación media absoluta:

$$\hat{\tau} = \Sigma \left| Y_i - \hat{\eta} \right| / n$$

Se recomienda utilizar el rango intercuartil (IQR) como medida de la confianza en torno a la mediana. Se trata de la diferencia entre el primer y el tercer cuartil: $IQR = Q_3 - Q_1$. Pueden encontrarse las fórmulas en el § 4.1.2. Si la distribución de los resultados es normal, el IQR representa dos veces la desviación media absoluta.

Se recomienda identificar la significancia estadística a partir de un nivel de significancia del 95%. Las pruebas no paramétricas de aleatorización sirven para determinar la significancia estadística. Contrariamente a los análisis estadísticos paramétricos, estas pruebas no se basan en hipótesis sobre la distribución subyacente de los datos y son menos sensibles a los muchos problemas que conlleva la utilización de una muestra más pequeña.

Con una prueba no paramétrica de aleatorización (prueba de permutación) robusta se puede identificar la probabilidad de que una diferencia observada entre dos condiciones de prueba ocurriese si los datos fueran verdaderamente aleatorios, como se supone en la hipótesis nula. La probabilidad medida con esta prueba es una medida real determinada a partir de la distribución de los datos reales, en lugar de una medida inferida que supone una distribución subyacente de forma específica [5]. Este tipo de prueba necesita de técnicas de remuestreo común, como las técnicas de simulación autodocimante y Monte-Carlo, que pueden utilizarse fácilmente gracias a la mayor velocidad de las computadoras actuales [6]. En el Adjunto 3 puede encontrarse una descripción detallada de este método de prueba.

9.2 Análisis de potencia

El análisis de potencia puede servir para estimar el tamaño de muestra necesario para las pruebas de escucha, si se aplica como un análisis *a priori*, y para estimar la potencia o el error de tipo II de la prueba, si se efectúa como un análisis *a posteriori.* El análisis *a priori* da como resultado el tamaño de la muestra necesario para el experimento dado un tamaño de efecto $d = \frac{\bar{x}}{s}$, un nivel de significancia α y una potencia estadística de $1 - \beta$.

Por otra parte, aplicando el análisis *a posteriori* se obtiene la potencia $1 - \beta$ o el error de tipo II β de la prueba con un tamaño de efecto $d = \frac{\bar{x}}{s}$, un nivel de significancia α y un tamaño de muestra N. El error de tipo II β es la probabilidad de que el efecto d exista en la población, aunque la prueba no lo haya considerado importante. Si, por ejemplo, una prueba sostiene que la calidad no se ve afectada por el sistema, $1 - \beta$ es la probabilidad de que la degradación se demuestre con la prueba.[2]

9.3 Aplicación y utilización de ANOVA

9.3.1 Introducción

En esta sección se examinan los requisitos para realizar estadísticas paramétricas utilizando el análisis de varianza (ANOVA). Dada la solidez del modelo ANOVA (véanse [7]; [8]; [12]; [13]) y su potencia estadística[3], es un método adaptado a los datos obtenidos utilizando la Recomendación UIT-R BS.1534. Dado que la estadística F de ANOVA es bastante robusta ante las distribuciones de datos no normales y la varianza heterogénea, la hipótesis de prueba se centra en la naturaleza del error o en los valores residuales.

---

[2] Hay muchas herramientas, como G*Power [16] para realizar automáticamente el análisis de potencia de las distribuciones de población conocidas, mientras que resulta más difícil estimar la potencia de las poblaciones desconocidas.

[3] Generalmente se aconseja seleccionar el método de análisis estadístico más potente que permitan los datos [9] [10].

Puede encontrarse más información sobre las hipótesis generales de las estadísticas paramétricas en el Adjunto 4.

9.3.2 Especificación de un modelo

Se aconseja vivamente que, al diseñar el experimento (véase el § 3), el modelo se especifique detalladamente en cuanto a las variables independientes (por ejemplo, MUESTRA, SISTEMA, CONDICIÓN, etc.) y a las variables dependientes (por ejemplo, calidad de audio básica o esfuerzo de escucha, etc.). También se definirán en la fase de especificación del modelo los niveles de cada variable dependiente.

Al definir un modelo de análisis (por ejemplo, utilizando ANOVA para análisis de varianza o ANOVA para repetición de la medición), es importante incluir todas las variables significativas. La omisión de variables significativas, por ejemplo, interacciones bidireccionales o tridireccionales de factores independientes, puede llevar a la especificación errónea del modelo, lo que, a su vez, dará una explicación insuficiente de la varianza (R^2) y podrá causar una interpretación errónea del análisis de los datos.

9.3.3 Lista de verificación para el análisis estadístico paramétrico

La siguiente lista de verificación sirve de breve guía para la comprobación de los datos, las hipótesis básicas (paramétricas y no paramétricas), así como los pasos básicos de la estadística paramétrica. Esta lista se centra en los requisitos del análisis de varianza, pues es un método adecuado para analizar datos obtenidos a partir de experimentos conformes a la Recomendación UIT-R BS.1534. Puede encontrarse una guía completa en los manuales de estadística (por ejemplo [8]; [11]; [9]).

- Estadísticas exploratorias[4]
  - Verificar que la estructura de los datos es correcta y es la esperada.
  - Verificar la ausencia de datos.
  - Estudiar la normalidad de la distribución de datos.
  - Examinar otras posibles distribuciones de datos (unimodal, bimodal, sesgada, etc.).
- Unidimensionalidad
  - Verificar que todos los evaluadores utilizan la misma escala[5].
  - Verificar que los datos son de naturaleza unidimensional.
  - Análisis de los componentes principales, gráficos Tucker-1 o alfa de Cronbach.
- Independencia de las observaciones
  - Se suele definir en la metodología experimental y no es fácil medirla estadísticamente. Se ha de garantizar que los datos se han obtenido de manera independiente, es decir, con técnicas de experimentos en doble ciego y garantizando que los evaluadores no se han influido mutuamente.

---

[4] Se aplica a las estadística paramétricas y no paramétricas por igual.

[5] Se han observado casos de multidimensionalidad cuando hay subpoblaciones con distintas opiniones relativas a la evaluación de efectos perturbadores particulares.

- Homogeneidad de la varianza[6]
  - Probar el supuesto de que cada variable independiente muestra una varianza semejante.
    - Examen visual de gráficos de caja en paralelo para comprobar el nivel de las variables independientes. Por lo general la heterogeneidad puede variar por un factor máximo de cuatro.
    - Se pueden utilizar la prueba de Brown y Forsythe o la estadística de Leven para evaluar la homogeneidad de la varianza.
- Distribución normal de los valores residuales
  - Probar la distribución normal de los valores residuales.
    - Prueba Kolmogorov-Smirnov D, prueba K-S Lillefors o prueba de Levene.
    - También pueden utilizarse el gráfico de probabilidad normal (denominado gráfico P-P) o el gráfico cuantil por cuantil (denominado gráfico Q-Q) como prueba visual de la distribución normal.
- Detección de valores extremos
  - Se han de detectar, y descartar cuando sea necesario, los valores extremos. Puede encontrarse más información al respecto en el § 4.1.2.
- Análisis
  - Análisis de varianza (ANOVA) – Modelo lineal general o modelo ANOVA para repetición de medición.
    - Emplear el modelo ANOVA adecuado, por ejemplo el modelo lineal general (GLM) o el modelo ANOVA para repetición de medición. Pueden encontrarse más detalles al respecto en el Adjunto 4.
    - Especificar el modelo en función del diseño del experimento.
      - Incluir, cuando sea posible, las interacciones bidireccionales y tridireccionales.
    - Analizar los datos con el modelo y los resultados.
      - Examinar la varianza explicada (R^2) del modelo utilizado para describir la variable dependiente.
      - Examinar la distribución del error residual.
      - Examinar los factores significativos y no significativos.
    - El modelo puede repetirse para eliminar valores extremos y factores no significativos.
  - Pruebas post hoc
    - Realizar pruebas post hoc para determinar la importancia de la diferencia entre medias, cuando el factor dependiente (o la interacción del factor) es importante en el ANOVA.
    - Existen diversas pruebas post hoc con distintos niveles de discriminación, por ejemplo, diferencia menos significativa de Fisher (LSD), diferencia verdaderamente significativa de Tukey (HSD), etc.
    - Se recomienda consignar el tamaño de los efectos junto con los niveles de importancia.

---

[6] Necesaria para la aplicación de ANOVA, pero no para la de rmANOVA (ANOVA para repetición de mediciones) (véase el Adjunto 4).

- Extracción de conclusiones
  - Una vez realizado el análisis, resumir las conclusiones mediante un gráfico de las medias y los intervalos de confianza del 95% asociados de los datos brutos o sometidos al modelo ANOVA (en ocasiones denominados media marginal estimada).
  - Cuando las interacciones del factor (por ejemplo, bidireccional o tridireccional) son importantes, deberán incluirse en un gráfico para ofrecer un panorama completo de los datos. En tales casos, la inclusión en el gráfico sólo de los efectos principales dará un panorama de los datos con confusión del efecto de la interacción.

Pueden encontrarse más orientaciones sobre la utilización de los modelos ANOVA en el Adjunto 4 y en los textos sobre estadística general y aplicada, por ejemplo, [11]; [13]; [15].

10 Informe de prueba y presentación de resultados

10.1 Generalidades

La presentación de los resultados debe efectuarse de manera que sea fácil para el usuario, a fin de que todo lector, no iniciado o experto, pueda obtener la información pertinente. En principio, todo lector desea ver el resultado experimental global, preferentemente de forma gráfica. Una presentación de este tipo puede estar apoyada por información cuantitativa con más detalle, aunque los análisis numéricos detallados deben figurar en los adjuntos.

10.2 Contenido del informe de prueba

El informe de prueba debe incluir, de la forma más clara posible, los fundamentos del estudio, los métodos utilizados y las conclusiones obtenidas. Debe presentarse un detalle suficiente de forma que la persona con conocimientos pueda, en principio, realizar de nuevo el estudio para verificar el carácter empírico del resultado. Aun así, no es necesario que el informe incluya todos los resultados individuales. Un lector informado debe poder comprenderlo y elaborar una crítica de los detalles importantes del ensayo, así como sobre los motivos subyacentes del estudio, los métodos del diseño experimental, la ejecución, el análisis y las conclusiones.

Debe prestarse especial atención a lo siguiente:

- una presentación gráfica de los resultados;
- una presentación gráfica de la selección y la especificación de los *evaluadores experimentados* escogidos;
- la definición del diseño experimental;
- la especificación y selección del material de prueba;
- una información general sobre el sistema utilizado para procesar el material de prueba;
- los detalles de la configuración de la prueba;
- los detalles físicos del entorno de audición y del equipo, incluyendo las dimensiones y características acústicas de la sala, los tipos y emplazamientos de los transductores y la especificación del equipo eléctrico (véase la Nota 1);
- el diseño experimental, la capacitación, las instrucciones, las secuencias experimentales, los procedimientos de prueba y la generación de datos;
- el tratamiento de los datos, incluyendo los detalles de las estadísticas descriptivas y su inferencia analítica;

– la utilización de patrones en la prueba;

– los métodos de post selección utilizados en el análisis de los resultados, incluidos los métodos de exclusión de valores extremos o de oyentes no formados;

– si la prueba se completó utilizando la Recomendación UIT-R BS.1534 o la Recomendación UIT-R BS.1534-1; se habrá de indicar claramente en el documento donde se describan las condiciones de patrón empleadas;

– la definición adecuada y la generación del código necesario para permitir a un nuevo usuario producir cualquiera de los patrones empleados en la prueba que no se describan explícitamente en esta Recomendación;

– la base detallada de todas las conclusiones obtenidas.

NOTA 1 – Como hay una cierta evidencia de que las condiciones de la audición, por ejemplo la reproducción con altavoz o con auriculares, puede influir en los resultados de las evaluaciones subjetivas, se pide a los experimentadores que informen explícitamente de las condiciones de audición y del tipo de equipo reproductor utilizado en los experimentos. Si se desea realizar un análisis estadístico combinado de los distintos tipos de transductores, se debe verificar si es posible dicha combinación de los resultados.

10.3 Presentación de los resultados

Para cada parámetro en prueba debe indicarse la mediana y el IQR de la distribución estadística de las notas de evaluación.

Los resultados deben darse junto con la información siguiente:

– descripción de los materiales de prueba;

– número de evaluadores;

– una presentación gráfica de los resultados; gráficos de caja que muestren el IQR, además de una presentación de las medias y los intervalos de confianza del 95%; se incluirán las diferencias notables entre los sistemas que se prueban así como el método de análisis estadístico aplicado.

Además, los resultados también se pueden presentar en formas adecuadas, como medias e intervalos de confianza, cuando los datos lo permitan, según la visualización del gráfico de caja.

10.4 Notas absolutas

Una presentación de las notas medias absolutas para el sistema en pruebas, de la referencia oculta y de los patrones ofrece una buena panorámica del resultado. No obstante, se debe tener presente que ello no ofrece información alguna sobre el detalle del análisis estadístico. En consecuencia, las observaciones no son independientes y el análisis estadístico sólo de las notas, sin tener en cuenta la población subyacente de la muestra observada, no conduce a una información significativa. Además, se ha de informar de la aplicación de los métodos estadísticos propuestos en el § 9.

10.5 Nivel de significación e intervalo de confianza

El informe de prueba debe aportar al lector información sobre el carácter inherentemente estadístico de todos los datos subjetivos. Deben indicarse los niveles de significación, así como otros detalles sobre los métodos estadísticos y los resultados que facilitarán la comprensión por parte del lector. Dichos detalles pueden incluir intervalos de confianza o barras de error en gráficos.

Evidentemente, no hay un nivel de significación «correcto». No obstante, se elige tradicionalmente el valor de 0,05. En principio, es posible utilizar una prueba de una rama o de dos, dependiendo de las hipótesis formuladas.

Referencias

[1] Stevens, S. S. (1951). Mathematics, measurement and psychophysics, in Stevens, S. S. (ed.), Handbook of experimental psychology, John Wiley & Sons, New York.

[2] EBU [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, January.

[3] EBU [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, June.

[4] Soulodre, G. A., & Lavoie, M. C. (1999, August). Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

[5] Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527-542.

[6] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[7] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. Lawrence Erlbaum Associates, Inc.

[8] Keppel, G. and Wicken., T. D. (2004). Design and Analysis. *A Researcher's Handbook*, 4th edition. Pearson Prentice Hall.

[9] Garson, D. G. Testing statistical assumptions, Blue Book Series, Statistical Associates Publishing, 2012.

[10] Ellis, P. D. (2010). The essential guide to effect sizes. *Cambridge: Cambridge University Press*, 2010, 3-173.

[11] Howell., D.C. (1997). Statistical methods for psychology, 4th Edition, Duxbury Press.

[12] Kirk., R.E., (1982). Experimental Design: Procedures for the Behavioural Sciences, 2nd edition. Brooks/Cole Publishing Company 1982.

[13] Bech, S., & Zacharov, N. (2007). Perceptual audio evaluation-Theory, method and application. John Wiley & Sons.

[14] Khan, A. and Rayner, G. D. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, *Journal of Applied Mathematics & Decision Sciences*, 7(4), 187-206.

[15] ITU-T. Practical procedures for subjective testing, International Telecommunication Union, 2011.

[16] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,(4),* 1149-1160.

Adjunto 1
al Anexo 1
(Normativo)

Instrucciones que han de darse a los evaluadores

A continuación figura un ejemplo del tipo de instrucciones que deben darse o leerse a los evaluadores para instruirles en cuanto a la forma de realizar la prueba.

1 Familiarización o fase de adiestramiento

El primer paso en las pruebas de audición es familiarizarse con el proceso de pruebas. Esta fase se denomina de adiestramiento y precede a la fase de evaluación formal.

El objetivo de la fase de adiestramiento es permitir al evaluador lograr los dos objetivos siguientes:

– **Parte A**: familiarizarse con todos los pasajes sonoros en pruebas y con sus gamas de nivel de calidad; y

– **Parte B**: aprender la forma de utilizar el equipo de prueba y la escala de valoración.

En la Parte A de la fase de adiestramiento se podrán escuchar todos los pasajes sonoros que se hayan seleccionado para las pruebas, a fin de ilustrar la gama completa de calidades posibles. Los elementos sonoros que se escucharán serán más o menos críticos, dependiendo de la velocidad binaria y de otras «condiciones» utilizadas. La Fig. 3 muestra la interfaz de usuario. Apretando los distintos botones se oyen los diferentes pasajes sonoros, incluidos los pasajes de referencia. De esta manera se puede aprender a apreciar una gama de niveles distintos de calidad para los diferentes elementos de programa. Los pasajes sonoros se agrupan sobre la base de condiciones comunes. Se identifican tres grupos de este tipo en cada caso. Cada grupo incluye cuatro señales procesadas.

En la Parte B de la fase de adiestramiento se aprenderá a utilizar la reproducción disponible y el equipo de valoración que se empleará para evaluar la calidad de los pasajes sonoros.

Durante la fase de capacitación debe saberse la forma en que, como individuo, se interpretan las degradaciones audibles en términos de escala de graduación. No debe discutirse la interpretación personal de la escala con los otros evaluadores en ningún momento durante la fase de adiestramiento. No obstante, se le alienta a explicar los efectos perturbadores a los demás evaluadores.

En las pruebas reales no se tendrán en cuenta las notas otorgadas durante la fase de adiestramiento.

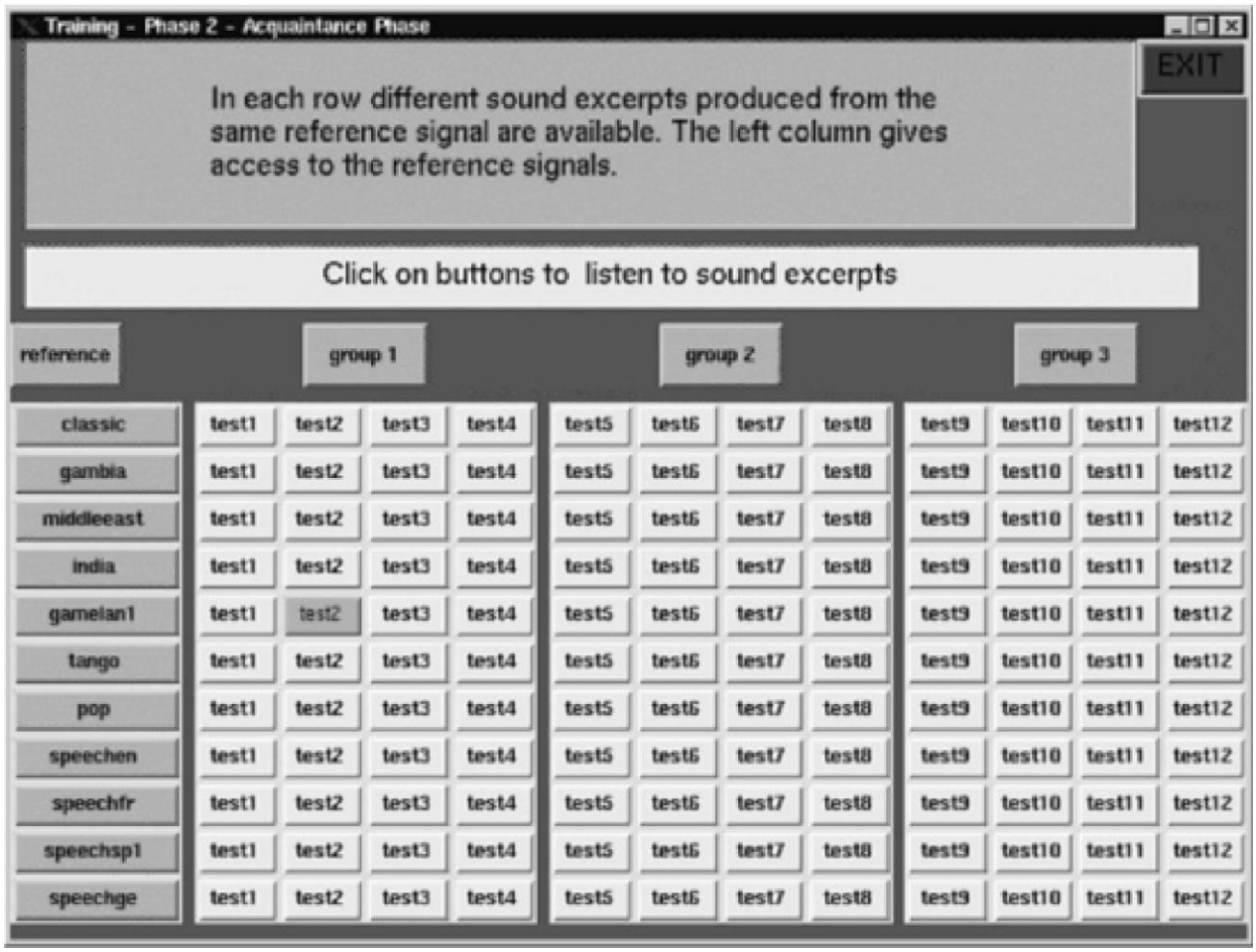
2 Fase de valoración ciega

El objetivo de la fase de valoración ciega es asignar notas utilizando la escala de calidades. Las notas del observador reflejarán su evaluación subjetiva del nivel de calidad para cada uno de los pasajes sonoros presentados. Cada tentativa contiene 9 señales que han de evaluarse. Cada uno de los elementos dura aproximadamente 10 s. Se debe escuchar la referencia, el patrón y todas las condiciones de prueba apretando en los botones respectivos. Se pueden escuchar las señales en cualquier orden y el número de veces que se desee.

Se utiliza la regla para cada señal, indicando la opinión de su calidad. Cuando se está satisfecho de la valoración de todas las señales, se debe apretar el botón «register scores», al final de la pantalla.

FIGURA 3

Imagen que muestra un ejemplo de interfaz de usuario para la Parte A de la fase de adiestramiento



BS.1534-03

Al asignar notas se utilizará la escala de calidad dada en la Fig. 1:

La escala de valoraciones es continua de «excelente» a «mala». Una nota 0 corresponde al mínimo de la categoría «mala» mientras que una nota de 100 corresponde al máximo de la categoría «excelente».

Al evaluar los pasajes sonoros, véase que no se ha de dar necesariamente una nota en la categoría «mala» al pasaje sonoro que tenga la categoría mínima en la prueba. No obstante, puede darse una nota 100 a uno o más pasajes, porque se incluye la referencia no procesada como uno de los pasajes que hay que evaluar.

FIGURA 4

Ejemplo de interfaz de usuario utilizada en la fase de valoración ciega

BS.1534-04

Adjunto 2
al Anexo 1
(Informativo)

Orientaciones sobre el diseño de interfaces de usuario

Las siguientes sugerencias están dirigidas a aquellos que probablemente tengan la intención de:

a) generar sistemas de prueba para realizar pruebas subjetivas de acuerdo con el método MUSHRA;

b) realizar dichas pruebas.

Estas sugerencias tienen por objeto aumentar la fiabilidad de los resultados de las pruebas y facilitar el análisis de cualquier irregularidad que pueda surgir durante el procesamiento de las puntuaciones de las pruebas.

La interfaz de usuario debe estar diseñada de modo tal que permita la posibilidad de que los participantes asignen una puntuación que no se ajuste a sus verdaderas intenciones. Para ello, habría que adoptar medidas con el fin de garantizar que la interfaz de usuario indique con claridad las versiones procesadas del fragmento de la prueba que está escuchando el participante en un determinado momento. Esto puede mejorarse si se eligen con cuidado los colores y brillos de los indicadores en pantalla (por ejemplo, botones sensibles) con el fin de evitar posibles dificultades si uno de los participantes no reconoce ciertos colores.

Asimismo, habría que cerciorarse de que el participante sólo pueda ajustar la puntuación asignada al fragmento que está escuchando. Se ha visto que algunos evaluadores escuchan de manera consecutiva dos versiones procesadas de un fragmento, con el fin de asignar una puntuación a la primera, y no a la última. En este caso, se podrían cometer errores (especialmente cuando aparece un gran número de controles en pantalla) y se podría asignar la puntuación a una señal distinta de la que se pretendía. Para tratar de reducir tal riesgo, se sugiere que se active en cada escucha un solo control, esto es, el correspondiente a la señal que se esté escuchando. Se deberían desactivar los controles que sirven para asignar puntuaciones a otras señales que no se estén escuchando.

Adjunto 3
al Anexo 1
(Normativo)

Descripción de la comparación estadística no paramétrica entre dos muestras utilizando técnicas de remuestreo y métodos de simulación Monte-Carlo

Se pueden utilizar pruebas no paramétricas de aleatorización con técnicas de remuestreo comunes, como los procedimientos autodocimantes, para determinar la importancia de casi cualquier resultado estadístico. Por ejemplo, la importancia de la diferencia de respuesta mediana observada entre dos señales de prueba (de tamaño de muestra = $N1$ y $N2$) puede calcularse como se explica a continuación. Debe anotarse la diferencia real entre las medianas de cada muestra, que se denominará $Diff_{ACT\_1}$. Todos los datos de estas muestras se agregarán en un único fichero o vector. Se utilizará entonces un procedimiento autodocimante de manera que cada iteración del conjunto se permutará con muestras extraídas de tamaño $N1$ y $N2$ sin sustitución. La diferencia entre las medianas de dos muestras aleatoriamente extraídas se denominará $Diff_{EST\_1}$. Este procedimiento puede repetirse 10 000 veces y la relación del número de veces en que $Diff_{EST\_N}$ es superior a $Diff_{ACT\_N}$, dividido por 10 000, arrojará el correspondiente valor p. Si el número total de veces en que $Diff_{EST\_N}$ es superior a $Diff_{ACT\_N}$ es inferior a 500 (500/10 000 = 0,05), se puede decir que la diferencia entre las dos medias es significativa a un nivel de 0,05, $p < 0,05$.

Adjunto 4
al Anexo 1
(Informativo)

Orientaciones para el análisis estadístico paramétrico

1 Introducción

En el § 9 puede encontrarse la descripción de un análisis estadístico paramétrico básico de los resultados de las pruebas MUSHRA. Sin embargo, sobre todo cuando se comparan muchas condiciones entre ellas, es más conveniente utilizar un test global, como ANOVA, para efectuar múltiples comparaciones entre pares. En este Adjunto se explica cómo llevar a cabo este procedimiento, se incluyen los requisitos previos al análisis, y se ofrecen alternativas para cuando no se cumplen esos requisitos.

La prueba MUSHRA utiliza *medidas repetidas* o un *diseño intrasujeto* (puede encontrarse una excelente introducción a estos conceptos en Maxwell y Delaney, 2004), donde se combinan completamente dos factores intrasujeto (condición y material de audio), y se obtiene al menos una calificación para cada combinación de oyente, material de audio y condición. También puede pasar que se presenten las mismas combinaciones de material de audio y condición a dos o más grupos de evaluadores distintos, por ejemplo, en distintos laboratorios. En tal caso, habrá un *grupo* de factores intersujeto adicional que deberán tenerse en cuenta en el análisis.

Es necesario recurrir a la estadística inferencial para generalizar los resultados obtenidos en una muestra pequeña de oyentes comparable a la población de todos los oyentes. Por ejemplo, si en la prueba de escucha las calificaciones indican una diferencia entre la calidad de audio percibida de un nuevo codificador y la del codificador tradicional, conviene averiguar si también podría esperarse esta diferencia si un grupo completamente distinto de oyentes calificase la calidad de audio de los dos sistemas. En lo que respecta al diseño específico de las pruebas de escucha MUSHRA, hay al menos tres cuestiones que conviene aclarar (o, en términos estadísticos, hipótesis que conviene comprobar), y la estadística inferencial descrita aquí ofrece respuestas válidas. La primera cuestión de mayor interés será normalmente si la calidad de audio percibida difiere entre los sistemas que se prueban (por ejemplo, la referencia y tres codificadores distintos). En segundo lugar si, suponiendo que en la prueba de escucha los sistemas de audio se evaluarán con materiales de prueba distintos, las calificaciones de la calidad de audio dependerían del material de audio. En tercer lugar, cabe determinar si el efecto del sistema de audio en la calidad de audio percibida difiere entre los materiales de prueba. La mejor manera de responder a estas cuestiones es obtener en primer lugar pruebas de importancia del principal efecto de la condición (sistema de audio), el principal efecto del material de audio y la interacción condición × material de audio, realizando un análisis de la varianza (ANOVA). Hay interacción cuando las diferencias entre la calidad percibida de los sistemas de audio dependen del material de audio. Téngase en cuenta que, a causa de las posibles interacciones, no conviene agregar las calificaciones de cada sistema de audio en función de los materiales de audio, aun cuando no se está particularmente interesado en el efecto del material de audio o en efecto de la interacción. Con comparaciones adicionales pueden probarse hipótesis más concretas, por ejemplo sobre la diferencia percibida entre un par de sistemas de audio.

Siempre que se hayan de comparar más de dos condiciones del experimento, por ejemplo, cuatro codificadores diferentes, no conviene basar la estadística inferencial en múltiples comparaciones por pares. Por ejemplo, si se incluyen en la prueba $K = 5$ sistemas de audio (4 codificadores más la referencia) habrá $\binom{K}{2} = K(K-1)/2 = 10$ pares de condiciones. Probar las diferencias entre cada uno de estos 10 pares mediante pruebas *t* de 10 parejas-muestras a un nivel α de 0,05 dará como resultado una inflación de la denominada tasa de errores de Tipo I por familia. Para cada prueba *t* individual, la probabilidad de rechazar erróneamente la hipótesis nula de no diferencia entre la calidad de audio percibida de los dos codificadores es α.

A lo largo de C pruebas de este tipo, la probabilidad de cometer al menos un error de Tipo I es $1 - (1 - \alpha)^C$, lo que para nuestro ejemplo de $C = 10$ da 0,40, lo que es muy superior al nivel α deseado de 0,05. La tasa de errores por familia puede controlarse efectuando las correcciones adecuadas para las pruebas múltiples, como la corrección de Bonferroni o el procedimiento de Hochberg (1988), que se describe a continuación. Sin embargo, las pruebas *t* por pares con corrección aún ocultan información pertinente, en parte porque la ejecución de múltiples pruebas *t* en todos los pares implica la utilización redundante de la información (cada media aparece en varias pruebas). Normalmente, las pruebas por pares serán menos potentes (es decir, menos sensibles a la hora de detectar una diferencia entre las condiciones) que una prueba global adecuada, que en el caso de MUSHRA es un análisis de repetición de medidas de la varianza (rmANOVA). En las siguientes cláusulas se describe cada uno de los pasos del análisis de datos para una

prueba MUSHRA sin factores intersujeto. Dicho de otro modo, se supone que sólo se ha probado un grupo de evaluadores y que se han presentado a cada evaluador todas las combinaciones de condición y material de audio al menos una vez. La extensión a un diseño con más de un grupo (por ejemplo, cuando la prueba se realiza en dos laboratorios) se describirá más adelante.

2 Prueba de normalidad

Resulta prudente considerar los efectos de una posible desviación con respecto a la medida de normalidad de la respuesta en la validez de la prueba estadística. Para un diseño intersujeto, donde cada evaluador se prueba en una sola condición experimental, las ANOVA realizadas en el marco del modelo lineal general son sorprendentemente robustas con respecto a la medida de no normalidad de la respuesta (por ejemplo [11]; [13]; [25]; [35]).

Para el diseño de la repetición de medidas, como en la prueba MUSHRA, en primer lugar hay que decir que hay una manera alternativa de probar la hipótesis nula de que en la población la calidad de audio percibida es idéntica para todas las condiciones. Esto es equivalente a calcular los contrastes ortogonales de $K - 1$, por ejemplo formando variables de diferencia entre las K condiciones, y probando entonces la hipótesis de que la media de población de todas esas variables de diferencia es igual a 0. Por ejemplo, si las pruebas comprenden la referencia y dos codificadores, se pueden crear dos variables de diferencia, D_1 *y* D_2, calculando para cada sujeto la diferencia entre la calificación de la referencia y la calificación del codificador A (D_1) y la diferencia entre la calificación del codificador A y la calificación del codificador B (D_2). ANOVA de repetición de mediciones asume que estas variables de diferencia están multinormalmente distribuidas. Por desgracia, a diferencia del caso del diseño intersujeto, la no normalidad puede resultar en tasas de errores de Tipo I demasiado conservadoras o demasiado liberales ([5]; [22]; [30]; [39]). Esto significa que para un determinado nivel α (por ejemplo, $\alpha = 0{,}05$), la proporción de casos en que ANOVA arroja un valor p significativo ($p < \alpha$), aunque la hipótesis nula de medias idénticas para todas las condiciones sea verdadera, será inferior o superior al valor nominal de α. También a diferencia de lo que ocurre con el diseño intersujeto, el simple incremento del tamaño de la muestra no soluciona este problema [30]. Hay cada vez más pruebas de que las desviaciones de la simetría tienen efectos mucho más graves que las desviaciones de la distribución normal en términos de curtosis ([4]; [18]). El grado de desviación de la simetría puede expresarse en términos de *sesgo* de la distribución, que es el tercer momento normalizado [8]. Para una distribución simétrica, como la distribución normal, el sesgo es 0. La *curtosis* es el cuarto momento de población normalizado alrededor de la media y describe el peso de los valores máximo y mínimo (véase la ilustración en [9]). Anteriores estudios de simulación indican que, para pequeñas desviaciones de la simetría, la rmANOVA seguirá controlando la tasa de error de Tipo I. Sin embargo, por ahora no se pueden formular normas precisas sobre el grado aceptable de desviación de la normalidad, por lo que se recomienda probar la normalidad multivariada y consignar las estimaciones empíricas del sesgo y la curtosis.

Es importante señalar que el modelo lineal general subyacente a rmANOVA no asume que las respuestas brutas (es decir, la calificación en la prueba MUSHRA) están normalmente distribuidas. Al contrario, el modelo asume que los *errores* están normalmente distribuidos. Por este motivo, se han de aplicar las pruebas de normalidad o las medidas de sesgo y curtosis de los *valores residuales* del modelo, en lugar de las de los datos brutos. Por suerte, la mayoría de software estadísticos pueden salvar los valores residuales de cada condición experimental analizada, lo que, en este caso, es cada combinación de sistema de audio y material de audio. Se obtendrá entonces un vector de valores residuales para cada condición experimental. En cada vector, cada valor representa a un evaluador.

Existen varias pruebas de normalidad multivariada, como, por ejemplo, la prueba Shapiro-Wilk multivariada propuesta por Royston [34], las pruebas basadas en sesgo y curtosis multivariados [10] y otros [14]. Las macros para aplicar estas pruebas pueden obtenerse en SPSS (http://www.columbia.edu/~ld208/normtest.sps) y SAS (http://support.sas.com/kb/24/983.html), como probablemente también para otros software. Las estimaciones invariadas de sesgo y curtosis, que pueden calcularse por separado para los valores residuales de cada combinación de sistema de audio y material de audio, se encuentran en todos los principales software estadísticos. La macro SPSS de DeCarlo [9] (http://www.columbia.edu/~ld208/normtest.sps) también calcula el sesgo y la curtosis multivariados [26]. Se han de consignar las estimaciones de sesgo y curtosis invariados y multivariados, así como el resultado de la prueba de normalidad multivariada.

Si la prueba de normalidad multivariada no es significativa, o si ninguna de las pruebas multivariadas o invariadas muestra una desviación importante del sesgo o la curtosis con respecto a los valores previstos en una distribución normal, se cumplen los supuestos de rmANOVA.

No obstante, si cualquiera de esas pruebas indica una desviación importante con respecto a la normalidad, o si el sesgo de cualquiera de las condiciones experimentales supera un valor de 0,5 (como norma general preliminar), habrá que saber cuáles serán las consecuencias de esas conclusiones. Se presentan dos problemas generales y ambos están asociados a la ya comentada falta de normas sobre la desviación aceptable de la normalidad para rmANOVA. En primer lugar, las pruebas de normalidad multivariada son bastante sensibles y con frecuencia detectarán desviaciones de la normalidad de minutos. También detectarán no sólo una asimetría en la distribución de los valores residuales, sino también la curtosis o demás aspectos de la distribución que también influyen, mientras que hay bastantes probabilidades de resultados de asimetría en las tasas de error de Tipo I no robustas en rmANOVA. En segundo lugar, si se estiman las medidas del sesgo y la curtosis multivariados a partir de los datos [26], esta información no permitirá tomar una decisión sobre si se puede o no aplicar rmANOVA al no haber reglas relativas a la desviación aceptable de la normalidad. Por eso es aún más importante consignar las medidas de sesgo y curtosis, así como los resultados de la prueba. En cuanto se disponga de normas válidas sobre la desviación aceptable de la normalidad, podrán reevaluarse los resultados de las pruebas rmANOVA con esa información. Si la desviación de la normalidad parece importante, por ejemplo por estimaciones de sesgo superiores a 1,0 [29], podrán considerarse las alternativas no paramétricas a rmANOVA como, por ejemplo, las pruebas que utilizan técnicas de remuestreo o la prueba de Friedman. Sin embargo, aún no está claro en qué situaciones las técnicas de remuestreo solucionan el problema de la no normalidad [38]. La prueba de Friedman no asume la normalidad multivariada, pero sí que las varianzas son idénticas para todas las condiciones experimentales [36], lo que con frecuencia no será el caso de los datos experimentales. Además, la prueba de Friedman es una prueba invariada, por lo que, aunque se suponga que se cumple la igualdad de varianzas, la prueba de Friedman puede utilizarse para detectar un efecto del sistema de audio mediado en todo el material de audio, pero no se puede utilizar para analizar la interacción entre el sistema de audio y el material de audio.

3 Selección del método rmANOVA

Para los datos de diseños de repetición de mediciones hay muchos distintos métodos para probar los efectos de los factores intrasujeto e intersujeto [21]. Como ahora estamos considerando un diseño sin factores intersujeto (agrupación), y asumimos que no faltan datos (es decir, que se dispone de una calificación para cada combinación de oyente, material de audio y condición), se pueden recomendar dos métodos. Ambos ofrecen pruebas válidas de las hipótesis cuando los datos son normales multivariados, pero pueden diferir en cuanto a potencia estadística (es decir, sensibilidad para detectar una desviación de la hipótesis nula), en función del tamaño de la muestra, entre otros factores.

Las dos variantes de análisis son a) el *método invariado con corrección Huynh-Feldt para los grados de libertad*, y b) el *método multivariado*. Pueden encontrarse descripciones detalladas de estos métodos en [21]; [28]. Ambas variantes están disponibles en los principales software estadísticos (por ejemplo, R, SAS, SPSS, Statistica).

Habida cuenta de la estructura con repetición de mediciones de los datos, las calificaciones obtenidas en las diferentes combinaciones de condición y material de audio están correlacionadas. Por ejemplo, si un oyente asigna una calificación inusualmente elevada a un patrón de baja calidad, probablemente sus calificaciones de los codificadores también tenderán a ser superiores que las de otros evaluadores. El método invariado asume que la estructura varianza-covarianza de los datos es esférica, lo que equivale a decir que las variables de diferencia descritas a continuación tendrán todas la misma varianza [16]; [33]. Sin embargo, este supuesto se ve contradicho por prácticamente todos los conjuntos de datos empíricos [21]. Para solucionar el problema se aplica un factor de corrección a los grados de libertad cuando se calcula el valor *p* en función de la distribución *F*. Para ello, se estima a partir de los datos la magnitud de la desviación de la esfericidad. Se recomienda el factor de corrección Huynh-Feldt, denominado $\tilde{\varepsilon}$, porque el factor de corrección Greenhouse-Geisser [12] alternativo tiende a producir pruebas conservadoras (por ejemplo [17]; [30]). Cuando los datos son normales, el método invariado con la corrección Huynh-Feldt produce tasas de errores de Tipo I válidas hasta para muestras de muy pequeño tamaño ($N = 3$). Todos los principales software estadísticos ofrecen el factor de corrección, $\tilde{\varepsilon}$, y los valores *p* corregidos.

El *método invariado* utiliza una formulación alternativa, aunque equivalente, de la hipótesis nula. Por ejemplo, considérese la hipótesis nula de que en la población la calidad de audio percibida es idéntica en todas las condiciones. Esto equivale a calcular los contrastes ortogonales de $K - 1$, por ejemplo, formando variables de diferencia entre las K condiciones, y probando entonces la hipótesis de que el vector μ de las medias de población de todos los contrastes $K - 1$ es igual al vector nulo, $\mu = 0$. Por ejemplo, si se presentan la referencia y dos codificadores, pueden crearse las dos variables de diferencia, D_1 y D_2, calculando para cada evaluador la diferencia entre la calificación de la referencia y la calificación del codificador A (D_1), y la diferencia entre la calificación del codificador A y la calificación del codificador B (D_2). El rmANOVA que utiliza el método multivariado se basa en las variables de diferencia y utiliza una prueba multivariada de la hipótesis $\mu = 0$. Este método no necesita supuestos sobre la matriz varianza-covarianza. Para los datos que siguen una distribución normal multivariada, esta prueba es exacta, pero necesita, como mínimo, tantos evaluadores como número de niveles de factor. Así, no se puede utilizar si, por ejemplo, se presentan 9 condiciones (8 codificadores y la referencia) a sólo 8 evaluadores.

La potencia relativa de los dos métodos depende, entre otras muchas cosas, del tamaño de la muestra y del número de niveles de factor del factor intrasujeto. De acuerdo con Algina y Keselman (1997), una norma de selección simple sería utilizar el método invariado con la corrección Huynh-Feldt, si $\tilde{\varepsilon} > 0{,}85$ y $N < K + 30$, donde N es el número de evaluadores y K es el número máximo de niveles de factor intrasujeto. En los demás casos, debe utilizarse el método multivariado. Téngase en cuenta que, si el experimento se realiza en laboratorios distintos, N será el número total de evaluadores participantes en el estudio (por ejemplo, 10 evaluadores en el laboratorio A y 10 evaluadores en el laboratorio B hacen que $N = 20$).

4 Realización de la rmANOVA seleccionada y de las pruebas post hoc optativas

Llegados a este punto se realizan las pruebas globales de los efectos de las condiciones, el material de audio y su interacción utilizando la variante rmANOVA. Para calcular la rmANOVA, la mayoría de software, como SAS, SPSS y Statistica, exigen que los datos estén disponibles en formato «una fila por evaluador». Por tanto, la configuración del cuadro de datos será de una fila para cada evaluador y en las columnas («variables») se representarán las calificaciones de todas las combinaciones de condiciones y materiales de audio.

La rmANOVA bifactorial da información sobre tres efectos.

1) Principal efecto de la condición

En la mayoría de los casos esta prueba tendrá un gran interés. Si ANOVA indica un efecto de condición importante, podrá rechazarse la hipótesis nula de que en la población la calidad de audio percibida es idéntica en todas las condiciones (referencia, codificadores 1 a *k*). Dicho de otro modo, la prueba indica que en la población hay diferencias entre la calidad de audio percibida de los sistemas de audio. Como medida del tamaño de la magnitud del efecto, no es posible utilizar la *d* de Cohen [6] o cualquiera de sus análogos, porque *d* no se define para una comparación de más de dos medias. En el contexto de ANOVA, se suele utilizar una medida de la fuerza de asociación. Estas medidas dan información sobre la proporción de la varianza en los datos donde se observa el efecto en cuestión. Lo mismo puede decirse del coeficiente de determinación, R^2. La mayoría de software estadísticos pueden calcular el η^2 parcial, que se calcula como la relación de la varianza causada por el efecto a la suma de la varianza del efecto y la varianza del error (residual). Puede encontrarse más información sobre medidas alternativas de la fuerza de asociación en Olejnik y Algina [31].

Una vez obtenido un resultado significativo para un efecto principal, con frecuencia convendrá determinar el origen de ese efecto, lo que se puede hacer calculando contrastes específicos. Por ejemplo, se puede estar interesado en si la calidad de sonido de un nuevo codificador difiere de la calidad de sonido de tres sistemas tradicionales. Para responder a esta pregunta, en primer lugar hay que calcular la calificación media de los tres codificadores tradicionales otorgada por cada evaluador, en función de todo el material de audio. Así, por cada evaluador habrá a) una clasificación para el nuevo codificador, y b) una calificación media de los tres otros codificadores. A continuación se comparan esos dos valores con una prueba *t* de pares-muestra. Téngase en cuenta que, dado que los datos proceden de un diseño con repetición de medidas, es importante no utilizar la varianza combinada [27]. Téngase también en cuenta que este contraste también puede haberse probado como un contraste planificado, en lugar de con ANOVA. Suele recomendarse utilizar pruebas de significancia de dos colas. Sin embargo, si, por ejemplo, hubiese una hipótesis *a priori* de que el nuevo codificador debería recibir mejores calificaciones que los codificadores existentes, se podría utilizar una región de rechazo de una cola.

Se pueden calcular otros contrastes específicos con esta misma lógica. Hay una manera más general de probar los contrastes, que es calcular una combinación lineal de calificaciones obtenidas en las distintas condiciones del experimento y, posteriormente, utilizar una prueba *t* de una muestra para decidir si este contraste varía notablemente de 0. Para cada evaluador *i* se calcula un valor de contraste

$$\Psi_i = \sum_{j=1}^{a} c_j Y_{ij}, \quad \sum_{j=1}^{a} c_j = 0$$

donde Y_{ij} es la calificación otorgada por el evaluador *i* en la condición *j* (mediada en todo el material de audio), *a* es el número de condiciones consideradas en este contraste y c_j son los coeficientes. Para el ejemplo anterior, si el nuevo codificador corresponde a $j = 1$ y los tres otros codificadores a $j = 2 \ldots 4$, la selección de $c_1 = -1$ y $c_2 = c_3 = c_4 = 1/3$ dará una prueba de la hipótesis de que la calidad de audio del nuevo codificador difiere de las de los tres otros codificadores.

Si se calcula más de un contraste post hoc, como ya se ha dicho antes, se crea el problema de las pruebas múltiples. Para resolverlo se recomienda aplicar el procedimiento de Bonferroni con mejora de aceptación secuencial de Hochberg [15]. Este procedimiento controla la tasa de errores de Tipo I por familias y es, al mismo tiempo, más potente que muchos otros procedimientos alternativos [20]. En el procedimiento de Hochberg, en primer lugar se calculan los contrastes *m* de interés y se ordenan en función del valor *p*. A continuación se van examinando los valores *p* empezando por el más grande. Si este valor *p* es inferior a α, se rechazan todas las hipótesis (es decir, todos los

contrastes son significativos). En caso contrario, la prueba *t* con valores *p* más grandes no es significativa y se pasa a comparar el siguiente valor *p* con $\alpha/2$. Si es más pequeño, esta prueba y todas las pruebas con valores *p* más pequeños son significativas. En caso contrario, la prueba con el segundo valor *p* más grande no es significativa y se pasa a comparar el siguiente valor *p* con $\alpha/3$. En términos más formales, si p_i con $i = m, m - 1, \ldots, 1$ son los valores *p* en orden decreciente, para cualquier $i = m, m - 1, \ldots, 1$, si $p_i < \alpha/(m - i + 1)$, y todas las pruebas con $i' \leq i$ son significativas.

En principio, también es posible efectuar comparaciones por pares post hoc entre las calificaciones para todas las condiciones. En el caso del diseño con repetición de medición, se necesitarían para ello pruebas *t* por pares-muestra entre todos los pares de condiciones. Sin embargo, no se recomienda este método. Considérese un experimento con 7 codificadores y una referencia. Para este conjunto de 8 condiciones, se pueden calcular $8 \cdot 7/2 = 28$ pruebas por pares y no resultará fácil extraer información interesante de tantas pruebas. Si se prueban todas las diferencias por pares, a causa del elevado número de pruebas revestirá especial importancia la aplicación del procedimiento de Hochberg [15] para corregir las pruebas múltiples. Téngase en cuenta de que, si hay pruebas de desviación de la normalidad de las calificaciones de diferencia en que se basan las pruebas *t* por pares, la prueba alternativa que no asume la normalidad es la prueba de signo.

Cabe señalar que tras un efecto principal significativo puede que ninguno de los contrastes post hoc o de las diferencias por pares sea significativo [28] debido a la distinta información estadística que utilizan rmANOVA y las pruebas post hoc. Insistimos en que rmANOVA es la prueba más adecuada. Por consiguiente, un efecto significativo indicado por ANOVA sigue siendo válido aun cuando ninguna de las pruebas post hoc resulte significativa. Si tras una prueba global (ANOVA) significativa ningún contraste post hoc resulta significativo, puede concluirse que los sistemas de audio difieren en cuanto a calidad de sonido percibida. Las diferencias entre los sistemas de audio también pueden compararse unas con otras. Por ejemplo, para los pares de sistemas de audio que muestran la mayor diferencia en las calificaciones de calidad de sonido, probablemente esas diferencias por pares se volverían significativas si la muestra fuera de mayor tamaño. No obstante, debe concluirse que, en este estudio, ninguna de las diferencias por pares fue significativa.

Si la rmANOVA no muestra *ningún* efecto principal significativo de condición, quiere decir que las diferencias entre los sistemas que se prueban son pequeñas. Sin embargo, dado que la muestra es finita, no se puede concluir que en la población no hay *ninguna* diferencia en la calidad de audio percibida entre las condiciones [3]. Es posible que las diferencias de población sean cero o que la magnitud de los efectos haya sido demasiado pequeña para poder detectarla en una muestra de ese tamaño. Si se realiza un análisis de potencia *a priori*, es decir, si se selecciona una muestra de tamaño suficiente para detectar un efecto de tamaño determinado con una probabilidad determinada, puede concluirse que los datos son pruebas contra un efecto del tamaño predeterminado.

Esta conclusión puede tomarse como una definición de la transparencia de los codificadores. Si no se ha realizado un análisis de potencia *a priori*, se habrá de tener cuidado a la hora de inferir que los codificadores eran transparentes, por los motivos explicados antes. Una solución de aproximación post hoc habitual es comparar el valor *p* a [0,2] en vez de 0,05. Si la prueba sigue sin ser significativa, queda bastante más claro que no hay diferencias en la calidad de audio percibida de las condiciones.

2) Principal efecto del material de audio

Siguiendo la misma lógica y los mismos pasos de antes, la prueba del principal efecto del material de audio da información sobre los cambios sistemáticos de las calificaciones en función del material de prueba. En la mayoría de configuraciones de la prueba MUSHRA, este efecto no tendrá gran interés, porque no tiene relación con la diferencia entre los sistemas de audio.

3) Interacción entre la condición y el material de audio

Si rmANOVA muestra una interacción significativa entre la condición y el material de audio, el efecto del sistema de audio en la calidad de audio percibida diferirá de un material a otro. Por ejemplo, puede ser que la referencia y un codificador obtengan la misma calificación para una canción pop muy comprimida, donde los efectos perturbadores de la codificación quedan ocultos por los componentes de distorsión del material. Por otra parte, la calificación de la calidad de sonido del codificador puede ser inferior a la de la referencia en el caso de una grabación de alta gama dinámica de un gran concierto. Esta interacción suele ser de interés para una prueba MUSHRA, pues indica que la diferencia entre los sistemas de audio depende del material.

Tras aplicar una prueba global significativa del efecto de interacción, podrá analizarse más a fondo la naturaleza de la interacción con pruebas post hoc. Se suelen probar los denominados *principales efectos simples*, que pueden, por ejemplo, calcularse realizando un rmANOVA monofactorial con la condición de factor intrasujeto distinto para cada material de audio. Estos análisis mostrarán qué materiales de audio tienen un efecto significativo de condición. También en este caso debe utilizarse el procedimiento de Hochberg para corregir las pruebas múltiples.

Como en el caso anterior, todas las diferencias por pares entre las combinaciones de condición y material de audio pueden probarse, en principio, utilizando pruebas *t* por pares-muestras individuales y el procedimiento de Hochberg. No obstante, el número de comparaciones por pares será incluso mayor que en el caso de los efectos principales. Si, por ejemplo, se combinan 8 sistemas de audio con 4 materiales de prueba, habrá 24 combinaciones de sistemas de audio y materiales de prueba, lo que corresponde a $24 \cdot 23/2 = 276$ pruebas por pares. Evidentemente, no es éste el método recomendado.

5 Extensión a los diseños con variable intersujeto (agrupación)

Hasta ahora hemos considerado un diseño sin factores intersujeto. Ahora veremos qué análisis se han de realizar si la prueba se llevó a cabo con distintos grupos de evaluadores, por ejemplo en dos laboratorios, o con músicos y no músicos.

Si hay factores intersujeto, es de vital importancia saber si el número de evaluadores en todos los grupos es idéntico (diseño equilibrado) o distinto (diseño desequilibrado).

Diseño equilibrado. Si el número de evaluadores es idéntico para todos los niveles del factor intersujeto, o si el tamaño del grupo no difiere en más del 10%, nuevamente se pueden utilizar para realizar la rmANOVA el método invariado con la corrección Huynh-Feldt para los grados de libertad o el método multivariado [21]. El diseño contendrá ahora las condiciones y materiales de audio de los factores intrasujeto y, al menos, un factor intersujeto (por ejemplo, el laboratorio). Por consiguiente, la rmANOVA ofrecerá una prueba adicional de los efectos intersujeto, así como de las interacciones entre todos los efectos intrasujeto e intersujeto.

Por ejemplo, puede resultar que la condición × la interacción del laboratorio sea significativa, lo que implicaría que las diferencias en la calidad de audio percibida de los sistemas de audio difieren del laboratorio A al laboratorio B. Téngase en cuenta que se supone que se han presentado a todos los grupos exactamente las mismas combinaciones de condición y material de audio. Si, por ejemplo, en cada laboratorio se han presentado materiales de audio diferentes, no pueden utilizarse los métodos presentados aquí. En ese caso, se habrán de utilizar los denominados modelos de efectos aleatorios [28] que quedan fuera del alcance de este Adjunto.

Diseño desequilibrado. Si el tamaño de los grupos difiere en más de un 10%, por desgracia ni el método invariado ni el método multivariado arrojarán resultados válidos [21]. Por consiguiente, se recomienda vivamente prever grupos de idéntico tamaño para evitar este problema. Si los grupos tienen distinto tamaño, se recomienda efectuar dos procedimientos de análisis. El primero es la

Prueba de aproximación general mejorada (IGA) [1] y el segundo es una variante específica de un análisis de modelo mixto basado en la máxima probabilidad [23]. La prueba IGA está disponible como una macro de SAS. El análisis de modelo mixto puede realizarse, por ejemplo, en SAS PROC MIXED. Para este último análisis hay dos opciones importantes. En primer lugar, los grados de libertad se han de calcular con el método de [19], que en SAS se realiza activando la opción `ddfm=KR` en la declaración `model`. En segundo lugar, se ha de ajustar una estructura de covarianza no estructurada intersujeto heterogénea (UN-H) [23] utilizando las opciones type=UN group=groupingvar en la declaración repetida, siendo groupingvar el nombre de la variable que contiene la clasificación del grupo.

Referencias

[1] Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,(2),* 243-252.

[2] Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*(2), 208-218.

[3] Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485-485.

[4] Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879. doi: 10.3758/s13428-012-0306-x.

[5] Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892.

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

[7] Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). New York: Wiley.

[8] Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

[9] DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi: 10.1037//1082-989x.2.3.292.

[10] Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70,(s1),* 927-939. doi: 10.1111/j.1468-0084.2008.00537.x.

[11] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237.

[12] Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

[13] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315.

[14] Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods, 19*(10), 3595-3617. doi: 10.1080/03610929008830400.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

[16] Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*(332), 1582-1589.

[17] Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational aand Behavioral Statistics, 1*(1), 69-82. doi: http://dx..org/10.2307/1164736.

[18] Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*(3), 813-825. doi: 10.2307/2530060.

[19] Kenward M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

[20] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational and Behavioral Statistics, 19*(2), 127-162.

[21] Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54, (1),* 1-20.

[22] Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical & Statistical Psychology, 53,(2),* 175-191.

[23] Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64*(2), 224-242. doi: 10.1177/0013164403260196.

[24] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, N.C.: SAS Institute, Inc.

[25] Lix, L. M., Keselman, J. C., & Keselman H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi: 10.2307/1170654.

[26] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.2307/2334770.

[27] Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational and Behavioral Statistics, 5*(3), 269-287. doi: 10.3102/10769986005003269.

[28] Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

[29] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

[30] Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and non-normal data. *Behavior Research Methods, 45*(3), 792-812. doi: http:/dx.doi.org/10.3758/s13428-012-0281-2.

[31] Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi: 10.1037/1082-989x.8.4.434.

[32] Rasmussen, J. L. (1987). Parametric and Bootstrap Approaches to Repeated Measures Designs. *Behavior Research Methods Instruments & Computers, 19*(4), 357-360.

[33] Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*(2), 147-163.

[34] Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C, 32*(2), 121-133. doi: 10.2307/2347291.

[35] Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 6*(4), 147-151. doi: 10.1027/1614-2241/a000016.

[36] St. Laurent, R., & Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics-Simulation and Computation, 42*(7), 1596-1615. doi: 10.1080/03610918.2012.671874.

[37] Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

[38] Seco, G. V., Izquierdo, M. C., García, M. P. F., & Díez, F. J. H. (2006). A comparison of the bootstrap-F, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35-62.

[39] Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical & Statistical Psychology, 53*, 69-82.

Adjunto 5
al Anexo 1
(Informativo)

Requisitos para el comportamiento de patrón óptimo

A continuación se enumeran los descriptores clave para cuya captura óptima deberán diseñarse los patrones.

Un patrón óptimo:

1) producirá datos que no muestran cambios sustanciales en la ordenación relativa de los sistemas que se prueban cuando se compara con los datos obtenidos utilizando las especificaciones de patrón de la Recomendación UIT-R BS.1534;
2) estará asociado a las calificaciones de los oyentes, que utilizan una escala más amplia de calificaciones para los sistemas que se prueban cuando se comparan con los datos obtenidos utilizando las especificaciones de patrón de la Recomendación UIT-R BS.1534;
3) será percibido por los oyentes como mucho más semejante a los sistemas que se prueban que los patrones descritos en las especificaciones de la Recomendación UIT-R BS.1534, lo que, a su vez, puede hacer que el tiempo de evaluación del patrón sea más largo;
4) permitirá realizar una comparación sensible de los sistemas de prueba de gama media;
5) producirán diferencias de unos 20-30 puntos, entre las puntuaciones patrones de gama baja y de gama media;
6) producirán degradaciones de calidad en los patrones con dependencia limitada del contenido.