RECOMMANDATION UIT-R BS.1534

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

(Cuestión UIT-R 220/10)

(2001)

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 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;

f) 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;

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

recomienda

1 que se utilicen 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.

ANEXO 1

1 Introducción

Esta Recomendación describe un nuevo 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 [EBU, 2000a; Soulodre y Lavoie, 1999; EBU, 2000b].

Esta Recomendación incluye los siguientes puntos y Apéndice:

Punto 1: Introducción

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

Punto 3: Diseño experimental

Punto 4: Selección de sujetos

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

Apéndice 1: Instrucciones que han de darse a los sujetos

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 systemas 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 [Soulodre y Lavoie, 1999] 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 UIT-T, como las 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 sujetos 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 nuevo método 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 [EBU, 2000a; Soulodre y Lavoie, 1999; EBU, 2000b]. 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. Si los sistemas ensayados pueden mejorar la calidad subjetiva de una señal, deben utilizarse otros métodos de prueba.

3 Diseño experimental

En un dominio de interés científico se utilizan varias clases de estrategias de investigación para recopilar información fiable. Para la evaluación sujetiva 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 recapilación 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 resultades de la prueba de audición, sean minimizados. Por ejemplo, si la secuencia real de elementos de audio fuese idéntica para todos los sujetos en una prueba de audición, no se podría estar seguro de si las evaluaciones efectuadas por los sujetos 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 sujetos 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 sujetos. 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.

4 Selección de los sujetos

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 sujetos 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 sujetos

Aunque el método de prueba MUSHRA no está concebido para pequeñas degradaciones, sigue siendo aún recomendable el empleo de oyentes experimentados. 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.

En ocasiones hay motivos para introducir una técnica de rechazo ya sea anterior (preselección) o posterior (postselección) a la prueba real. En algunos casos pueden utilizarse ambos tipos de rechazo. En el caso que nos ocupa, el rechazo es un proceso en el que se dejan de lado todas las evaluaciones de un sujeto particular.

Todo tipo de técnica de rechazo, analizado y aplicado minuciosamente puede conducir a un resultado sesgado. Es por ello extremadamente importante que siempre que se realice una eliminación de datos, el informe de la prueba describa claramente el criterio aplicado.

4.1.1 Preselección de los sujetos

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 sujetiva 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.

Puede utilizarse el procedimiento de adiestramiento como instrumento de la preselección.

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.

4.1.2 Postselección de los sujetos

Los métodos de postselección pueden básicamente dividirse al menos en dos clases:

- una se basa en la capacidad de los sujetos para efectuar valoraciones congruentes repetidas;
- la otra se basa en las incongruencias de una apreciación individual comparándola con el resultado medio de todos los sujetos para un elemento determinado.

Se recomienda examinar la dispersión individual y la desviación respecto a la valoración media de todos los sujetos.

El objetivo de ello es obtener una evaluación justa de la calidad de los elementos de prueba.

Si unos pocos sujetos utilizan alguno de los extremos de la escala (Excelente, Mala) y la mayoría se centran en otro punto de ella, cabe apreciar que los primeros se apartan de la norma y puede rechazárseles.

Dado que lo que se prueba es la «calidad intermedia», un participante debe ser capaz de identificar muy fácilmente la versión codificada y, por tanto, dar una valoración que se encuentre dentro de la gama de la manifestada por la mayoría de los participantes. Es probable que los oyentes con valoraciones en el extremo superior de la escala sean menos críticos y probablemente los sujetos que sólo den valoraciones del extremo inferior de ella sean demasiados críticos. Rechazando estos participantes de los extremos, cabe esperar una evaluación más realista de la calidad.

Los métodos se utilizan principalmente para eliminar participantes que no pueden efectuar discriminaciones adecuadas. 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 participantes 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 participante individual, con lo que se reducirá considerablemente la necesidad de rechazar los datos de un participante.

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 sujetos 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 sujetos 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 participantes, 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 sujetos 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 sujetos 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 al menos un patrón oculto. Pueden utilizarse patrones adicionales bien definidos, tales como los que se describen en el § 5.1.

5.1 Descripción de las señales de prueba

La longitud de las secuencias no debe generalmente rebasar 20 s para evitar la fatiga de los participantes y reducir la duración total de la prueba de audición.

El grupo de señales procesadas consta de todas las señales de prueba y de al menos una señal adicional (patrón) que es una versión filtrada en paso bajo de la señal no procesada. La anchura de banda de esta señal adicional debe ser de 3,5 kHz. Dependiendo del contexto de la prueba, pueden utilizarse como opción patrones adicionales. Pueden utilizarse otros tipos de patrones que muestren clases de degradaciones similares a las de los sistemas en prueba. Estos tipos de degradaciones pueden ser:

– limitación de la anchura de banda a 7 kHz o 10 kHz;

– imagen estereofónica reducida;

– ruido adicional;

– desvanecimientos;

– pérdidas de paquetes;

– otras.

NOTA 1 – 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 participantes 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 Apéndice 1.

5.3 Presentación de los estímulos

El MUSHRA es un método de prueba doblemente ciega multiestímulo con referencia oculta y patrón o 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 [Soulodre y Lavoie, 1999].

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 experimentador 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 sujeto 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 sujetos tengan grandes dificultades para discriminar de forma precisa entre las diversas señales degradadas. Por ejemplo, en una comparación directa por pares, los participantes 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, 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 instantáneamente entre 11 señales (1 referencia + 8 degradadas + 1 referencia oculta + 1 patrón 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 participantes obtendrán su valoración de un sistema determinado comparando dicho sistema con la señal de referencia, así como con las otras señales de cada tentativa.

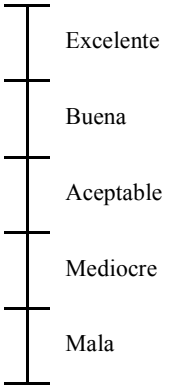
Se recomienda no incluir en cada tentativa más de 15 señales (por ejemplo, 12 sistemas en prueba, 1 referencia conocida, 1 patrón oculto y 1 referencia oculta).

En una prueba de la Recomendación UIT-R BS.1116, los sujetos tendían a abordar un ensayo determinado empezando con un proceso de detección, al que seguía un proceso de valoración. La experiencia en la realización de ensayos según el método MUSHRA muestra que los sujetos 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.

5.4 Proceso de valoración

Se pide a los sujetos 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.

FIGURA 1

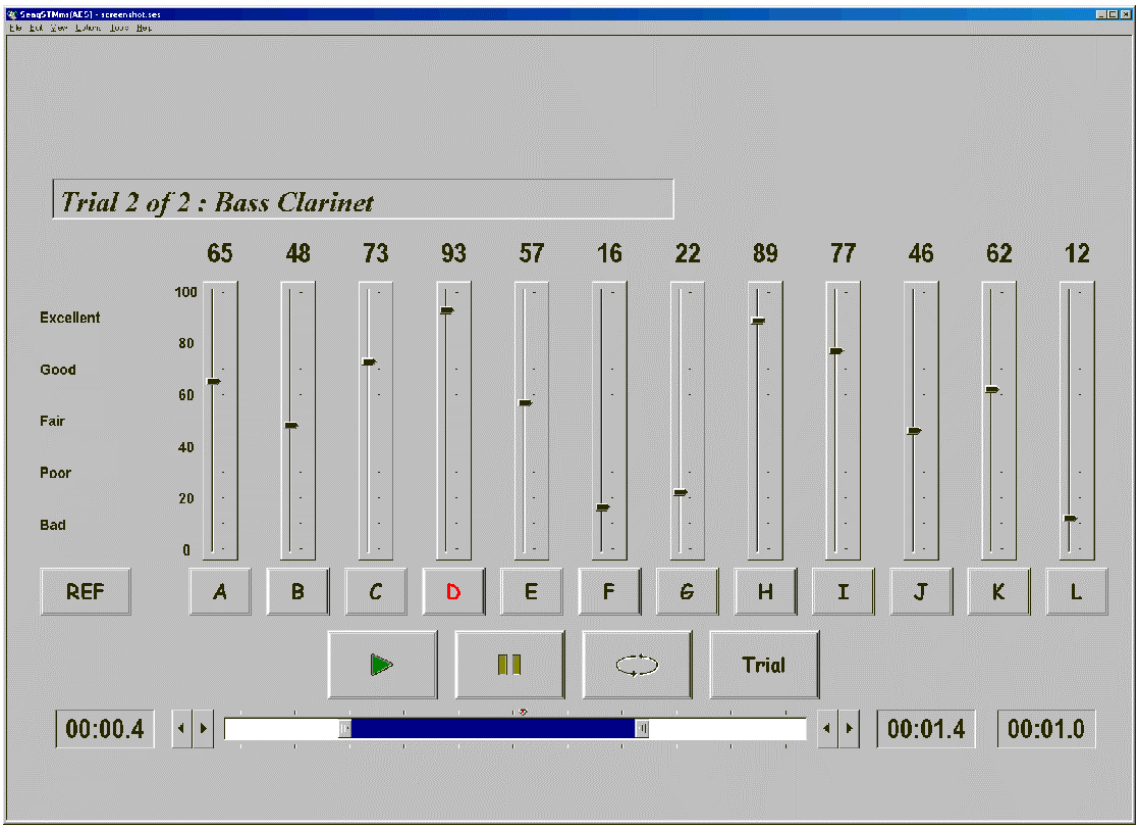


Esta escala se utiliza también para la valoración de la calidad de la imagen (véase la Recomendación UIT-R BT.500 – Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión).

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. Se pide al participante que evalúe 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 todos los estímulos al mismo tiempo, de forma que el sujeto puede verificar cualquier comparación entre ellos directamente. Los resultados son más congruentes, lo que da lugar a intervalos de confianza más pequeños. 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.

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 sujetos 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.

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.

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 sujetos 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 sujetos 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 sujetos 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 R 68 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. Los pasajes de audio tendrán generalmente

una longitud de 10 s a 20 s. Dada la complejidad de la tarea, debe disponerse de los sistemas sometidos a prueba. Sólo puede efectuarse una selección adecuada si se define un programa de tiempos apropiado.

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):

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

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

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 utilizarán 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 sujetos 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 R 68 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:

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

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 sujetos 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.

El cálculo de los promedios de las valoraciones normalizadas de todos los oyentes restantes después de la postselección serán las notas medias subjetivas.

El primer paso del análisis de los resultados es el cálculo de la nota media, $\overline{u}_{jk}$ para cada una de las presentaciones:

$$\overline{u}_{jk} = \frac{1}{N}\sum_{i=1}^{N} u_{ijk} \tag{1}$$

donde:

u_i: nota del observador i para una condición de prueba j y secuencia de audio k determinadas

N: número de observadores.

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

Al presentar los resultados de una prueba, todas las otras medias tienen que tener un intervalo de confianza asociado que se obtiene a partir de la desviación típica y del tamaño de cada muestra.

Se propone utilizar un intervalo de confianza del 95% que viene dado por:

$$\left[\overline{u}_{jk} - \delta_{jk}, \overline{u}_{jk} + \delta_{jk}\right]$$

donde:

$$\delta_{jk} = t_{0,05}\,\frac{S_{jk}}{\sqrt{N}} \tag{2}$$

y $t_{0,05}$ es el valor de t para un nivel de significación del 95%.

La desviación típica para cada presentación, S_{jk}, viene dada por:

$$S_{jk} = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_{jk} - u_{ijk})^2}{(N-1)}} \tag{3}$$

Con una probabilidad del 95%, el valor absoluto de la diferencia entre la nota media experimental y la nota media verdadera (para un número muy elevado de observadores) es inferior al intervalo de confianza del 95%, a condición de que la distribución de las notas individuales satisfaga ciertos requisitos.

De forma similar, puede calcularse una desviación típica, S_j para cada condición de prueba. No obstante, se señala que esta desviación típica, en los casos en que se utiliza un número pequeño de secuencias de prueba, estará influida por las diferencias entre las secuencias de prueba empleadas más que por las variaciones entre los asesores que participen en la evaluación.

La experiencia ha demostrado que las notas obtenidas por las distintas secuencias de prueba dependen del grado crítico del material de prueba utilizado. Puede obtenerse una comprensión más completa del comportamiento del sistema presentando por separado los resultados de las distintas secuencias de prueba, más que agregando únicamente los promedios de todas las secuencias utilizadas en la evaluación.

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 Apéndices.

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, tal como sobre los motivos subyacentes del estudio, los métodos del diseño experimental y la ejecución, y el análisis y las conclusiones.

Debe prestarse especial atención a lo siguiente:

- una presentación gráfica de los resultados;
- la especificación y selección de los sujetos (véase la Nota 1);
- 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 2);
- 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 base detallada de todas las conclusiones obtenidas.

NOTA 1 – Se ha demostrado que las variaciones del nivel de aptitud del grupo de oyentes puede influir en los resultados de las evaluaciones de audición. Para facilitar más el estudio de este factor, se pide a los experimentadores que informen lo más posible sobre las características de sus grupos de oyentes. Los factores pertinentes pueden ser los de edad y género en la composición del grupo.

NOTA 2 – 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 (por ejemplo, utilizando ANOVA).

10.3 Presentación de los resultados

Para cada parámetro en prueba debe indicarse la media y el intervalo de confianza del 95% 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 asesores;

– nota media total de todas las unidades de prueba utilizadas en el experimento;

– únicamente las notas medias y el intervalo de confianza del 95% tras la postselección de los observadores, es decir tras eliminar los resultados conforme al procedimiento que figura en el § 4.1.2 (postselección).

Además, hay que presentar también los resultados en forma adecuada, tal como la de histogramas, medianas u otras.

10.4 Notas absolutas

Una presentación de las notas medias absolutas para el sistema en pruebas, de la referencia oculta y del anclaje 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 de estas notas absolutas no conduce a una información significativa y no se debe utilizar ésta como tal.

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 Bibliográficas

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

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

SOULODRE, G. A. y LAVOIE, M. C. [septiembre de 1999] Subjective evaluation of large and small impairments in audio codecs, AES 17th International Conference, Florence, p. 329-336.

APÉNDICE 1

AL ANEXO 1

Instrucciones que han de darse a los sujetos

A continuación figura un ejemplo del tipo de instrucciones que deben darse o leerse a los sujetos 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.

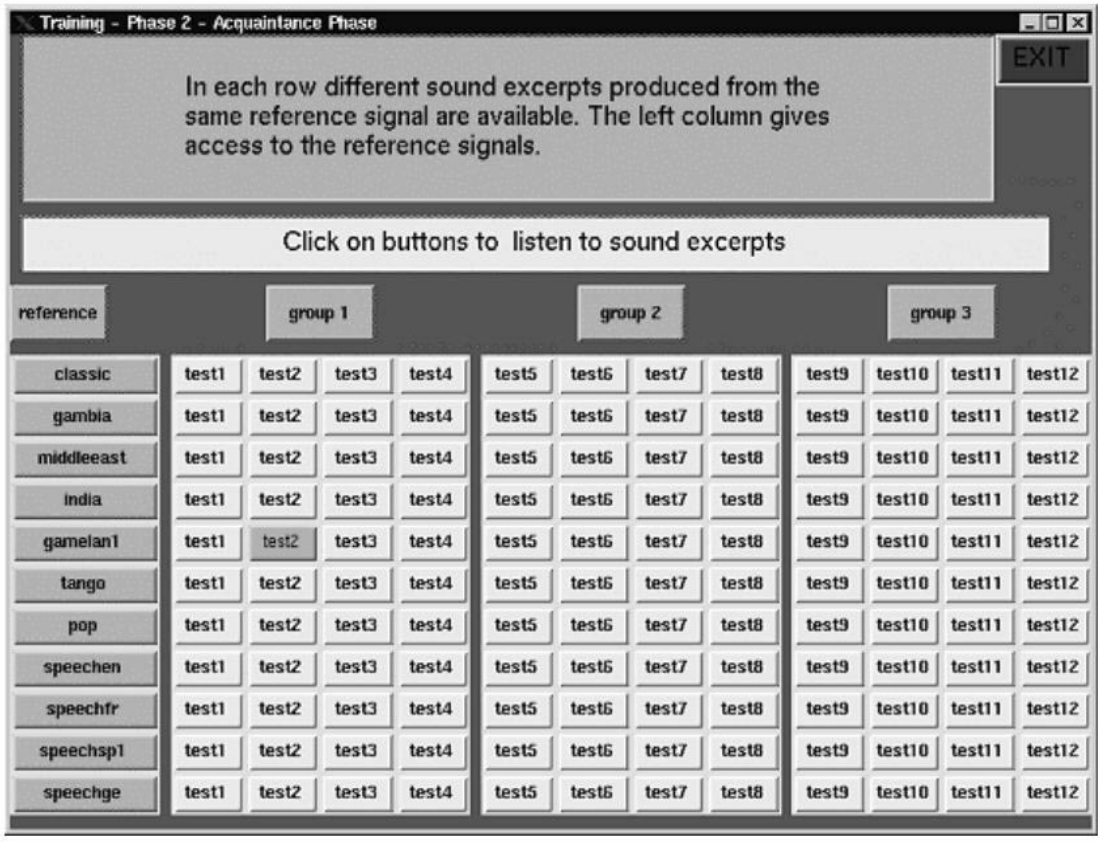
En la Parte B de la fase de adiestramiento se aprenderá a utilizar la reproducción disponible y el equipo de valoración que se empleará para evaluar la calidad de los pasajes sonoros.

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 11 señales que han de evaluarse. Cada uno de los elementos dura aproximadamente 10 a 20 s. Se debe escuchar la referencia 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



1534-03

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

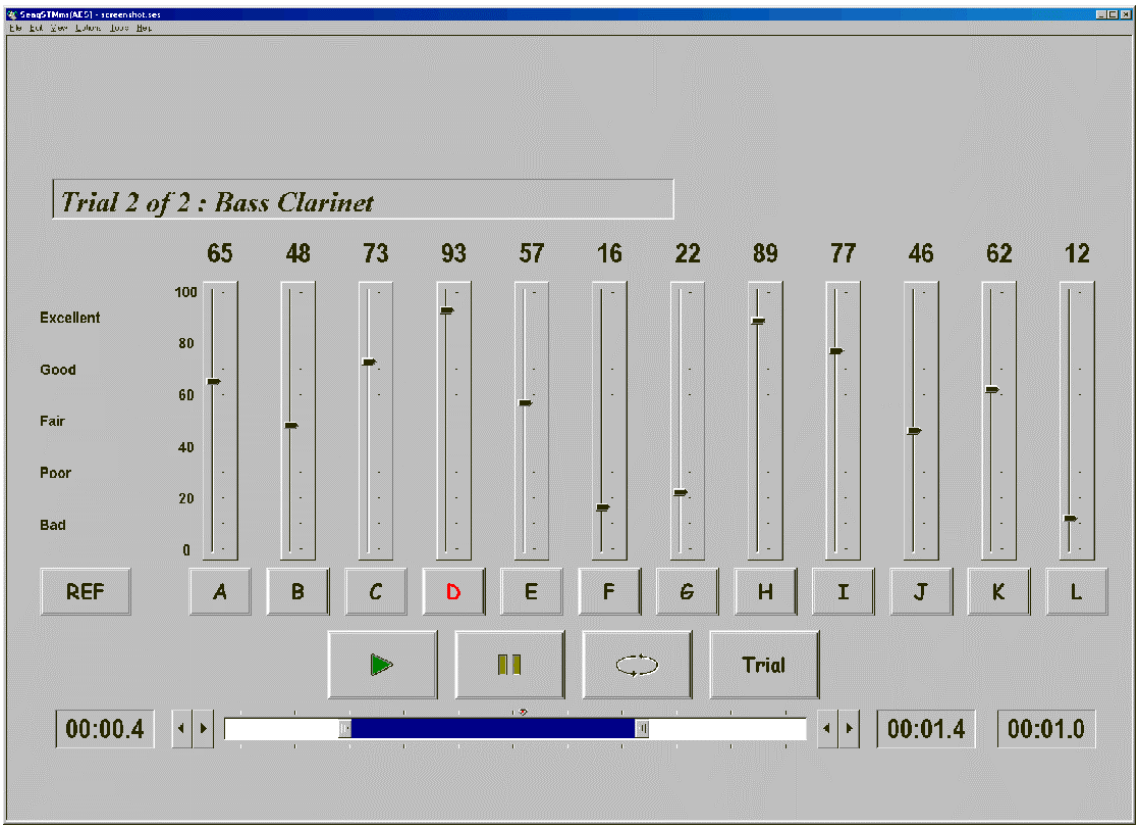
La escala de valoraciones es continua de «Excelente» a «Mala». Una nota 0 corresponde al mínimo de la categoría «Mala» mientras que una nota de 100 corresponde al máximo de la categoría «Excelente».

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.

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 con los otros sujetos en ningún momento durante la fase de adiestramiento.

FIGURA 4

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



1534-04

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