Unión Internacional de Telecomunicaciones

Procedimientos prácticos para pruebas subjetivas

Manual del UIT-T

2011

EL SECTOR DE NORMALIZACIÓN DE LAS TELECOMUNICACIONES DE LA UIT

La Unión Internacional de Telecomunicaciones (UIT) es el organismo especializado de las Naciones Unidas en el campo de las telecomunicaciones y de las tecnologías de la información y la comunicación (TIC). El Sector de Normalización de las Telecomunicaciones de la UIT (UIT-T) es un órgano permanente de la UIT. Este órgano estudia los aspectos técnicos, de explotación y tarifarios, y publica Recomendaciones sobre los mismos con miras a la normalización de las telecomunicaciones en el plano mundial. Aunque las Recomendaciones del UIT-T no son vinculantes, su utilización está muy extendida porque garantizan la interconectividad e interoperabilidad de las redes y permiten la prestación de servicios de telecomunicación a escala mundial.

La Asamblea Mundial de Normalización de las Telecomunicaciones (AMNT) y el Grupo Asesor de Normalización de las Telecomunicaciones (GANT), con la colaboración de las Comisiones de Estudio y sus Grupos de Trabajo, cumplen las funciones reglamentarias y políticas del UIT-T.

Procedimientos prácticos para pruebas subjetivas

Manual del UIT-T
2011

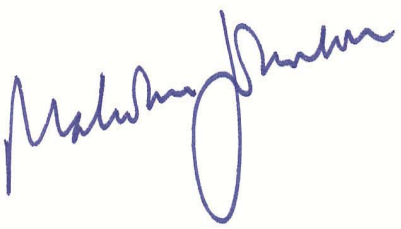
Prefacio

Malcolm Johnson
Director
Oficina de Normalización de las Telecomunicaciones

Hace más de cien años que se viene reconociendo que la calidad audio de los sistemas de comunicaciones debe evaluarse subjetivamente para que dicha evaluación tenga sentido. Si bien existen aplicaciones de audio conocidas que permiten predecir con exactitud los juicios de los usuarios sobre la base de mediciones objetivas, hay muchas otras aplicaciones –por ejemplo, los nuevos sistemas de tratamiento de las señales vocales– cuya calidad sólo puede evaluarse de una manera significativa con pruebas subjetivas.

Sin embargo, no basta con comprender la necesidad de realizar pruebas subjetivas para evaluar las nuevas tecnologías. Puesto que los seres humanos intervienen en dichas pruebas subjetivas, un aspecto esencial de las pruebas es la metodología aplicada. En última instancia, las conclusiones que se extraigan de los datos de las pruebas subjetivas dependerán de cierto número de factores, con inclusión del material vocal utilizado, la tarea asignada, la escala de calificación y el entorno acústico. Por otro lado, la prueba debe diseñarse con miras a permitir un análisis estadístico de los datos compilados. Por consiguiente, es importante que expertos diseñen la prueba subjetiva, así como su planificación y ejecución, y que éstos se encarguen del análisis de sus datos.

Así pues, es sumamente conveniente que la elaboración de un manual completo sobre pruebas subjetivas de la calidad audio quede en manos de un grupo de expertos de nivel mundial, y por lo tanto me complace que los expertos de la Comisión de Estudio 12 del UIT-T se hayan encargado de elaborar dicho manual. Este "Manual sobre procedimientos prácticos para pruebas subjetivas" proporciona la orientación que tanto se necesitaba. El contenido ha sido concebido para ayudar a los expertos en calidad vocal y audio a instalar laboratorios de pruebas subjetivas equipados con los instrumentos y dispositivos necesarios para llevar a cabo pruebas subjetivas armonizadas. La finalidad es generar datos de pruebas subjetivas capaces de guiar todas las actividades, desde la evaluación (y la eventual normalización) de los nuevos sistemas de tratamiento de la palabra, hasta la planificación de los servicios de comunicación de extremo a extremo.



Malcolm Johnson

Director
Oficina de Normalización de las Telecomunicaciones

ÍNDICE

Sección 1

Generalidades

1.0 Generalidades

En el presente Manual se describen los procedimientos prácticos que se han de utilizar para cuantificar la calidad audio percibida de los sistemas de comunicaciones vocales y audio de extremo a extremo.

Dichos procedimientos se pueden utilizar, por ejemplo, para:

- diseñar y llevar a la práctica pruebas subjetivas con el fin de calificar, seleccionar, caracterizar y/o verificar la calidad de funcionamiento de los códecs vocales y demás equipos de procesamiento de señales de red; o
- proporcionar información con fines de planificación y para mecanismos de evaluación objetiva.

Por el momento los procedimientos de prueba subjetiva audiovisual, según se describen en las Recomendaciones UIT-T de la serie P.9xx, están fuera del alcance de este Manual.

El presente Manual también contiene información para ayudar a los expertos en calidad vocal y audio a instalar laboratorios de pruebas subjetivas equipados con los instrumentos y dispositivos adecuados para llevar a cabo experimentos subjetivos. El objetivo es obtener resultados apreciables, mediante las pruebas de escucha y/o conversacionales, para clasificar los algoritmos de codificación vocal existentes y/o nuevos y/o determinar la calidad de extremo a extremo de los sistemas de comunicación completa, desde el punto de vista de la calidad y/o la degradación percibidas subjetivamente.

1.1 Comentarios sobre los métodos subjetivos utilizados en laboratorio

La elección del método de prueba subjetiva más adecuado entre las pruebas normalizadas disponibles influye de manera fundamental en los resultados generales de un experimento. Esa elección puede ser responsabilidad de un experimentador (por ejemplo, con fines internos dentro de una organización) o bien decidirse por consenso, si el ejercicio tiene su origen en un comité de normas. En todo caso, el profesionalismo y la experiencia del o de los diseñadores de la prueba desempeñan una función fundamental en este aspecto delicado pero esencial de una prueba en laboratorio, puesto que hay diferentes maneras de evaluar la calidad percibida de una comunicación, y cada método tiene sus propias capacidades y peculiaridades. Dependiendo del objetivo de la prueba, cada método subjetivo debe seguir las directrices estipuladas en la correspondiente recomendación de prueba, con inclusión de: la elaboración de materiales de prueba específicos, siguiendo instrucciones detalladas para la presentación de estímulos vocales a los sujetos, la adopción de unas escalas de calificación precisas y, lo que es más importante, la utilización de fórmulas e instrumentos adecuados para el análisis de los datos, ya que esto último es de importancia capital para obtener resultados válidos y repetibles dentro de una gama predecible.

1.2 Utilización de los resultados de los experimentos

La realización de una prueba subjetiva de la calidad vocal o audio en el laboratorio o en el terreno puede tener diversos objetivos, a saber:

a) calificar/seleccionar un códec vocal (para una Recomendación) entre diferentes candidatos;

b) caracterizar la calidad de funcionamiento y verificar si se cumplen los requisitos de uno (o más) códec/s seleccionados para una Recomendación;

c) comprobar el impacto de los diferentes parámetros o los efectos (por ejemplo, atenuación, retardo, eco, ruido, errores de transmisión, pérdidas de paquetes, etc.) en la calidad de la llamada;

d) detectar y remediar problemas o degradaciones concretos de la conexión (sobre la base de la calidad percibida).

1.3 Recomendaciones UIT-T

Se dispone de algunas Recomendaciones UIT-T de las series G y P sobre sistemas y medios de transmisión, sistemas y redes digitales para evaluar la calidad percibida de los medios audio transportados por redes RTPC, RMTP o IP. Revisten particular importancia a tales efectos las Recomendaciones UIT-T de las series P.800 y G.700, así como las Recomendaciones UIT-T G.107, 109, 114, 131 y 191.

1.4 La continua necesidad de pruebas subjetivas

La calidad percibida de una telecomunicación varía según las degradaciones introducidas durante el transporte por la red (diferencia entre la señal original y la señal recibida), así como con las experiencias acumuladas por los particulares. Diferentes tipos de redes introducen diferentes degradaciones; por ejemplo, las llamadas a través de la RTPC o redes móviles o de paquetes pueden alcanzar diferentes niveles de calidad. La telefonía de paquetes integra el tráfico de voz y de datos, y permite prestar nuevos servicios multimedios. Entonces las expectativas del usuario van evolucionando continuamente, y un usuario puede ser más tolerante a causa de las ventajas de la movilidad en comparación con la telefonía tradicional cableada, o debido a la disponibilidad de servicios multimedios. Puesto que las reacciones subjetivas del usuario cambian según los diferentes contextos, los juicios subjetivos de un grupo de escuchas puede variar considerablemente con el correr de tiempo.

De los ejemplos y consideraciones que anteceden se desprende claramente que es necesario realizar continuamente pruebas subjetivas para mantenerse al corriente de las opiniones y percepciones de los usuarios del mundo real, así como de sus preferencias y expectativas cambiantes respecto de los nuevos servicios multimedios.

1.5 Pruebas en laboratorio frente a pruebas en el terreno

Las ventajas fundamentales de las pruebas en laboratorio son la definición y el control exactos de las variables experimentales (condiciones específicas de ensayo) y la posibilidad de repetir los resultados obtenidos con el experimento. Además, una vez que se han acordado y terminado el diseño y el plan de prueba, se puede estimar "a priori" con bastante exactitud el costo del ejercicio de prueba.

Las ventajas de las pruebas en el terreno consisten en no excluir "a priori" ningún aspecto de un servicio (o comportamiento de algoritmo), es decir que se someten a prueba las condiciones del mundo real, lo que debería conducir a una evaluación realista de la calidad de funcionamiento del códec o la red, a condición de que se investiguen todos los casos posibles y se efectúe y repita un número suficiente de pruebas. Las desventajas de estos ejercicios en el mundo real es la posibilidad de repetición menos segura del experimento, debido a las variables no controladas o los errores imprevistos en el ejercicio de comunicación.

Esos errores dan lugar a datos que podrían no ser totalmente seguros y a resultados que exigen retrasos y un mayor volumen de gastos para la nueva prueba. En resumen, las pruebas en tiempo real en el terreno consumen más tiempo que las pruebas en laboratorio y son más exigentes en lo tocante al costo y al establecimiento de la prueba, pero ofrecen la ventaja de reproducir y someter a prueba escenarios del mundo real.

1.6 Métodos recomendados

En las Recomendaciones que se indican a continuación están consignados los principales métodos utilizados actualmente en las pruebas subjetivas:

- **UIT-T P.800**

Esta es una Recomendación fundamental para la realización de pruebas subjetivas de la calidad de transmisión.

En esta Recomendación se describen métodos y procedimientos para realizar evaluaciones subjetivas de la calidad de transmisión, tales como el procedimiento de evaluación de los índices por categorías absolutas (*Absolute Category Rating*, ACR), evaluación de los índices por categorías de degradación (*Degradation Category Rating*, DCR), evaluación de los índices por categorías de comparación (*Comparison Category Rating*, CCR), las comparaciones apareadas y las pruebas conversacionales.

- **UIT-T P.805**

En esta Recomendación se describen métodos y procedimientos para realizar pruebas conversacionales con el fin de evaluar la calidad de la comunicación. En la metodología se utilizan ejemplos de hipótesis, escalas de puntuación y procedimientos de análisis para estimar la calidad subjetiva de los servicios de telecomunicaciones interactivos o bidireccionales. Las pruebas conversacionales permiten simular situaciones más realistas próximas a las condiciones de servicio reales experimentadas por los usuarios de telecomunicaciones. Además, las pruebas conversacionales están diseñadas para evaluar los efectos de las degradaciones que pueden causar dificultades durante la conversación (tales como retardo, pérdida de paquetes, eco, interrupciones, ruido, recortes vocales, etc.) y pueden utilizarse para estudiar los efectos generales del sistema o las degradaciones específicas.

- **UIT-T P.830**

En esta Recomendación se describen métodos y procedimientos para realizar evaluaciones subjetivas de la calidad de funcionamiento de los códecs vocales digitales. Basada en el anexo B a UIT-T P.800, en esta Recomendación se define un método de prueba específico para evaluar los procesos digitales, de modo que los efectos de distorsión de cuantificación de estos procesos en la calidad de funcionamiento de la transmisión puedan tenerse en cuenta en la red internacional en evolución.

- **UIT-T P.832**

En esta Recomendación se describen procedimientos para evaluar la calidad de funcionamiento subjetiva de los terminales manos libres. Los métodos que se definen pueden utilizarse para evaluar el grado en el cual un terminal manos libres funciona eficazmente para conversación.

- **UIT-T P.835**

En esta Recomendación se describen procedimientos para evaluar la calidad de funcionamiento subjetiva de los algoritmos, incluido el algoritmo de supresión de ruido (*Noise Suppression Algorithms*, NSA). Con un método de puntuación de una sola escala, como el ACR, cada sujeto individual pondera la señal y los

componentes del fondo para determinar su puntuación de la calidad vocal general. Este proceso de ponderación introduce una varianza del error adicional en las puntuaciones subjetivas de la calidad global, lo que las hace menos fiables. La metodología que se describe en esta Recomendación reduce la incertidumbre del oyente dado que a éste se le exige que preste atención y puntúe por separado cada componente de la forma de onda: la *señal de voz,* el *ruido de fondo,* y el *efecto global: voz + fondo.*

Esta metodología ha demostrado ser fiable y válida para la evaluación de los NSA, y debería utilizarse en otras aplicaciones. Podría utilizarse siempre que se desee evaluar la voz en presencia de ruido de fondo. Su aplicación es especialmente adecuada cuando no se sabe si el sistema incluye un preprocesador de ruido.

- **UIT-T P.840**

El objeto de la presente Recomendación es describir un método para las pruebas de escucha subjetiva que pueda utilizarse para evaluar la calidad de la señal vocal en los equipos de multiplicación de circuitos (*circuit multiplication equipment*, CME). Esta Recomendación está destinada a utilización con sistemas CME, como los descritos en las Recomendaciones UIT-T G.763, 767, 768 (DCME), 765 (PCME) y 769/Y.1242 (IP-CME), que utilizan técnicas de interpolación digital de la palabra (*digital speech interpolation*, DSI). En las evaluaciones de pruebas subjetivas de sistemas CME que no utilizan técnicas de interpolación digital de la palabra.

- **UIT-T P.851**

En esta Recomendación se describen métodos y procedimientos para llevar a cabo experimentos de evaluación subjetiva de servicios telefónicos que emplean sistemas conversacionales. Los sistemas correspondientes permiten una interacción natural empleando lenguaje oral y disponen de capacidades de reconocimiento e interpretación vocal, gestión del diálogo y emisión vocal. Se describe la configuración y la realización de los experimentos de interacción adecuados, y se proponen cuestionarios para cuantificar los parámetros de calidad pertinentes que percibe el usuario.

- **Suplemento 24 a las Recomendaciones UIT-T de la serie P**

En el presente Suplemento se describen parámetros que facilitan información sobre la interacción con servicios basados en sistemas de diálogo oral, desde el punto de vista del ingeniero de sistema y del operador del servicio. Los parámetros describen la calidad de funcionamiento del sistema desde la perspectiva del ingeniero de sistemas; por esa razón, contienen información complementaria a la que se obtiene mediante los experimentos de evaluación subjetiva de los sistemas de diálogo oral, que se especifican en la Recomendación UIT-T P.851.

- **UIT-T P.880**

En esta Recomendación se describe la metodología denominada evaluación continua de la calidad vocal que varía con el tiempo (*Continuous Evaluation of Time Varying Speech Quality*, CETVSQ) que sirve para evaluar los efectos de las fluctuaciones temporales de la calidad vocal en la calidad instantánea percibida (que se percibe en cualquier instante de una secuencia vocal) y de la calidad percibida general (al final de la secuencia vocal). El método consta de dos partes: en primer lugar, la valoración instantánea que se regula mediante un cursor mientras que se escucha la secuencia vocal en una escala continua y, en segundo lugar, la valoración global en una escala de cinco categorías convencionales al final de la secuencia vocal.

- **UIT-T BS.1116-1**

La idea es servirse de esta Recomendación para evaluar sistemas que introducen degradaciones tan pequeñas que no son detectables si no se controlan rigurosamente las condiciones experimentales y se efectúa el correspondiente análisis estadístico. La aplicación de esta Recomendación a los sistemas que introducen

degradaciones relativamente importantes y fácilmente detectables, conduciría a un derroche de tiempo y de esfuerzos y podría producir, por otra parte, resultados menos fiables que una prueba más simple.

- **UIT-T BS.1534**

La finalidad de esta Recomendación es evaluar la calidad audio intermedia y su método, comúnmente conocido como "Ensayo multiestímulo con referencia y patrón ocultos" (*Multi stimulus test with hidden reference and anchor*, MUSHRA), está destinado a aportar una medición fiable y repetible de los sistemas con calidad audio que caerían en la mitad inferior de la escala de degradación utilizada en UIT-T BS.1116-1.

- **UIT-T BS.1285**

Esta Recomendación está basada en la Recomendación UIT-T BS.1116. Es una variación de la misma que da lugar a una presentación a múltiples sujetos en una sesión de prueba, y difiere básicamente de ésta por el contenido de su punto referente a los métodos de prueba. Para no efectuar dichas pruebas cuando ello no se requiere, convendría disponer de una metodología de preselección con arreglo a la cual se pudieran rechazar fiablemente los sistemas que introducen degradaciones considerables. Aunque esta reducción del ámbito del método de prueba puede aminorar su sensibilidad, es necesario que dicho método permita seguir distinguiendo los sistemas que introducen grandes degradaciones de aquellos que no la introducen.

- **Otros**

Entre los demás métodos de prueba subjetivos cabe incluir las pruebas de habla simultánea y las pruebas de escucha de tercera parte, tal como están consignadas en las Recomendaciones UIT-T P.83 y UIT-T P.832.

En los ejercicios de prueba subjetiva se utilizan algunas Recomendaciones (UIT-T de la serie G: Sistemas y medios de transmisión, sistemas y redes digitales) como condiciones del sistema de referencia. La base de datos *Media Coding Summary Database* que mantiene la CE 16 del UIT-T contiene una lista completa de las mismas
(http://www.itu.int/en/ITU-T/studygroups/com16/multimedia/Pages/mcsd.aspx).

Para este Manual sobre procedimientos prácticos para pruebas subjetivas se han utilizado algunas partes del Manual sobre Telefonometría (1992), Ginebra, ISBN 92-61-04911-7, en particular:

Secciones 1.1 – *Definiciones y unidades* y 1.3 – *Conceptos básicos de estadística*, que siguen siendo aplicables para la definición de términos y unidades de medida y para los métodos estadísticos utilizados en el análisis de los resultados de las pruebas subjetivas.

Sección 2 – *Pruebas subjetivas*, que contiene gran parte de la filosofía sobre la que se apoya la utilización actual de estos métodos

Anexo B (al § 2.4) – *Atenuación equivalente para la nitidez (AEN)*:

2.5 *Pruebas de opinión*, que contiene detalles que aún se aplican a las pruebas de escucha únicamente y a las pruebas conversacionales.

2.6 *Otros métodos de pruebas subjetivas*, que contiene detalles que aún se aplican a las pruebas de comparaciones pares.

También revisten interés las siguientes secciones de la publicación del UIT-T *"Adiciones al Manual de Telefonometría"* (1999):

- Sección 2.3.6 – *Algunas repercusiones del efecto local*, indica la importancia de considerar al tono lateral como un factor en las pruebas subjetivas, tanto para pruebas conversacionales como al preparar los materiales básicos para las pruebas de escucha únicamente.

- Sección 3.2.5 – *Voz artificial de conversación*, no es pertinente para las pruebas subjetivas, con la posible excepción de la aplicabilidad de las distribuciones estadísticas del estado conversacional.
- Sección 4.8 – *Modelos de cálculo para la estimación de la opinión de los clientes sobre la calidad de las comunicaciones de conversación en las redes telefónicas*, contiene un excelente examen del concepto "opinión del cliente" en relación con las pruebas subjetivas y las encuestas de clientes.

El documento actual trata de ser coherente con estas secciones del Manual sobre Telefonometría y de enriquecer dichas secciones con la experiencia adquirida en la realización de pruebas subjetivas desde su publicación.

1.6.1 Pruebas de opinión conversacionales

Se recomienda hacer pruebas de opinión conversacionales en todos los ejercicios que entrañan interacción entre los comunicadores, como por ejemplo: interrupciones vocales, retraso, eco, borrado de trama, pérdida de paquetes, etc. Dado que las pruebas conversacionales son más exigentes en cuanto a la organización y el costo, a menudo han sido sustituidas por pruebas de opinión únicamente de escucha, que se consideraron adecuadas, por ejemplo, para calificar/seleccionar cierto número de algoritmos de codificación vocal (véanse las Recomendaciones UIT-T de la serie G.700 y la mayoría de los códecs vocales aprobados por las organizaciones normativas (*Standards Development Organizations*, SDO).

Se está propagando de manera generalizada el acceso IP en banda ancha utilizando líneas de abonado digital (DSL) y fibra óptica, y se prevé que en un futuro éste se difundirá aún más. A medida que la banda ancha empiece a ser algo común, las comunicaciones vocales se utilizarán más frecuentemente en las comunicaciones persona a persona. Según las previsiones, el sector de las comunicaciones manos libres marchará a la vanguardia en lo tocante a las comunicaciones vocales de alta calidad en banda ancha. Conforme a estas pautas, se prevé que las pruebas de opinión conversacional serán cada vez más pertinentes para evaluar correctamente la forma según la cual los usuarios perciben la calidad.

1.6.2 Pruebas de opinión de escucha

Tradicionalmente, las pruebas de opinión de escucha han sido las más utilizadas. Ello obedece a la simplicidad y eficacia del método para comparar y evaluar la calidad de funcionamiento sobre la base de la calidad percibida de los algoritmos de codificación del habla y/o sistemas de comunicación vocal más complejos. El razonamiento subyacente para el uso de las pruebas de opinión de escucha únicamente, en vez de las conversacionales, siempre ha sido que, con fines de selección/comparación, un algoritmo o un sistema más eficaz que otro confirmaría ese ordenamiento de la puntuación en ambos tipos de prueba. Por consiguiente, la UIT y los órganos normativos regionales se mostraron partidarios de utilizar pruebas de opinión de escucha para los ejercicios de calificación, selección, caracterización y/o verificación.

Actualmente en algunos laboratorios especializados del mundo se dispone de procedimientos automatizados para preparar materiales vocales, realizar experimentos y obtener resultados fiables de manera oportuna.

1.6.3 Pruebas de entrevista y encuesta

A menudo los operadores y los fabricantes recurren a las pruebas de entrevista y encuesta para evaluar el grado de aceptación de nuevos productos y servicios por el público en general. Es esencial que la muestra de los oyentes sea representativa de la población de usuarios de telecomunicaciones para los cuales se proyectan los resultados. En esta categoría de pruebas los principales aspectos son: cómo se formulan las preguntas y a quiénes se elije para responderlas. A veces estas pruebas se utilizan como complemento de las pruebas de escucha y conversacionales, y pueden resultar útiles para interpretar correctamente los resultados de esas pruebas.

1.6.4 Otros métodos de prueba subjetiva

Se podría necesitar otros métodos de prueba subjetiva (por ejemplo, pruebas de inteligibilidad, versiones modificadas de pruebas normalizadas, pruebas de aplicación específica, etc.) para evaluar la calidad de funcionamiento de los sistemas de comunicación audio, pero éstos no se abordan en el presente Manual.

El Manual sobre Telefonometría y ciertas recomendaciones de otros órganos normativos (por ejemplo, ANSI S3.2 (2009), *Method for measuring intelligibility of speech over communication systems*) contienen descripciones de las pruebas de inteligibilidad y de otros métodos subjetivos tradicionales.

Sección 2

Objetivos de la prueba

2.0 Introducción

En los siguientes párrafos se describe el proceso aplicado por la UIT durante una actividad de normalización códec típica. Hay otros métodos para realizar investigaciones, actividades de desarrollo e implementación de sistemas de comunicación vocal de extremo a extremo, pero en general éstos son una variación del método del UIT-T y pueden deducirse de las siguientes descripciones de los objetivos de la prueba.

A tenor de la política actual del UIT-T, los resultados de las fases de prueba descritas en los puntos 2.1-2.4 se incluyen en un informe técnico.

2.1 Calificación

El objetivo de una prueba de calificación es demostrar la calidad de funcionamiento adecuada de un codificador en un subconjunto de condiciones requeridas que representan la aplicación para la cual se diseña el codificador.

Por lo general la fase de calificación tiene lugar al comienzo de una competencia para preseleccionar n códecs candidatos entre las m > n propuestas recibidas por un comité normativo. En la fase de selección, que por lo general viene después de la fase de calificación, sólo se admiten los n candidatos que "prometen cumplir con los requisitos de calidad de funcionamiento". Las ventajas de este procedimiento son las siguientes:

a) permitir que el comité normativo reduzca el número de candidatos a un número razonable o "fijo con antelación" de algoritmos que vale la pena incluir en una fase de selección;

b) permitir que los proponentes candidatos extraigan sus propias conclusiones sobre las posibilidades que tienen de "ganar la competencia";

c) permitir que los diseñadores de la prueba de calificación elaboren un plan de prueba reducido y razonable, centrándose en los requisitos esenciales que debe cumplir un candidato para no ser descartado de la competencia. Puesto que en cada prueba de calificación interviene un único códec candidato, eso deja lugar para ensayar en esta fase un gran número de condiciones experimentales diferentes, pero que siguen siendo pertinentes.

En general las pruebas de calificación las realizan los proponentes del códec, ya sea en sus propias instalaciones de prueba calificadas o contratando a un laboratorio de prueba externo calificado. Por lo tanto, es esencial que los métodos y procedimientos para llevar a cabo las pruebas estén definidos con precisión en el plan de prueba, con miras a reducir el margen de discreción a que puede dar lugar este tipo de pruebas "internas". Dependiendo del consenso del comité normativo, se puede otorgar un breve periodo para mejorar los puntos débiles de cada códec candidato antes de iniciar la fase de selección. En esta última fase se incluirán a todos los candidatos que se considere que han "pasado" la fase de calificación.

2.2 Selección

En general el objetivo de la prueba de selección es elegir un solo códec del grupo de candidatos "calificados". La elección se basa en la calidad de funcionamiento en un conjunto de condiciones que representan los requisitos de la aplicación para la cual se consideran los códecs.

En el UIT-T la fase de selección siempre es un ejercicio coordinado a nivel internacional, en el que participan laboratorios de pruebas subjetivas muy calificados. Además, se designan laboratorios huéspedes para procesar el material vocal original seleccionado y elaborar la "base de datos" de pruebas de escucha. Esta base de datos se transmite a los laboratorios de escucha designados para realizar las pruebas subjetivas. Uno o más laboratorios huéspedes pueden compartir la responsabilidad de procesar las muestras vocales o audio proporcionadas por los diferentes laboratorios de escucha que intervienen en las pruebas de selección. La ventaja de contar por lo menos con dos laboratorios huéspedes es que un laboratorio hace una verificación cruzada del procesamiento efectuado por el otro y suministra sus resultados. Este procedimiento, que puede ser realizado incluso por organizaciones independientes, asegura que no se incurra en errores, o por lo menos reduce al mínimo el número de errores posibles. Este proceso es vital para evitar la necesidad de realizar pruebas subjetivas costosas y que consumen mucho tiempo pese a ser en última instancia inútiles, pues producen resultados incorrectos o que inducen a error.

A menudo el proceso de selección incluye la designación de uno o más laboratorios de análisis globales (*Global Analysis Laboratories*, GAL) que tienen la responsabilidad de compilar los datos brutos (los "puntos") de los diferentes laboratorios de escucha y proporcionar los resultados para todo el ejercicio. Normalmente esos resultados incluyen una "puntuación de opinión media", intervalos de confianza, gráficos, cuadros, análisis de varianzas y otros datos estadísticos. La finalidad del análisis global es proporcionar conclusiones basadas en los resultados procedentes de todos los diversos laboratorios de escucha que participan en la prueba de selección.

Por último, el comité normativo responsable de la fase de selección revisa las conclusiones del informe del GAL y recomienda, sobre la base de unos criterios de selección previamente acordados, al "ganador de la competición" (a condición, por supuesto, de que se haya llegado a un consenso sobre un candidato que cumple con todos los requisitos).

Es indispensable que el comité normativo responsable de la evaluación de la calidad de funcionamiento y la verificación de los requisitos de calidad de funcionamiento sea independiente del comité que establece estos últimos requisitos para un nuevo algoritmo de codificación. Dicha independencia de juicio (así como diferentes ámbitos de competencia) asegura que la recomendación final no esté influida por ningún posible conflicto de interés. El único objetivo es cumplir con el mandato de recomendar como candidato ganador al "mejor códec".

2.3 Caracterización

La finalidad de la caracterización es demostrar la calidad de funcionamiento de un solo códec en un conjunto de condiciones y modalidades de utilización para las cuales ha sido diseñado el códec.

Por lo general las pruebas de caracterización son realizadas en una versión acordada de un códec vocal elegido en una prueba de selección. Cierto número de laboratorios de escucha realizan una serie exhaustiva de pruebas subjetivas de escucha en diferentes idiomas (tanto en idiomas tonales como no tonales) y en diferentes condiciones, aplicaciones o modalidades de utilización simuladas. En los ejercicios en los cuales se designa un GAL, los laboratorios de escucha deben entregar los datos brutos obtenidos de las pruebas de escucha que realizaron a la organización responsable del análisis global de los resultados. Según proceda, los laboratorios huéspedes y los laboratorios de escucha también pueden proporcionar su propio informe y análisis para dar cumplimiento a sus acuerdos contractuales.

Las pruebas de caracterización exploran la serie completa de hipótesis de aplicaciones para el códec que es objeto de prueba, y los resultados forman el contenido básico de una publicación permanente del comité normativo que ha desarrollado el nuevo algoritmo de codificación.

Los resultados contenidos en el informe de la fase de caracterización revisten una importancia fundamental para la planificación, pues permiten a los operadores de red tomar decisiones sobre las compensaciones entre la calidad y los aspectos relacionados con la capacidad de la red.

El informe de caracterización también puede incluir resultados de la fase de verificación.

2.4 Verificación

El objetivo de la prueba de verificación es asegurar el cumplimiento de una implementación de codificador o una variante con su normalización. Esto se logra sometiendo a prueba al codificador en una serie de condiciones y midiendo su conformidad mediante la comparación de los resultados con la norma.

Para completar la normalización se puede hacer participar a algunos voluntarios en la fase de verificación, mediante la presentación de contribuciones que sirven de base para el informe de verificación. La fase de verificación puede consistir en la solución de posibles problemas y la medición, mediante pruebas subjetivas, de la calidad de funcionamiento frente a otros aspectos concretos (por ejemplo, dependencia lingüística), o por medio de pruebas basadas en herramientas objetivas (por ejemplo, calidad de funcionamiento con tonos DTMF, o con señales de entrada especiales), evaluación de la complejidad, consumo de memoria, retardo, comportamiento en condiciones de canal en reposo, respuesta de frecuencia, etc.

2.5 Investigación

A menudo las pruebas subjetivas se realizan con fines de investigación. En tales casos, la organización que investiga un aspecto concreto (empresa pública o privada, operador de telecomunicaciones, universidad o centro de investigaciones científicas) financia las actividades tendientes a organizar y realizar la prueba, así como el análisis de sus resultados. Se propone que, siempre que sea posible, se adopten procedimientos normalizados, para evitar desviaciones apreciables con respecto a resultados bien conocidos y a los resultados obtenidos con ejercicios similares realizados en el pasado. Una excepción es el caso en el cual el objetivo de la prueba consiste específicamente en estudiar nuevos métodos de pruebas subjetivas o revisar/ampliar los métodos existentes. Se propone asimismo que los resultados de esa investigación se transmitan a las organizaciones normativas para la recomendación de nuevas normas y/o la mejora de los procedimientos existentes. En el pasado se han logrado muchos avances gracias a iniciativas y contribuciones que aportaron las instituciones de investigación a la comunidad internacional.

2.6 Otros

En algunas disciplinas, como por ejemplo la psicoacústica, las relacionadas con factores humanos, la ergonomía y la medicina, se proponen con carácter facultativo u obligatorio algunos procedimientos de pruebas subjetivas un poco similares, para atender toda una amplia gama de diversos intereses, pero todos ellos reflejan el deseo de comprender mejor el comportamiento humano y/o la reacción de los usuarios cuando se enfrentan a problemas de calidad. Este Manual está limitado a los procedimientos de pruebas subjetivas que apuntan a evaluar la calidad de funcionamiento de los sistemas de telecomunicaciones.

Sección 3

Diseños experimentales

3.0 Introducción

Los programas de pruebas subjetivas sirven para uno o varios de los fines descritos en la Sección 2. La selección de un diseño experimental en concreto depende de los objetivos del experimento o los experimentos que componen el programa de pruebas. Los experimentos deben diseñarse con arreglo a los objetivos y fines del programa de pruebas subjetivas y deben organizarse de modo que los resultados obtenidos permitan alcanzar dichos objetivos. La correcta utilización de elementos y principios de diseño garantiza que los resultados de los experimentos subjetivos resulten fiables y válidos para los objetivos y fines del programa de pruebas.

3.1 Etapas para diseñar una prueba subjetiva

Las tres etapas para garantizar que el diseño de una prueba subjetiva o un programa de pruebas sea adecuado son las siguientes:

- determinar las preguntas a las que se pretende dar respuesta con los resultados de la prueba;
- especificar los métodos y procedimientos estadísticos para responder a esas preguntas; y
- diseñar la prueba subjetiva o el programa de pruebas.

3.2 Elementos y principios de diseño

Los elementos de diseño de experimentos subjetivos son:

- la metodología de pruebas;
- el número de condiciones de prueba (es decir, los códecs de referencia y de prueba, las condiciones de referencia, etc.);
- la base de datos de voz de origen (es decir, idioma, número de hablantes, número de muestras de voz/hablantes, etc.);
- el número de ensayos (es decir, número de hablantes x número de condiciones de prueba); y
- el número de grupos de oyentes y el número de oyentes/grupo.

Los principios de diseño son las reglas y convenios utilizados para diseñar los experimentos que componen el programa de pruebas con el fin de controlar las diversas fuentes de variabilidad. Junto con los elementos de diseño, estas reglas determinan la estructura de los experimentos subjetivos-el diseño experimental.

3.2.1 Fuentes de variabilidad (idiomas, hablantes, material vocal, sujetos,…)

En las pruebas subjetivas, cada respuesta de un oyente es la suma de varias fuentes de variabilidad. Estas fuentes pueden clasificarse de manera sistemática o aleatoria. El objetivo del diseño experimental es identificar y eliminar todas las fuentes de variación sistemáticas, salvo las correspondientes a los factores objeto del experimento. Esas fuentes de variabilidad comprenden todos los factores que dan lugar a una variación de las respuestas obtenidas en pruebas de escucha subjetivas, en particular:

- *factores de prueba* –efectos predeterminados– factores para los que se ha diseñado el experimento (por ejemplo, códecs, tasa de errores, nivel de entrada, ruidos de fondo);
- *factores de muestreo* –efectos aleatorios– factores que permiten la evaluación estadística de los factores de prueba (por ejemplo, sujetos, hablantes, muestras de voz); y
- *factores extrínsecos* –efectos predeterminados o aleatorios– factores que introducen una variabilidad no deseada y/o inexplicable (por ejemplo, efectos de tiempo/orden, incertidumbre del sujeto, efectos contextuales).

En las pruebas subjetivas, cada respuesta de un sujeto es una combinación de las fuentes de variabilidad antes descritas. El objetivo del diseño experimental es maximizar la variabilidad en las puntuaciones debido a los factores de prueba, minimizar la variabilidad debida a los factores de muestreo, y eliminar o controlar la variabilidad debida a factores extrínsecos.

3.2.2 Factores – efectos aleatorios o predeterminados

Los experimentos se diseñan normalmente para evaluar un número de factores de prueba específicos (por ejemplo, códecs, condiciones de ruido, hablantes, sujetos, etc.). Cada factor de prueba consiste en dos o más niveles o valores que se seleccionan como parte del proceso de diseño experimental –ese factor se denomina *hecho predeterminado*. La varianza debida a los niveles del factor se denomina *efecto predeterminado*. Sin embargo, si los niveles de un factor se *muestrean* en una población de individuos, entonces el factor se denomina *factor aleatorio* y la varianza entre cada nivel se denomina *efecto aleatorio* o *de error*. Ejemplos típicos de *efectos predeterminados* son códecs, tasa de errores en las tramas, ruido de fondo, retardo, y supresión del ruido "activada" o "desactivada". Ejemplos típicos de *efectos de error* son los sujetos y las mediciones repetidas.

En la Sección 6 figuran ejemplos de métodos y procedimientos específicos utilizados por el UIT-T para controlar los factores experimentales en las pruebas subjetivas.

3.3 Tipos de diseños de pruebas

El diseño experimental puede clasificarse en dos niveles. El primer nivel comprende el diseño general o global del experimento y consiste en la determinación de los factores que se han de analizar y las hipótesis que se desea verificar. El segundo nivel comprende los elementos más locales que permiten al diseñador de la prueba aplicar y realizar realmente un experimento válido y fiable. El *diseño factorial* es un ejemplo de diseño experimental global mientras que los *bloques equilibrados/aleatorizados* es un ejemplo de diseño local para llevar a cabo el experimento.

3.3.1 Diseño factorial

El diseño experimental más popular para pruebas subjetivas es el diseño factorial. Permite al experimentador diseñar una prueba para evaluar uno o más efectos predeterminados y resulta adecuado en las técnicas estadísticas destinadas a evaluar dichos efectos. Por ejemplo, el diseño factorial es especialmente adecuado para experimentos en los que el objetivo es evaluar la calidad de funcionamiento de múltiples códecs en una gama de condiciones de prueba. En el cuadro 3.1 se muestra un ejemplo de diseño experimental factorial con tres factores. Dos de los factores, *Códecs (2)* y *Tasa de errores en las tramas (2)*, conllevan efectos predeterminados, y el otro factor, *Oyentes (n=10)*, entraña un efecto aleatorio o de error.

Cuadro 3.1 – Diseño factorial con tres factores

Códecs (2)	Tasa de errores en las tramas (FER) (2)	Oyentes (10)
Ref	5%	L_1
Prueba	10%	L_2
		…
		L_{10}

Un adecuado diseño experimental permitiría al experimentador responder a las siguientes preguntas:

- ¿Qué códec recibe el mayor índice?
- ¿Se debe la diferencia en el índice a un efecto sistemático o a un error de muestreo aleatorio, es decir, la diferencia es estadísticamente significativa?
- ¿La tasa de errores en las tramas afecta considerablemente a los índices?
- ¿Se produjo una interacción del códec y la tasa de errores en las tramas, es decir, la FER tiene un efecto diferente en los índices correspondientes a los dos códecs?

El diseño simplificado que se ilustra en el cuadro 3.1 implicaría la recopilación de respuestas para cuatro ensayos de prueba (2 *códecs* x 2 *FER*) para cada uno de los *n* oyentes. En el cuadro 3.2 se muestra un orden de presentación de muestra para cuatro ensayos.

Cuadro 3.2 – Orden de presentación de ensayos

Ensayo	Códec	Tasa de errores en las tramas (FER)	Oyentes
1	Ref	5%	L_1,,...., L_{10}
2	Prueba	10%	L_1,,...., L_{10}
3	Ref	10%	L_1,,...., $L10$
4	Prueba	5%	L_1,,...., L_{10}

Para garantizar la fiabilidad y validez de los análisis estadísticos que se utilizan habitualmente en tales experimentos (por ejemplo, análisis de la varianza, pruebas t, pruebas de gama múltiple post hoc), debemos garantizar que los experimentos se diseñan para eliminar, o al menos controlar, las diversas fuentes de error experimental. Estos errores experimentales son inherentes a todas las pruebas subjetivas –es precisamente este *error* el que nos permite realizar comparaciones y análisis estadísticos. El objetivo del diseñador del experimento es garantizar que el error experimental represente las fluctuaciones aleatorias en los índices asignados por el sujeto sobre los que se basan estos análisis estadísticos, y que nuestras estimaciones de este error no adolezcan de efectos sistemáticos que pudieran distorsionar o invalidar los resultados de esos análisis.

Desde hace ya bastante tiempo se sabe que las respuestas de un sujeto a pruebas individuales no son independientes, sino que se ven influenciadas por diversos factores que pueden afectar a la reacción del sujeto a los estímulos de esa prueba. Algunos de estos factores son extrínsecos a la respuesta propiamente dicha, aunque, no obstante, contribuyen al error experimental global. Dos de las fuentes primarias de estos errores son las asociadas al tiempo y orden de presentación, que a menudo recibe el nombre de *efectos de*

orden y tiempo. Cada respuesta del individuo se ve influenciada no solamente por el estímulo presentado en ese ensayo particular, sino también por los estímulos experimentados en pruebas anteriores, es decir, se trata de efectos ligados al *orden de presentación*. Por otra parte, los estímulos que se experimentan más tarde en un conjunto de presentaciones serán más sensibles a los efectos de la fatiga y el aburrimiento que los experimentados al principio del conjunto, es decir, son efectos ligados al *tiempo de presentación*. Al diseñar conjuntos de presentación aleatorios, podemos controlar el error experimental relacionado con estos *efectos de orden y tiempo*.

Un experimento subjetivo característico organizado en el UIT-T consistiría en una serie de condiciones de prueba, varios hablantes, y un número de muestras de voz por cada voz. Las condiciones de prueba se eligen para responder a las preguntas que constituyen la esencia del experimento. Se utilizan múltiples hablantes de ambos sexos para que los resultados sean aplicables a una amplia gama de personas. También se utilizan varias muestras de voz para que los resultados sean aplicables, en general, al idioma utilizado en la prueba. El número total de muestras de voz que se han de evaluar puede llegar a ser muy elevado, a saber, número total de muestras = #Condiciones × #Hablantes × #Muestras por voz. Por ejemplo, en un experimento subjetivo característico que conste de 36 condiciones, 6 hablantes y 4 muestras por voz habría que evaluar 864 muestras de prueba. Una prueba subjetiva de este tamaño resultaría demasiado larga para que un solo sujeto evaluara todas las muestras de voz (es decir, 864 en este ejemplo). Se ha de realizar un diseño experimental que permita repartir las numerosas muestras entre sujetos o grupos de sujetos, sin que ello menoscabe la integridad del análisis estadístico y teniendo presente los *efectos de orden y tiempo*. Este diseño, es decir el diseño experimental de bloques equilibrados, se describe en la siguiente sección.

3.3.2 Diseño experimental de bloques equilibrados

El diseño experimental de bloques equilibrados es una de las clases más generales de diseños factoriales que se utiliza ampliamente en las pruebas subjetivas para actividades de normalización.

A los efectos del análisis estadístico de datos de pruebas subjetivas, el experimento ideal consistiría en hacer pruebas con todos los oyentes y en todas las condiciones de prueba, utilizando todo el corpus de material vocal para cada condición de prueba. Sin embargo, en la práctica, las limitaciones del tiempo necesario para cada oyente restringe el número de ensayos por oyente. Por consiguiente, los experimentos deben diseñarse de modo que las muestras se repartan entre los oyentes sin menoscabar la validez de las comparaciones estadísticas. La solución común a este problema es separar los oyentes en grupos de modo que se puntúen todo el corpus de material vocal, aunque cada grupo sólo evalúe un subconjunto del corpus. Asimismo, cada grupo de oyentes escucha un subconjunto equivalente, sin ser idéntico, de dicho material. En cada condición de prueba, grupos separados de oyentes valorarán cada hablante utilizando muestras diferentes. Una vez determinado el número de grupos de oyentes, el diseño experimental de bloques equilibrados exige seis muestras por voz.

Por consiguiente, el primer principio de diseño de pruebas experimentales de bloques equilibrados es que el número de muestras por oyente debe ser igual al número de grupos de oyentes.

En la terminología estadística de las técnicas de análisis de la varianza (*analysis of variance*, ANOVA), la división de *hablantes* x *corpus de material de muestra* resulta en una confesión de los efectos de *materiales vocales* con los efectos de interacción de *condiciones de prueba x Hablantes x Grupos de oyentes*. Dado que no cabe esperar que el efecto de interacción específico sea significativo en un ANOVA (y no se sabría cómo interpretarlo si así fuera), esta confesión es una concesión que estamos dispuestos a aceptar para garantizar la integridad y validez de los efectos estadísticos, que representa el interés primordial del experimento –por ejemplo, el principal efecto de *Condiciones* y del efecto interacción para *Condiciones x Hablantes*.

Según los principios antes mencionados, los factores o *principales efectos* en un ANOVA serán: *Condiciones*, *Hablantes*, *Material vocal* y *Oyentes*. El efecto principal de *Oyentes* consta, en realidad, de dos componentes, *Grupos* y *Oyentes del grupo*. Ahora bien, si aceptamos el principio de confesión antes mencionado, podemos ignorar o mancomunar estos componentes en un solo efecto *Oyentes*, y simplificar el ANOVA para analizar los efectos de *Condiciones*, *Hablantes* y *Oyentes*. Así se obtiene un ANOVA resumido en el cuadro de recursos de varianza que se ilustra en el cuadro 3.3.

En el cuadro 3.3, C = número de Condiciones, T = número de Hablantes y L = número de Oyentes. Los dos efectos de mayor interés son *Condiciones* y *Condiciones x Hablantes*. Las relaciones F (es decir, relaciones de varianzas o *valores cuadráticos medios*[1]) utilizadas para comprobar estos efectos serán:

$$F_{\text{Condiciones}} = MS_{\text{Condiciones}} / MS_{\text{Condiciones x Oyentes}}$$

$$F_{\text{Condiciones x Hablantes}} = MS_{\text{Condiciones x Hablantes}} / MS_{\text{Condiciones x Oyentes x Hablantes}}$$

Cuadro 3.3 – Cuadro de fuentes de varianza simplificado para el análisis de la varianza de los efectos de Condiciones, Hablantes y Oyentes

Fuente de variación [efecto]	Grados de libertad	Término de error
Condiciones [predeterminado]	C-1	Condiciones x Oyentes
Hablantes [predeterminado]	T-1	Hablantes x Oyentes
Oyentes [aleatorio]	L-1	- - -
Condiciones x Hablantes [predeterminado]	(C-1)(T-1)	Condiciones x Hablantes x Oyentes
Condiciones x Oyentes [aleatorio]	(C-1)(L-1)	- - -
Hablantes x Oyentes [aleatorio]	(T-1)(L-1)	- - -
Condiciones x Hablantes x Oyentes [aleatorio]	(C-1)(T-1)(L-1)	- - -
Total	(C x T x L)-1	

Por otra parte, dependiendo del diseño experimental, los efectos de *Condiciones* pueden dividirse para permitir pruebas de factores adicionales integradas en el diseño factorial, por ejemplo códigos posibles, nivel de entrada, tasa de errores, etc.

Una vez más, si el *corpus* del material vocal para un experimento es el conjunto total de las muestras vocales de todos los hablantes, el experimento se considera *equilibrado* si todos los grupos de oyentes se exponen al *corpus* vocal un número idéntico de veces. Por lo tanto, el segundo principio de diseño es que en un experimento de bloques equilibrados, el número de condiciones de prueba totales (N) debe ser igual a un múltiplo entero del número de hablantes y a un múltiplo entero del número de muestras por voz. Con estas restricciones, cada grupo de oyentes puntuará el corpus del material el mismo número de veces, aunque dicho corpus se distribuya de manera diferente entre las condiciones de prueba.

[1] En el ANOVA, el término cuadrático medio se refiere a una varianza de la población basada en la variabilidad de un determinado conjunto de mediciones.

En la práctica, se ha identificado un número de principios de diseño experimentales para ayudar a controlar los *efectos de tiempo y orden*, entre los que cabe citar los siguientes:

- El experimento se organiza en bloques de ensayos en los que el número de bloques suele ser igual al número de hablantes.
- Cada bloque contiene una muestra de cada una de las N condiciones de prueba previstas en el experimento y el orden de dichas muestras se aleatoriza, con restricciones, dentro del bloque.
- Una de esas restricciones es que las muestras sucesivas (es decir, ensayos) de un bloque correspondan a diferentes hablantes (en la práctica se alternan voces masculinas y femeninas).
- Deben utilizarse múltiples muestras para cada voz, las cuales han de ser únicas entre los diferentes hablantes y para un mismo hablante.
- El conjunto total de muestras de prueba (número de hablantes x número de muestras por hablantes) se mantiene equilibrado a lo largo de todo el experimento, de modo que las comparaciones entre códigos y conjuntos de códigos se basan en series de material equivalentes.
- Las condiciones de prueba se ordenan dentro de bloques de tal forma que el orden promedio, para los bloques, sea aproximadamente igual a cada condición de prueba.
- Los oyentes deben estar sujetos a un tiempo de prueba máximo de 60 a 70 minutos, sin contar los descansos, para tener en cuenta los efectos de la fatiga y el tedio. En la práctica, el tiempo máximo es equivalente aproximadamente a 400 ensayos ACR y 200 ensayos DCR o CCR.

3.3.3 Cuadrado latino y diseños conexos

El cuadrado latino fue creado por Leonhard Euler, que utilizó como símbolos caracteres latinos.

En la ciencia combinatoria y la estadística, el cuadrado latino es un cuadro n × n que consiste en n símbolos diferentes de modo que cada símbolo aparece una sola vez en cada fila y exactamente una vez en cada columna. A continuación se muestra un ejemplo:

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

Se dice que el cuadro latino está reducido (o normalizado) si su primer fila y su primera columna están en orden natural. Por ejemplo, el cuadrado latino anterior está reducido, porque su primera fila y su primera columna son 1,2,3,4 (en lugar de 3,1,2,4 o en cualquier otro orden). Los cuadrados latinos pueden reducirse permutando (reordenando) las filas y columnas.

Aunque el cuadrado latino es un objeto simple para un matemático, tiene múltiples aplicaciones para el diseñador de experimentos. El mismo cuadrado latino puede utilizarse en muchas situaciones diferentes. Recuerde que un diseño experimental consiste en la atribución de condiciones experimentales que habrán de evaluar un grupo de sujetos.

3.3.4 Diseño del cuadrado grecolatino

Se dice que dos cuadrados latinos son ortogonales si uno puede superponerse sobre el otro, y cada una de las n^2 combinaciones de los símbolos (teniendo en cuenta el orden de la superposición) sólo se produce una vez en las n^2 células de la matriz. Estos pares de cuadrados ortogonales se suelen denominar cuadrados grecolatinos, dado que se acostumbra a utilizar letras latinas para los símbolos de un cuadrado y legras griegas para los del otro. A continuación figura un ejemplo de cuadrado grecolatino de orden 3.

Aα	Bγ	Cβ
Bβ	Cα	Aγ
Cγ	Aβ	Bα

Los cuadrados latinos y grecolatinos tienen una importante aplicación en la teoría estadística del diseño de experimentos.

3.3.5 Diseños anidados

Un diseño experimental en el que las variables tienen una jerarquía implícita se denomina diseño anidado. Otra definición es un tipo de diseño experimental en el que cada nivel de un determinado factor aparece en un solo nivel de cualquier otro factor. Los factores que no están anidados se dice que están cruzados.

Por ejemplo, un hospital consta de dos alas (I y II). Los pacientes en el ala I se asignan aleatoriamente al especialista A o al B. Los pacientes en el ala II se asignan aleatoriamente al especialista C o al D. Así los especialistas A y B están anidados respecto de los pacientes del ala I, mientras que los especialistas C y D están anidados respecto de los pacientes del ala II. Si, en cambio, todos los pacientes del hospital se han asignado aleatoriamente a uno de los cuatro especialistas, se dice que el diseño es cruzado.

Se recomienda el diseño anidado para estudiar el efecto de fuentes de variabilidad que se manifiestan a lo largo del tiempo. La recopilación de datos y el análisis son simples, y no hay razón para estimar la interacción cuando se analizan los errores dependientes del tiempo. Los diseños anidados pueden ejecutarse a varios niveles.

3.3.6 Otros diseños

Cuando hay más de un factor, podemos disponer de un diseño gráfico mixto o dividido.

Un diseño de parcela dividida consta de dos factores experimentales, A y B. Los niveles de A se asignan aleatoriamente a parcelas enteras (parcelas principales), y los niveles de B se asignan aleatoriamente a parcelas divididas (subparcelas) dentro de cada parcela entera. El diseño ofrece información más precisa sobre B que sobre A, y a menudo surge cuando A sólo puede aplicarse a unidades experimentales grandes.

Normalmente en los diseños de parcelas divididas cada bloque se divide en parcelas enteras y cada una de estas parcelas enteras en subparcelas (o parcelas divididas).

3.4 Escalas

En el texto clásico de J.P. Guilford sobre pruebas y mediciones, *Métodos psicométricos* (1954), el autor describe cuatro niveles generales de medición y las escalas relacionadas, a saber: *nominal*, *ordinal*, de *intervalo* y *razón*. Las reglas sobre la forma de asignar números a cada nivel de medida y las relaciones de estos números constituyen los criterios esenciales para definir las escalas. Cuanto mayor es el nivel de las escalas, más podemos hacer en la forma de operaciones estadísticas y matemáticas con los números obtenidos en la medición.

3.4.1 Escala nominal

La escala inferior de medición es la escala *nominal*. En la medición de escala nominal, los números se asignan a categorías o clases únicamente como etiquetas para dichas categorías. Todos los miembros de cada categoría se consideran iguales o equivalentes. Las escalas nominales constituyen la forma inferior de medición y, por ende, permite sólo el nivel inferior de tratamiento estadístico o matemático. En las escalas nominales, las operaciones estadísticas se limitan a las que implican *frecuencias*. Se puede contar el número de miembros asignados a cada categoría y luego ordenar las categorías por frecuencia. Además, si el número de categorías de cada factor se limita a dos, podemos describir la relación entre factores calculando un *coeficiente de contingencia*.

3.4.2 Escala ordinal

El siguiente nivel de escala de mediciones es la escala *ordinal*. En las mediciones con la escala ordinal, los números asignados a categorías ponen de manifiesto la propiedad de orden de clase. Los miembros asignados a una categoría con números inferiores se consideran de una clase inferior a los asignados a una categoría designada por un número más elevado. La distinción entre categorías se basa en alguna propiedad o cualidad de los miembros clasificados. Sin embargo, no se parte del supuesto de que las categorías están idénticamente espaciadas a lo largo de la escala o de que los intervalos entre categorías son iguales. Las operaciones estadísticas adecuadas comprenden todas las relacionadas con escalas nominales. Además, el principio de orden de clase permite calcular *medianas*, *centiles*, y *coeficientes de correlación de orden de clase*.

El término "efecto tope" se refiere a un efecto según el cual los datos no pueden tomar un valor más elevado que un determinado número, denominado "tope". Análogamente, el término "efecto base" se refiere a la situación en la que los datos no pueden tomar un valor inferior a un determinado número, denominado "base".

Los efectos tope y base en el ámbito de la recopilación de datos, cuando la varianza en una variable independiente no se mide ni calcula por encima de cierto nivel, son un problema práctico bastante corriente de recopilación de datos en muchas disciplinas científicas.

3.4.3 Escala de intervalo

El tercer nivel de escala de medición es la *escala de intervalo*. Las escalas de intervalos también se denominan escalas de *unidad idéntica* en las que las distancias numéricamente idénticas en la escala corresponden a distancias empíricamente idénticas en algún aspecto de los objetos que se miden. Por otra parte, las distancias entre puntos son aditivas, por ejemplo, la distancia de A a B más la distancia de B a C es igual a la distancia de A a C, o en forma de ecuación: $AB + BC = AC$. La única diferencia entre este nivel de escala de medición y el nivel más elevado, es decir, la escala de razón, es la ausencia de un cero absoluto. Sin embargo, la mayoría de las técnicas estadísticas pueden calcularse a partir de escalas de intervalo, en particular la *media*, la *desviación típica*, el *coeficiente de correlación producto-momento de Pearson*, y otros cálculos estadísticos que dependen de éstos.

3.4.4 Escala de razón

El nivel más alto de escala de medición es la escala de *razón*. La escala de razón tiene un cero absoluto, en la que cero representa *ninguna* de las propiedades representadas por la escala. Al igual que las de las escalas de intervalo, las mediciones sobre las escalas de razón cumplen la propiedad aditiva. Además, las razones de las mediciones son numéricamente equivalentes en las escalas de razón, es decir, 15/5 = 27/9 = 12/4 = 3/1, etc. En las escalas de razón, todas las técnicas y los métodos estadísticos son válidos.

Sección 4

Pruebas de opinión sobre conversación

4.0 Introducción

La Recomendación UIT-T P.805, *Subjetive evaluation of conversational quality* (Evaluación subjetiva de la calidad de la conversación) proporciona directrices, métodos y procedimientos para diseñar, llevar a cabo e informar de los resultados de las pruebas de conversación. Las pruebas de opinión sobre conversación permiten a los participantes que intervienen en las mismas encontrarse en una situación más realista que simula las condiciones reales de servicio experimentadas por los usuarios del teléfono. Además, estas pruebas están diseñadas para evaluar los efectos de la degradación que pueden causar dificultades durante la conversación (tales como retardos, pérdidas de paquetes, ecos, interrupciones, ruido, recortes, etc.). Pueden emplearse para estudiar los efectos globales del sistema o para determinar degradaciones específicas tales como el retardo.

Los participantes en la prueba se agrupan en pares de comunicadores. Se sientan en salas insonorizadas separadas y se les pide que mantengan una conversación a través de la cadena de transmisión y a continuación den su opinión según las diferentes escalas de calidad. Los entornos de ruido acústico pueden simularse en una o en ambas salas.

Dependiendo del objetivo de la prueba, pueden participar en la misma personas expertas, con experiencia o sin ninguna formación al respecto. Estas pruebas pueden ser útiles para los fabricantes, los operadores y los clientes, y constituyen una importante herramienta de evaluación porque proporcionan la simulación más cercana de las interacciones de telefonía real entre abonados. Los participantes sin entrenamiento específico entran en juego cuando es importante obtener una indicación de cómo va a estimar el público en general la calidad global del servicio telefónico y la dificultad en el uso de la conexión con el sistema sometido a prueba. Puede utilizarse para realizar una evaluación global de la calidad de funcionamiento en una cierta gama de condiciones. Sin embargo, los participantes sin entrenamiento específico no pueden describir e identificar con precisión los tipos de degradación asociada al sistema sometido a prueba.

A continuación se indican las principales características de las pruebas de opinión sobre conversación:

- Un parecido muy semejante a una conversación real, donde se pide a las personas que interactúen y puedan adaptar su comportamiento para acomodarse al sistema sometido a prueba.
- Una tarea para estimular una conversación con igual participación de ambas partes.
- Posible variación en el comportamiento de los distintos participantes durante una conversación (debido a diferencias de cultura, personalidad, etc.) que podría dar lugar a una mayor variación en las respuestas de los participantes que están evaluando la calidad de la señal vocal.
- Reconocimiento de que las medidas finales pueden ser menos sensibles que las pruebas de sólo escucha, ya que los participantes deben concentrarse en la conversación en la que intervienen y no específicamente en la evaluación de la calidad durante la conversación.
- Disponibilidad en el laboratorio de pruebas de dispositivos sometidos a prueba y herramientas de simulación que deben actuar en tiempo real.

- Las pruebas de conversación son el método más adecuado para medir el efecto sobre la aceptación de ciertas degradaciones del sistema, tales como el retardo.

Esta metodología de prueba de conversación puede adaptarse a pruebas en funcionamiento real; no obstante, está previsto que el control de algunas variables experimentales (por ejemplo, el retardo, la pérdida de paquetes, el ruido acústico, etc.) puede ser limitado.

4.1 Instalaciones de prueba

Una prueba de conversación debe proporcionar un entorno de comunicación que sea lo más realista posible. Todos los procesos en el enlace de comunicaciones deben serlo en tiempo real.

La conmutación entre condiciones que implican diferentes códecs y/o distintos parámetros de red debe ser transparente para los participantes. Ello puede exigir una instrumentación y unos procedimientos especializados.

La asimetría entre dos participantes en una comunicación es típica en muchas situaciones reales de comunicación vocal; una situación asimétrica puede estar definida por diferentes entornos de ruido acuático o diversas condiciones de transmisión. Puede que sea precisa una consideración especial para garantizar una simulación precisa de los entornos de ruido acústico. Por ejemplo, es muy significativo que se necesite una potencia de baja frecuencia para simular los entornos de automóviles.

En la figura 4.1 se ilustra una instalación de pruebas típica:

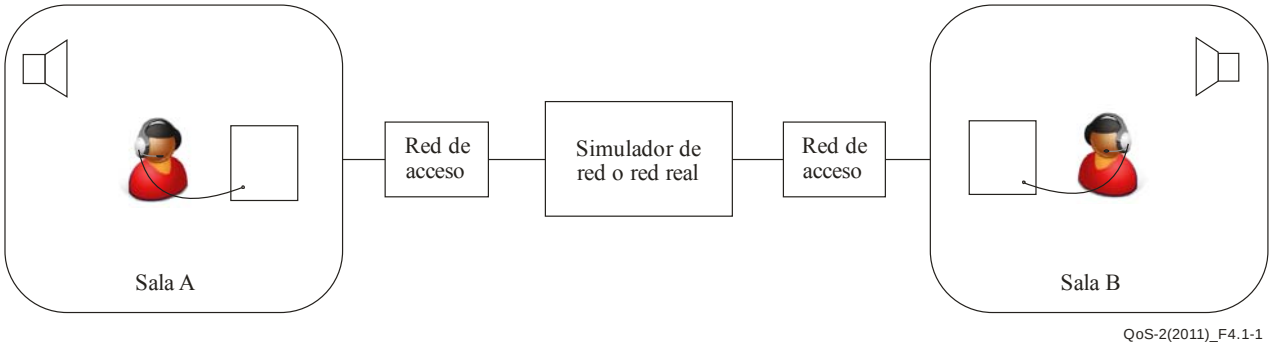


Figura 4.1 – Ejemplo de instalaciones de prueba

Cada participante está sentado en una sala insonorizada donde pueden simularse diversos entornos de ruido acústico. El entorno en ambas salas puede ser el mismo o distinto. Como ejemplos de diferentes entornos pueden citarse las salas silenciosas, las oficinas, los automóviles, las estaciones de ferrocarril, los trenes, las calles y las cafeterías. Una sala silenciosa puede simularse introduciendo un nivel adecuado de ruido de Hoth. Algunas salas también permiten considerar la reverberación como una variable experimental.

En la Recomendación UIT-R P.800 figura una descripción de las salas insonorizadas.

Además, los sensores de emisión y recepción utilizados por los participantes pueden ser iguales o distintos. Por ejemplo, puede utilizarse casco, casco con micrófono o micrófono y altavoz; la elección de los equipos depende de cada caso.

4.2 Diseño de la prueba

La mayoría de los temas relativos al diseño de las pruebas de sólo escucha también son pertinentes para las pruebas de conversación, por ejemplo, las condiciones de referencia y la presentación de los efectos de orden. Una limitación importante en el diseño de la prueba de conversación es la duración de cada tarea individual, o ensayo, necesaria para llevar a cabo cada condición experimental. Un sistema de comunicación adecuado exige que la conversación dure un mínimo de 2 minutos. Las pruebas típicas requieren de 4 a 5 minutos, en los que el periodo de conversación dura de 2 a 3 minutos y el periodo de respuesta otros 2 minutos. Ello limitaría el número total de condiciones en la sesión de un participante a unas 24. Por tanto, una sesión llevaría unas 3 horas, incluidas las instrucciones, las condiciones preliminares y las pausas. Las tareas establecidas para medir algunas de las degradaciones del sistema pueden exigir que las conversaciones duren más de 2 ó 3 minutos.

Debe llegarse a un compromiso entre la duración de la prueba y la elección de las condiciones. Si van a probarse más condiciones, la prueba debe dividirse en varias sesiones/experimentos y puede que sean necesarios varios grupos distintos de participantes.

A continuación aparece un ejemplo de programación para una prueba de 24 condiciones:

	Visita 1				Visita 2		
	Instrucción	Sesión 1	Pausa	Sesión 2	Sesión 3	Pausa	Sesión 4
Número de conversaciones		7 (incluidas las prácticas)		6	6		6
Tiempo	15 minutos	35 minutos	10 minutos	30 minutos	30 minutos	10 minutos	30 minutos

Las condiciones que son idénticas en ambos sentidos y utilizan los mismos sensores y el mismo ruido acústico se dcenominan condiciones simétricas. Cualquier otro caso se considera asimétrico. En condiciones asimétricas es necesario que los pares de participantes cambien de lugar en cada condición. Ello limita el número total a unas 12 condiciones asimétricas.

Es sabido que debe llegarse a una solución de compromiso entre la resolución de la prueba y el número de votos por condición. La relación entre estos dos parámetros viene dada por la siguiente ecuación general para la mínima diferencia significativa (MSD):

$$MSD = \pm C_{df,p}\sqrt{EMSq/n}$$

Siendo $C_{df, p}$ un valor similar a t determinado por la prueba estadística particular, el nivel de probabilidad (p) y el grado de libertad (df); EMSq es el error cuadrático medio obtenido del ANOVA y n es el número de votos por condición.

A fin de lograr una resolución suficiente entre las condiciones, se recomienda que el mínimo número de pares de participantes sea, en general, 16. Sin embargo, cabe señalar también que este número puede reducirse en algunas circunstancias con objeto de acortar el tiempo necesario para realizar la prueba, pero ello disminuiría la fiabilidad de los resultados.

4.3 Tarea de conversación

La Recomendación UIT-T P.800 y el Manual sobre Telefonometría del UIT-T proporcionan descripciones de *pruebas de conversación completas* que eran los métodos recomendados antes de elaborar la Recomendación UIT-T P.805. Dicha Recomendación describe la *tarea de conversación corta* que es el método recomendado para oyentes sin entrenamiento y el método más ampliamente utilizado en las actuales pruebas de conversación. Al desarrollar las *tareas de conversación corta* utilizadas en un experimento específico deben tenerse presente las siguientes consideraciones.

Debe seleccionarse la tarea que mejor satisfaga los requisitos del objetivo específico del experimento y los factores culturales del grupo de sujetos. Las características requeridas para seleccionar una tarea son las siguientes:

- Debe permitir la generación de un número suficiente de versiones equivalentes. Cada versión debe estimular un nivel equivalente de conversación e interacción.
- Debe estimular conversaciones semiestructuradas. Las conversaciones demasiado "abiertas" hacen imposible medir la eficacia de la comunicación mientras que las que están demasiado estructuradas no dejan margen para que los participantes se hagan una opinión equilibrada del canal.
- Debe aprenderse fácilmente.
- Debe ser intrínsecamente motivante.
- Debe permitir interrupciones por parte de los participantes.
- Debe ser insensible a los cambios en la estrategia o capacidad de los participantes para llevar a cabo la tarea.
- Debe representar un esfuerzo de cooperación entre los comunicadores en vez de un esfuerzo competitivo.
- Debe inducir a los participantes a hacer uso de un vocabulario rico y variado y alentar una interacción bidireccional suficiente.
- Debe inducir a debates fonéticamente ricos y con amplia distribución en el tiempo (expresiones e interrupciones cortas y largas).

La Recomendación UIT-T P.805 contiene una lista de ejemplos de *tareas de conversación corta* en tres idiomas, alemán, francés e inglés.

4.4 Procedimiento de prueba

4.4.1 Selección de los participantes en la prueba

La elección de participantes sin entrenamiento, con experiencia o expertos depende de las cuestiones y del grado de precisión requerido en los resultados.

4.4.1.1 Participantes sin entrenamiento

Los participantes sin entrenamiento están acostumbrados al uso diario de un teléfono. Sin embargo, no tienen experiencia en la metodología de pruebas subjetivas ni son expertos en implementaciones técnicas de los equipos sometidos a prueba. Idealmente, no tienen conocimiento específico sobre el dispositivo que van a evaluar. En coherencia con la Recomendación UIT-T P.800, los participantes no han tomado parte en ninguna prueba subjetiva de los seis meses anteriores. Para controlar la variabilidad experimental asociada con la familiaridad del par de participantes, puede que sea conveniente requerir que estos pares de participantes no se conozcan entre sí. La información sobre la familiaridad debe señalarse (desconocidos, conocidos casuales, amigos, miembros de una misma familia, etc.). A cada par de participantes se les da la

oportunidad de familiarizarse entre sí, durante un periodo de tiempo controlado. Debe emplearse el tiempo necesario para informar a los participantes sobre el procedimiento de la prueba y las tareas que deben llevar a cabo. Las condiciones prácticas [cuyos resultados no se incluirán en el análisis de los resultados] deben utilizarse al principio de la prueba para garantizar que los participantes están suficientemente familiarizados con el procedimiento de prueba y comprenden perfectamente las tareas que han de realizar. El conjunto de participantes debe ser representativo del grupo de usuarios de telecomunicaciones y de la aplicación que ha de medir el experimento.

4.4.1.2 Participantes con experiencia

Estos participantes tienen experiencia en pruebas subjetivas e incluyen a las personas que toman parte periódicamente en estas pruebas subjetivas, pero no a individuos que administran, diseñan o llevan a cabo de forma rutinaria evaluaciones subjetivas. Los participantes con experiencia pueden describir detalladamente un evento de auditoría y pueden separar distintos eventos basándose en degradaciones específicas. También son capaces de describir sus impresiones subjetivas con detalle. Sin embargo, los participantes con experiencia no tienen formación sobre las implementaciones técnicas de los equipos sometidos a prueba ni cuentan con un conocimiento detallado sobre la influencia de estas implementaciones en la calidad subjetiva.

4.4.1.3 Participantes expertos

Se trata de participantes con experiencia en pruebas subjetivas. Pueden describir un evento de auditoría y dar sus impresiones subjetivas con detalle así como separar distintos eventos basándose en degradaciones específicas. Tienen experiencia en implementaciones técnicas de los equipos sometidos a prueba y poseen conocimientos detallados de la influencia que esas implementaciones en concreto pueden tener sobre la calidad subjetiva. Las personas directamente implicadas en el diseño o desarrollo del sistema específico sometido a prueba deberán ser excluidas de esa prueba en particular.

4.4.1.4 Directrices generales de selección

Por regla general deben tenerse en cuenta las orientaciones indicadas en la Recomendación UIT-T P.800 a la hora de seleccionar a los participantes en la prueba.

Deben tomarse las precauciones necesarias cuando se seleccionen participantes para las pruebas de conversación. Al igual que sucede en los equipos de procesamiento de la señal vocal, algunos posibles participantes tendrán más experiencia que otros. Se sabe que la experiencia con tecnologías o equipos específicos oscila continuamente entre personas que no están en absoluto familiarizadas con el comportamiento técnico de los equipos sometidos a prueba (no expertos) e individuos totalmente competentes en el funcionamiento y mantenimiento de estos equipos (expertos).

La edad y el sexo de todos los tipos de participantes, junto con sus asociados, debe registrarse en todos los tipos de pruebas, pero especialmente en las pruebas de conversación formal distintas a las evaluaciones informales realizadas por expertos.

A menos que el sexo, la edad y otras características socioeconómicas sean factores de diseño de la prueba, toda prueba de conversación formal debe llevarse a cabo con una combinación aleatoria de participantes.

4.4.2 Criterios de evaluación subjetiva y escalas de opinión

Las escalas de opinión descritas en este punto han sido elaboradas en la Recomendación UIT-T P.805. La intención es que después de cada prueba (correspondiente a una condición de prueba específica), los participantes evalúen múltiples aspectos de su experiencia de comunicación. Se indican las siguientes cuestiones, C1-C6, como ejemplos y pueden considerarse representativas de los múltiples aspectos que han de considerarse. Estas cuestiones ejemplo incluyen escalas de notas de cinco puntos, así como una escala de

respuesta binaria en un caso (C5). La carga cognoscitiva sobre los participantes y, por tanto, el número de cuestiones planteadas debe minimizarse para reducir la fatiga de estos participantes y cualquier posible confusión.

C1 – *¿Cómo evalúa usted la calidad de sonido de la voz de la otra persona?*

5 Ninguna distorsión, natural
4 Mínima distorsión
3 Distorsión moderada
2 Considerable distorsión
1 Fuerte distorsión

C2 – *¿Con qué nitidez entiende lo que la otra persona le está diciendo?*

5 Ninguna pérdida de comprensión
4 Mínima pérdida de comprensión
3 Pérdida de comprensión moderada
2 Considerable pérdida de comprensión
1 Fuerte pérdida de comprensión

C3 – *¿Qué nivel de esfuerzo le supuso entender lo que la otra persona le estaba diciendo?*

5 Ningún esfuerzo
4 Un mínimo esfuerzo
3 Un esfuerzo moderado
2 Un esfuerzo considerable
1 Un gran esfuerzo

C4 – *¿Cómo evalúa su nivel de esfuerzo para mantener la conversación?*

5 Ningún esfuerzo
4 Un mínimo esfuerzo
3 Un esfuerzo moderado
2 Un esfuerzo considerable
1 Un gran esfuerzo

C5a – *¿Detectó (introduzca aquí la distorsión de interés)?* Sí o No

C5b – *En caso afirmativo, ¿qué grado de perturbación apreció?*

5 Ninguna perturbación
4 Mínima perturbación
3 Perturbación moderada
2 Perturbación considerable
1 Fuerte perturbación

C6 – *¿Cuál es su opinión sobre la calidad global de la conexión que ha estado utilizando?*

5 Excelente

4 Buena

3 Regular

2 Mediocre

1 Mala

Las cuestiones antes indicadas (C1-C6) sólo son ejemplos del tipo de cuestiones que deben utilizarse en las pruebas de conversación. El número y la redacción de las cuestiones específicas, y sus escalas de respuesta, deben ser revisadas a fin de adaptarse a las necesidades y requisitos del experimento. Cuando se utilizan múltiples cuestiones y escalas para evaluar el aspecto multidimensional de la calidad debe tomarse la precaución de garantizar que los participantes no conocen las respuestas precedentes.

4.4.3 Ejemplo de instrucciones dadas a sujetos que van a tomar parte en una prueba de conversación

INSTRUCCIONES A LOS SUJETOS

En este experimento estamos evaluando sistemas que pueden ser utilizados por servicios de telecomunicaciones.

Usted va a mantener una conversación con otro usuario. La situación de prueba simula comunicaciones entre dos equipos sometidos a prueba. La mayoría de las situaciones corresponderán a condiciones ambientales silenciosas, pero en algunas se simularán situaciones más específicas tales como el ruido de un automóvil, el de una estación de ferrocarril o el de un entorno de oficina donde otras personas hablan a una cierta distancia.

Tras completar cada conversación se le pedirá que dé su opinión sobre la calidad respondiendo a las seis siguientes cuestiones que aparecerán en la pantalla que está frente a usted. Sus respuestas serán almacenadas.

Dispone de ocho segundos para responder a cada cuestión. Tras pulsar el botón "siguiente" en la pantalla, aparecerá otra cuestión. Repita el procedimiento para cada una de las seis cuestiones siguientes.

¿Cómo juzga la calidad de sonido de la voz de la otra persona?

Ninguna distorsión, natural	Mínima distorsión	Distorsión moderada	Considerable distorsión	Fuerte distorsión

¿Qué nivel de esfuerzo le supone entender lo que la otra persona le esta diciendo?

Ningún esfuerzo	Un mínimo esfuerzo	Un esfuerzo moderado	Un esfuerzo considerable	Un gran esfuerzo

¿Cómo evalúa su nivel de esfuerzo para mantener la conversación?

Ningún esfuerzo	Un mínimo esfuerzo	Un esfuerzo moderado	Un esfuerzo considerable	Un gran esfuerzo

¿Detectó la presencia de eco?

No	Sí

En caso afirmativo, ¿qué grado de perturbación apreció?

Ninguna perturbación	Mínima perturbación	Perturbación moderada	Perturbación considerable	Fuerte perturbación

¿Cuál es su opinión sobre la calidad global de la conexión que ha estado utilizando?

Excelente	Buena	Regular	Mediocre	Mala

Posteriormente habrá una pausa de unos 30 minutos. La prueba durará aproximadamente 60 minutos.

Le rogamos que no comente sus opiniones con el resto de personas que participan en el experimento.

4.4.4 Recopilación de datos

Deben proporcionarse hojas de respuesta a los participantes para que puedan anotar sus opiniones sobre cada condición así como para responder a cualquier otra pregunta que se les haga. Alternativamente, pueden utilizarse técnicas de registro automático de datos (por ejemplo, un sistema de respuesta apretando un botón, un PC con teclado y pantalla o ratón y pantalla) para recopilar los datos en bruto.

4.4.5 Análisis de los datos e informe sobre los resultados

4.4.5.1 Métodos de análisis

Dependiendo del diseño experimental, se llevarán a cabo, según el caso, pruebas-t, pruebas de significado múltiple, análisis de la varianza (ANOVA) o análisis múltiple de la varianza (MANOVA). Un análisis de correlación entre las diversas cuestiones debe soportar los resultados de MANOVA. Este análisis ofrece una descripción de los resultados que puede entenderse con mayor facilidad.

4.4.5.2 Resultados obtenidos

El informe de la prueba de conversación debe describir las cuestiones particulares utilizadas en el experimento, incluidas las categorías de escalas empleadas para cada cuestión. El resumen de los resultados debe incluir, como mínimo, las notas medias y las desviaciones típicas para todas las condiciones probadas y para todas las cuestiones.

Sección 5

Pruebas de opinión de escucha

5.0 Introducción

Las pruebas de opinión de escucha han sido el método más comúnmente utilizado en los ejercicios de pruebas subjetivas. La experiencia ha demostrado que éstas son de administración sencilla y válidas para evaluar la calidad percibida en la mayoría de las situaciones hipotéticas de telecomunicación.

5.1 Material de base de datos

Los hablantes para las pruebas de escucha deben seleccionarse cuidadosamente. El idioma de la prueba debe ser su lengua materna y no han de tener anomalías o degradaciones manifiestas. Sus voces deben ser cribadas por *oyentes experimentados* para detectar degradaciones del habla y asegurar que sus voces no representan ningún extremo de las variables básicas del habla, como timbre fundamental y velocidad de conversación. Además, éstas deben estar exentas de todo acento regional o dialecto cultural pronunciado. En las pruebas subjetivas realizadas en inglés de América del Norte, por ejemplo, se determinó que el dialecto americano general era el dialecto más desprovisto de acentos culturales y/o regionales. Éste es el dialecto hablado que prevalece en la región del oeste medio de los Estados Unidos.

En general, la demografía de la muestra de hablantes debería ser representativa de la población de usuarios de la o las aplicaciones que son objeto de prueba. Normalmente en la muestra de hablantes se debería mantener un equilibrio de género y ésta debería ser suficientemente diversa como para representar la gama de variabilidad de individuos de habla normal en la población de que se trate.

Para ciertas pruebas concretas podría solicitarse cierta flexibilidad con respecto a las normas generales, pero éstas deben controlarse cuidadosamente.

5.1.1 Conversación

5.1.1.1 Estructura del fichero vocal

En general los ficheros vocales originales para las pruebas subjetivas se producen en el laboratorio de prueba. Normalmente los ficheros vocales procesados para las pruebas subjetivas que se entregan al laboratorio de escucha o instalación de prueba subjetiva proceden del laboratorio anfitrión, como un fichero digital en formato PCM, por ejemplo de 16 bit, PCM lineal, complemento de 2, notación "Little-endian", etc. La velocidad de muestreo depende de la finalidad de la prueba subjetiva y las aplicaciones particulares que se someten a prueba. Por lo general los ficheros vocales están hechos de muestras de conversación individuales para una combinación de hablante y vocalización. La longitud de las diferentes muestras también depende del objetivo y las aplicaciones. Por razones prácticas, todos los ficheros vocales fuente o de origen para un experimento deben tener la misma duración. La primera frase del fichero debe estar precedida de un periodo de silencio de aproximadamente 0,3-0,5 segundos, y un periodo de silencio similar debe seguir a la última frase del fichero. La o las frases de una muestra deben ser sencillas y significativas, según se describe en el anexo B.1.4 a la Recomendación UIT-T P.800.

Cabe señalar que el silencio de aproximadamente 0,4 segundos tras el final de la última frase en el fichero es de extrema importancia, dado que (para ciertas condiciones) hay una serie de filtros FIR con un gran número de coeficientes. Si no está presente el silencio prescrito se corre un riesgo considerable de mutilar la conversación al final del fichero.

Para evitar los efectos de contraste de ruido, cualquier laguna y/o pausa de silencio incorporada a los ficheros vocales para ajustarlos al formato especificado, no debería ser de silencio digital puro. El acolchonado debe hacerse añadiendo el ruido ambiental presente durante la grabación del material de conversación, de modo que ninguno de los ficheros vocales tenga silencio digital.

5.1.1.2 Duración de muestra corta (8 a 12 segundos)

En la mayor parte de las pruebas subjetivas los ficheros vocales están formados de muestras cortas (8 a 12 segundos), compuestas de dos frases breves concebidas específicamente para pruebas de conversación. En el inglés norteamericano (NAE) éstas se conocen como Frases Harvard [de "Prácticas recomendadas para la medición de la calidad de la conversación", en *IEEE Transactions on Audio and Electroacoustics* (1969), vol. 17, pág. 227-46]. En la figura 5.1.1.2-1 se ilustra la disposición de una muestra típica de conversación breve utilizada en pruebas subjetivas, y en el cuadro 5.1.1.2-1 se muestra una serie típica de pares de Frases Harvard utilizadas en muestras de conversación breve.

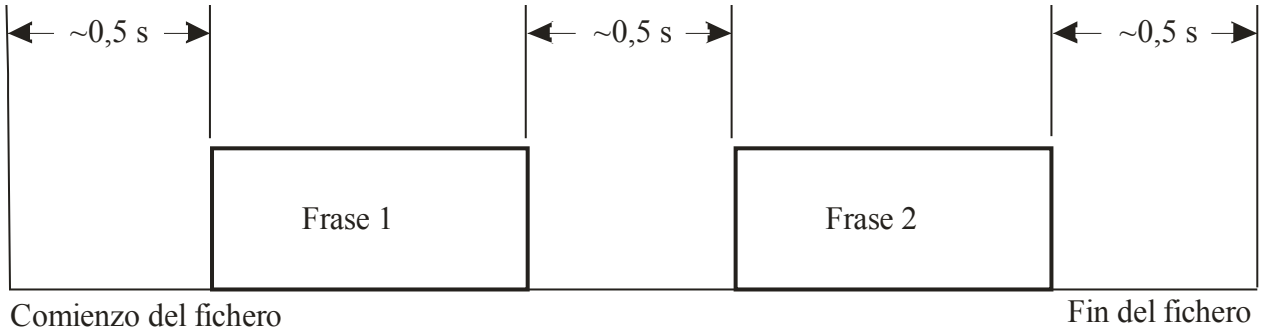


Figura 5.1.1.2-1 – Estructura de pares de frase para un fichero vocal

Cuadro 5.1.1.2-1 – Pares de frases de muestra compuestas de Frases Harvard

The birch canoe slid on the smooth planks.	Glue the sheet to the dark blue background.
It's easy to tell the depth of a well.	These days a chicken leg is a rare dish.
Rice is often served in round bowls.	The juice of lemons makes fine punch.
The box was thrown beside the parked truck.	The hogs were fed chopped corn and garbage.
Four hours of steady work faced us.	A large size in stockings is hard to sell.

5.1.1.3 Duración de muestra larga (12 a 24 segundos)

Se pueden utilizar muestras de conversación largas para presentar una distribución realista de las degradaciones, esto es, degradaciones con características de variación en el tiempo (por ejemplo, desvanecimiento radioeléctrico de los sistemas celulares, pérdida de paquetes para Internet y traspasos para GSM). Cada muestra contendrá cuatro frases diferentes y durará aproximadamente 16 segundos, con el intervalo de tiempo entre las frases descrito en el anexo B.1.4 a la Recomendación UIT-T P.800. Cada fichero vocal original de muestra debería contener conversación activa por lo menos durante 9 segundos y como máximo 20 segundos. Cabe señalar que en algunas bases de datos vocales puede resultar difícil de cumplir con este último requisito. La Frase Harvard NAE típica tiene una duración de menos de 2 segundos, de modo que cuatro de estas frases durarían menos de los 9 segundos de conversación activa requeridos. Por consiguiente, se tolerará una flexibilidad razonable en el cumplimiento de este requisito.

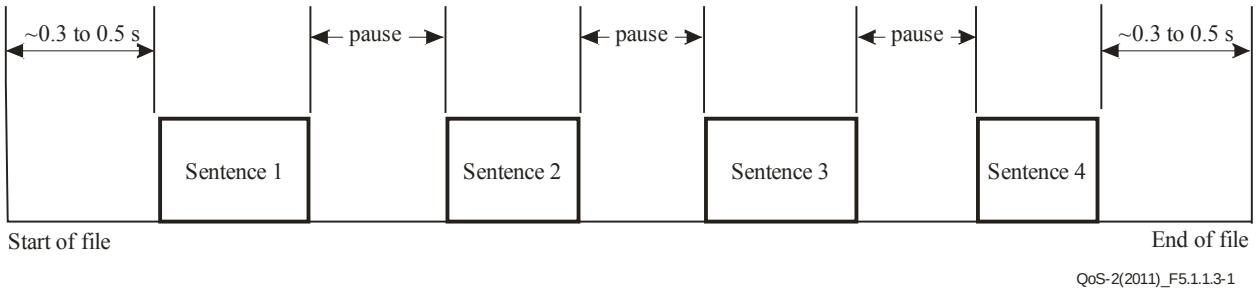


Figure 5.1.1.3-1 – Example of speech file structure for sentence-quadruples

5.1.1.4 Duración de muestra larga (45 segundos a 3 minutos)

Se puede recurrir a la Recomendación UIT-T P.880, Evaluación continua de la calidad vocal que varía con el tiempo (*Continuous Evaluation of Time Varying Speech Quality*, CETVSQ), para evaluar las degradaciones con características de extrema duración. En estas pruebas, la longitud de la muestra vocal estará determinada por la finalidad de la prueba. Las muestras vocales deben construirse como ampliaciones de las muestras largas antes mencionadas.

Puede adquirirse material vocal de fuente pregrabado según se describe en el punto 5.3.2.5.

De preferencia, el laboratorio de escucha debería proporcionar su propio material vocal de fuente. Así pues, se recomienda aplicar el procedimiento estipulado en el punto 5.3.

5.1.2 Base de datos sobre ruido

La finalidad de realizar la prueba en condiciones de ruido de fondo es evaluar la calidad de funcionamiento del códec o sistema en un entorno representativo del mundo real. Los entornos de ruido de fondo difieren en el grado de variabilidad en el tiempo, tanto en lo que respecta al nivel como al espectro. Los entornos con menor variabilidad de nivel y espectro se clasifican como entornos de ruido estático o estacionario. Los entornos con mayor variabilidad se clasifican como entornos de ruido dinámico o no estacionario. Puede grabarse una muestra de ruido de fondo en el entorno en el cual éste ocurre en realidad (por ejemplo, en un vehículo en movimiento, en una intersección de calles activas, en una cafetería muy ocupada y agitada), o bien éste puede simularse (por ejemplo, creando ruido de murmullos mediante la combinación de múltiples hablantes). En ambos casos la muestra de ruido debe tener una duración suficiente como para ser representativa del entorno real que se pretende simular.

ETSI STQ ha compilado una serie representativa de muestras de ruido de fondo, que se puede descargar y examinar en el sitio web http://docbox.etsi.org/STQ/Open/EG 202 396-1 Background noise database/.

Los entornos de ruido de fondo más comunes evaluados en pruebas pasadas realizadas por el UIT-T han sido de automóviles, murmullos, oficinas y hablantes interferentes. En los siguientes párrafos se proporcionan detalles sobre el contenido de cada una de esas muestras de ruido de fondo.

5.1.2.1 Ruido de automóvil

La finalidad de este ruido es someter a prueba la calidad de funcionamiento del códec en condiciones de ruido de fondo estático constante, y debe grabarse en un automóvil en movimiento, para lo cual se recomienda una velocidad constante entre 80 y 110 km/h (50-70 mph). La marca y el modelo de automóvil deben ser razonablemente comunes en el país donde se efectúa la grabación. Por lo general las ventanas del automóvil se mantienen cerradas y la radio apagada.

5.1.2.2 Ruido de murmullos

La finalidad de este ruido es someter a prueba la calidad de funcionamiento del códec en condiciones de ruido no estacionario con un espectro similar al espectro a largo plazo de la conversación. En la grabación se deberían combinar materiales de conversación continua entre un número suficiente de hablantes, de modo que un oyente de esa lengua materna no pueda entender ninguna palabra o frase con sentido. Este tipo de ruido no debe tener periodos de silencio. Además, la grabación no debe proporcionar a los oyentes de ese mismo idioma o de otro idioma ninguna indicación sobre el idioma de los hablantes. Las grabaciones utilizadas en el pasado contenían conversaciones simultáneas de 40 hablantes. El número de hablantes masculinos y femeninos debe ser aproximadamente igual.

5.1.2.3 Ruido de oficinas

La finalidad del ruido de oficinas es representar un entorno de trabajo clásico. Este tipo de ruido también debe contener sonidos típicos de oficina tales como los de teclados, ventiladores de ordenador, timbres de teléfonos, impresoras, acondicionamiento de aire, etc.

5.1.2.4 Ruido de hablante interferente

Este ruido está destinado a comprobar la calidad de funcionamiento del códec con un solo hablante interferente. En los experimentos en los que se utilicen muestras de ruido largas añadidas a materiales de frases concatenadas, se debería utilizar una grabación de conversación nominalmente natural entre un hombre y una mujer. Los periodos de silencio durante la conversación no deberían ser de más de 0,5 segundos.

El o los encargados del experimento deciden caso por caso si el hablante interferente debe hablar el mismo idioma que el material vocal original ("hablante rival"), o bien actuar en cambio como una "fuente de distracción ininteligible". Este tipo de ruido aún merece ser objeto de nuevos estudios; si se le impone el mismo idioma que el del hablante primario, la fuente de distracción pasa a ser inteligible y aporta algunas propiedades psicoacústicas que podrían no comprenderse cabalmente en el contexto de la Recomendación UIT-T P.835, mientras que una fuente de distracción ininteligible podría clasificarse más fácilmente como ruido genérico "de tipo conversación".

5.1.3 Procesamiento para añadir ruido y conversación

El laboratorio de escucha debe procesar todos los ficheros vocales antes de transferirlos al laboratorio anfitrión para su procesamiento en condiciones experimentales. Este procesamiento previo garantiza que la conversación se halla al nivel correcto y tiene las características de insumo adecuadas. En un Documento sobre procesamiento de ficheros se describe detalladamente el procesamiento requerido.

El *experimento x*, por ejemplo, exige que el material vocal de fuente esté mezclado con ruido de fondo. Esto es responsabilidad del laboratorio anfitrión. En el Documento sobre procesamiento se describe detalladamente el proceso que se ha de aplicar.

Puesto que en este experimento hipotético la conversación se presentará a individuos con auriculares, en una etapa posterior al procesamiento (responsabilidad del laboratorio anfitrión) se filtrará la conversación con el sistema intermedio de referencia modificado (*Modified Intermediate Reference System*, mod-IRS), según se especifica en el anexo D a la Recomendación UIT-T P.830. En el Documento sobre procesamiento de ficheros también se describe detalladamente el procesamiento ulterior.

5.2 Sistemas de referencia

Los sistemas de referencia caracterizados por condiciones conocidas proporcionan un método útil para hacer comparaciones significativas de los resultados de las pruebas subjetivas procedentes de distintos laboratorios, o del mismo laboratorio, en diferentes momentos. Entre estas condiciones figura la condición mejor posible o "*Directa*", así como condiciones en las cuales se han incorporado a los materiales vocales degradaciones controladas conocidas. En las pruebas subjetivas es indispensable la presencia de Sistemas de Referencia. A veces estos sistemas actúan como un punto de anclaje. La condición "*Directa*" es la mejor condición que puede lograrse en el experimento; en esta condición no se codifica la conversación de entrada, sino que tiene únicamente el mismo filtrado de entrada, nivel de entrada y filtrado de salida que la conversación que ha sido procesada por el códec que es objeto de prueba.

5.2.1 Sistema intermedio de referencia

En la Recomendación UIT-T P.48 se especifica el sistema intermedio de referencia (IRS) que se utilizó para definir la puntuación de la sonoridad en la telefonía de banda estrecha. La descripción debería bastar para permitir que se reproduzcan en los distintos laboratorios equipos con las características requeridas y una calidad de funcionamiento normalizada.

5.2.2 Sistema intermedio de referencia modificado

El UIT-T recomienda que se utilicen las características de frecuencia del sistema intermedio de referencia modificado, según se define en el anexo D a la Recomendación UIT-T P.830.

5.2.3 Sistemas de filtrado uniforme

Los sistemas de filtrado uniforme pueden obtenerse utilizando los instrumentos informáticos de la Recomendación UIT-T G.191.

5.2.4 Otros sistemas de filtrado

Para anchuras de banda más elevadas pueden encontrarse otros sistemas de filtrado en la biblioteca de herramientas informáticas de la Recomendación UIT-T G.191; para los ejercicios de banda ancha, por ejemplo, se utiliza la Recomendación UIT-T P.341.

5.3 Compilación de material de origen o fuente

5.3.1 Entorno de grabación

Al preparar los materiales de fuente, se recomienda aplicar el procedimiento descrito en esta sección. Las grabaciones vocales deberían realizarse en entornos acústicos y eléctricos que cumplan con los requisitos estipulados en el anexo B.1.1 a la Recomendación UIT-T P.800.

5.3.2 Sistema de grabación

El método recomendado consiste en grabar digitalmente la conversación con un micrófono lineal y un amplificador de bajo nivel de ruido con una respuesta de frecuencia uniforme. Para lograr una relación señal-ruido (SNR) óptima, el micrófono debe ubicarse a 15-20 cm de los labios del hablante. Si las exhalaciones de aliento del hablante son muy evidentes se ha de utilizar una pantalla antiviento.

La grabación debe efectuarse directamente en la computadora o con un sistema de grabación de alta calidad. En todo caso, el sistema A/D debe estar en conformidad con las especificaciones del anexo B a la Recomendación UIT-T P.800. Los sistemas A/D modernos producen muestras de 48 kHz, complementos de 2, de 16 bit, con anchura de banda de 20 a 24 000 Hz y una excelente calidad de gama dinámica y distorsión

5.3.2.1 Procedimientos de grabación

El Manual sobre Telefonometría del UIT-T (1992) contiene un ejemplo de los procedimientos de grabación.

Se pueden utilizar simultáneamente sistemas/canales de grabación separados para registrar la conversación en banda ancha/banda superancha/con anchura de banda completa en una pista, y/o la conversación telefónica (banda estrecha) en otra pista. A menudo se utiliza otro canal para registrar las señales de control.

Normalmente sólo se necesita una de esas grabaciones para cualquier experimento, pero en ocasiones es necesario utilizar más de una, y es conveniente en todos los casos poder hacer mediciones comparativas en diferentes anchuras de banda.

Si se necesitan nuevos materiales grabados, se recomienda grabar la conversación con anchura de banda completa a una frecuencia de muestreo de 48 kHz y utilizar instrumentos informáticos para muestrear hasta la anchura de banda necesaria.

5.3.2.2 Señal de calibración

A efectos de calibrar la conversación al nivel correcto, debe grabarse un tono de calibración que se almacena en un fichero digital. Normalmente esta es una señal sinusoidal de 30 segundos a 1 kHz con un nivel de RMS igual al nivel vocal nominal deseado, es decir, –26 dBov. El nivel de cresta de esta señal sinusoidal es ±2 322 y su RMS es de 1 642. Dependiendo del experimento, pueden utilizarse otras señales de calibración, según proceda.

5.3.2.3 Hablantes

Deben intervenir al menos dos hablantes masculinos y dos femeninos, pero el equilibrio del plan de prueba podría exigir más hablantes de cada sexo.

5.3.2.4 Niveles vocales

En la grabación se observa el nivel vocal activo consignado en la Recomendación UIT-T P.56. Durante el proceso de grabación hay que asegurarse de que el nivel vocal activo en ambos sistemas de grabación se encuentra entre 20 y 30 dB por debajo del nivel de sobrecarga del sistema de grabación para cada frase

medida separadamente. Se deberán volver a grabar cualesquiera grupos de frases que no cumplan con este requisito.

5.3.2.5 Base de datos vocal plurilingüe de NTT-AT en CD-ROM, 1994

En 1994 NTT-AT fabricó y puso a la venta una base de datos vocal (CD-ROM) con ejemplos de señales vocales en 21 idiomas.

Véase http://www.ntt-at.com/prdsvc/audiovisual.html.

5.3.3 Control de la base de datos vocal/musical

Los anexos a la Recomendación UIT-T P.800 contienen información sobre la preparación y el procesamiento del material vocal, la filosofía del experimento (incluida la elección de las condiciones del circuito), el procedimiento de la prueba de escucha y el tratamiento de los resultados.

Es conveniente efectuar algunas mediciones para caracterizar la base de datos vocal/musical utilizada para el ejercicio de pruebas subjetivas (por ejemplo, excluir artefactos o lugares con valores S/N muy bajos, o cuantificar/cualificar el ruido de fondo, etc.). Puede ser útil conocer las características de la base de datos para la interpretación de resultados peculiares observados a partir del análisis de las puntuaciones de las pruebas subjetivas.

5.3.4 Procesamiento de los ficheros vocales

Las etapas del procesamiento están relacionadas con las situaciones hipotéticas simuladas, y por lo general se describen en el plan de procesamiento del laboratorio anfitrión.

5.3.5 Convenio de designación del fichero vocal de fuente/procesado

Por ejemplo, los nombres de fichero de las muestras vocales son xxLGySz.cnn, siendo.

- xx el número del experimento
- L el designador del laboratorio
- G el género del hablante (es decir, M para mujer, H para hombre)
- y el número del hablante 1, 2, 3…
- S la muestra
- z el número de la muestra (1, 2, 3, 4 ó 5)

5.4 Procedimiento de prueba de escucha

Véase el anexo B.4 a la Recomendación UIT-T P.800.

5.4.1 Entorno de escucha

Véase el anexo B.4.1 a la Recomendación UIT-T P.800.

5.4.2 Sistema de escucha

Véase el anexo B.4.2 a la Recomendación UIT-T P.800.

Se reconoce que las actuales Recomendaciones UIT-T no contemplan la medición y ecualización de la respuesta de frecuencia del audífono para las pruebas subjetivas en banda superancha y con anchura de banda completa, y éste es un tema que ha de ser objeto de nuevos estudios. Las mediciones de la respuesta realizadas utilizando micrófonos de sonda ubicados en el canal auditivo de los individuos sometidos a

prueba, tal como se describe en UIT-R BS.708, resultan difíciles y poco prácticas para la mayoría de las organizaciones que desean efectuar pruebas subjetivas. Actualmente las mediciones que se realizan utilizando un simulador de cabeza y torso (*Head-and-Torso-Simulator*, HATS) sólo son eficaces hasta 10 kHz. Entretanto, en las directrices proporcionadas en los recientes planes de pruebas subjetivas con banda superancha en general se recomienda utilizar auriculares circumaurales de alta calidad (para que queden bien adheridos a la cabeza) y se incluyen ejemplos de modelos adecuados. Éstos deben tener baja distorsión y cumplir con los requisitos de respuesta de frecuencia en campo difuso consignados en UIT-R BS.708 para reducir al mínimo la coloración de la señal. Con el fin de evitar fugas a otros oyentes, se han de utilizar auriculares cerrados si varios oyentes se encuentran en la misma sala.

5.4.3 Nivel de escucha

Véase el anexo B.4.3 a la Recomendación UIT-T P.800.

5.4.4 Idoneidad de los participantes

Véase el anexo B.4.4 a la Recomendación UIT-T P.800.

Dependiendo de la finalidad de la prueba, los participantes pueden carecer de formación (cándidos), ser experimentados o expertos. Dado que estas pruebas pueden ser de utilidad para los fabricantes, operadores y proveedores de contenidos, constituyen una importante herramienta de evaluación ya que pueden proporcionar una simulación válida de los servicios telefónicos reales. Se recurrirá a participantes sin formación (cándidos) cuando sea importante obtener una indicación de la puntuación que le daría la población usuaria de teléfonos a la calidad general y la dificultad de utilización de la conexión con el sistema que se somete a prueba. Esto puede servir para proporcionar una evaluación "global" de la calidad de funcionamiento en toda una serie de condiciones. Sin embargo, según se indicó en el punto 4 sobre Pruebas de opinión conversacional, en general los participantes sin formación no son capaces de identificar y describir con precisión los tipos de degradación del sistema objeto de prueba.

Para las definiciones de oyentes cándidos, experimentados y expertos, véase el punto 4.4.1.

5.4.5 Criterios de evaluación subjetiva y escalas de opinión

Se han utilizado varias escalas de cinco notas para juzgar las categorías con diversos fines. La presentación y los textos de las escalas de opinión utilizadas por los participantes en los experimentos son muy importantes, y deben ajustarse a las normas establecidas tras años de experiencia y práctica.

Véase en el anexo B.4.5 a la Recomendación UIT-T P.800 las escalas de opinión recomendadas por el UIT-T.

5.4.6 Equivalencia de escalas de opinión entre idiomas

Se deberían utilizar textos equivalentes, compatibles con el uso normal dentro del idioma, lo que podría dar lugar a variaciones con respecto al texto inglés original.

5.4.7 Juicio de las categorías

Las siguientes escalas de opinión son las más utilizadas para las aplicaciones del UIT-T.

5.4.7.1 Escala de calidad de escucha

Nota de la calidad de la señal vocal:

Excelente	5
Buena	4
Regular	3
Mediocre	2
Mala	1

Las notas resultantes de la puntuación (nota media de opinión sobre la calidad de escucha o sencillamente nota media de opinión) se representan por el símbolo MOS. Se trata de una puntuación de categoría absoluta (ACR).

5.4.7.2 Escala de esfuerzo de escucha

La denominación de las escalas de opinión de esfuerzo en la escucha es muy importante. Sin ellas el resto de descripciones puede provocar graves equívocos.

Esfuerzo necesario para comprender el significado de las frases	Nota
Audición perfecta; ningún esfuerzo	5
Cierta atención es necesaria; ningún esfuerzo apreciable	4
Esfuerzo moderado	3
Esfuerzo considerable	2
Significado incomprensible, aun con el mayor esfuerzo	1

Las notas resultantes de la puntuación (nota media de opinión de esfuerzo en la escucha) se representan por el símbolo MOS_{LE}, pero cuando no se dispone de notación con sufijos, se utiliza el símbolo MOSle. Se trata de una puntuación de categoría absoluta (ACR).

5.4.7.3 Escala de sonoridad preferida

Sonoridad preferida	Nota
Mucho mayor que la preferida	5
Mayor que la preferida	4
Preferida	3
Menor que la preferida	2
Mucho menor que la preferida	1

La magnitud evaluada a partir de las notas (nota media de opinión de sonoridad preferida) se representa por el símbolo MOS_{LP}, pero cuando no se dispone de notación con sufijos se utiliza el símbolo MOSlp. Se trata de una puntuación de categoría absoluta (ACR).

NOTA – En 2.6 del Manual sobre Telefonometría, aparecen ejemplos de escalas subjetivas alternativas que pueden utilizarse únicamente si las tres escalas de opinión anteriores no satisfacen las necesidades del experimentador.

5.4.7.4 Escala de degradación

Véase el anexo D a la Recomendación UIT-T P.800.

El método de determinación de índices por categorías absolutas (ACR) descrito en el anexo B suele tener poca sensibilidad para diferenciar los circuitos de buena calidad. Con una versión modificada del

procedimiento ACR denominada procedimiento de determinación de índices por categorías de degradación (*degradation category rating*, DCR) se logra una sensibilidad más elevada. Este procedimiento se ha adaptado de las Recomendaciones del CCIR (Documento 11/17 del CCIR, Evaluación subjetiva de la calidad de las imágenes de televisión (VER), periodo de estudios 1978-1982) para efectuar evaluaciones de circuitos de buena calidad.

El procedimiento DCR, que utiliza en particular una escala de molestias y una referencia de calidad antes de evaluar cada configuración, parece adecuado para evaluar señales vocales en condiciones deterioradas (por ejemplo, ruidosas).

Deben darse instrucciones a los participantes para que clasifiquen las condiciones según una escala de categorías de degradaciones de cinco notas, a saber:

Degradación	Nota
Degradación inaudible	5
Degradación audible, pero no molesta	4
Degradación ligeramente molesta	3
Degradación molesta	2
Degradación muy molesta	1

Las notas resultantes de la puntuación (nota media de opinión de degradación) se representan por el símbolo DMOS.

5.4.7.5 Escala de degradación modificada

Conforme al procedimiento de determinación de índices por categorías de degradación modificado (DCR), deben darse instrucciones a los participantes para que clasifiquen las condiciones según una escala de categorías de degradaciones de cinco notas, a saber:

Degradación	Nota
No se percibe degradación, e incluso se percibe cierta mejora	5
Se percibe degradación, pero no es molesta	4
La degradación es ligeramente molesta	3
La degradación es molesta	2
La degradación es muy molesta	1

Este método se utilizó para el ejercicio UIT-T G.729EV (2005).

5.4.7.6 Escala de comparación

Véase el anexo E a la Recomendación UIT-T P.800.

El método de determinación de índices por categorías de comparación (*Comparison Category Rating*, CCR) es semejante al método de determinación de índices por categorías de degradación (DCR). En cada prueba se presentan a los oyentes un par de muestras vocales. En el procedimiento DCR, se presenta primero una muestra de referencia (no procesada), seguida por la misma muestra vocal procesada por alguna técnica. En el método DCR los oyentes siempre evalúan el grado en que se ha degradado la muestra procesada (la segunda) con relación a la muestra no procesada (la primera). En el procedimiento CCR, se elige al azar en cada prueba el orden de las muestras procesada y no procesada. En la mitad de las pruebas, la muestra no procesada va seguida por la muestra procesada. En las pruebas restantes se invierte el orden.

Los oyentes utilizan la escala siguiente para calificar la calidad de la segunda muestra con relación a la de la primera:

La calidad de la segunda comparada con la calidad de la primera es:

Calidad	Nota
Mucho mejor	+3
Mejor	+2
Ligeramente mejor	+1
Aproximadamente igual	0
Ligeramente peor	–1
Peor	–2
Mucho peor	–3

En realidad los oyentes proporcionan dos juicios con una sola respuesta: "¿qué muestra tiene mejor calidad?" y "¿en qué medida?". Los métodos DCR y CCR son particularmente útiles para evaluar el comportamiento de los sistemas de telecomunicación cuando su entrada está deteriorada por el ruido de fondo. No obstante, una ventaja del método CCR sobre el DCR es la posibilidad de evaluar el procesamiento de la señal vocal que o bien degrada o mejora la calidad de la voz. Las notas resultantes de las puntuaciones (nota media de opinión sobre las comparaciones) viene representada por el símbolo CMOS.

NOTA – Debe tenerse cuidado cuando se utiliza el método CCR. Ciertos laboratorios han encontrado que el método es útil para evaluar los sistemas de reducción de ruido. Sin embargo, cuando se utilizó este método en las evaluaciones subjetivas del códec UIT-T G.729 (8 kbit/s), se encontró que el método era demasiado sensible al evaluar el comportamiento del códec para señales vocales mezcladas con ruido de fondo.

5.4.7.7 Escala de comparación modificada

En el método de determinación de índices por categoría de degradación modificado se utilizan muestras de referencia procesadas, mientras que en el método CCR normal se utilizan muestras de referencia sin procesar, y se presenta a los participantes la escala de puntuación resultante (al participante se le formula la misma pregunta que en la escala de determinación de índices por comparación de categorías).

Este método ha sido utilizado para el interfuncionamiento de puntos de prueba fijos/flotantes (por ejemplo, anexo C a la Recomendación UIT-T G.722.1).

5.4.7.8 Método del umbral

Por comparación directa de un sistema de transmisión con uno de referencia es posible evaluar el comportamiento del sistema sometido a prueba en términos de característica de degradación del sistema de referencia que puede variarse y ajustarse a valores definidos. Un ejemplo de dicha característica es la relación señal/ruido, SNR (para la definición véase 8.2.3 de UIT-T P.830). Véase el anexo F a la Recomendación UIT-T P.800.

Se utiliza un procedimiento de prueba de escucha solamente. Se presenta a los oyentes un par de señales compuesto por una señal de referencia y una señal de prueba, y se les pide que indiquen cuál de las dos señales del par consideran que tiene más calidad (índice de preferencia). La equivalencia subjetiva se define como el valor de referencia correspondiente al punto de intersección de la curva de regresión de las notas de preferencia al nivel de preferencia del 50%.

5.4.7.9 Escala de preferencia (comparación par)

Los métodos de escala directa dan una imagen satisfactoria del modo según el cual el observador percibe los estímulos que se le presentan, pero al mismo tiempo se le exige mucho a su capacidad de juicio. A diferencia de éstos, los métodos de escala indirecta sólo exigen una puntuación del tipo "mayor que" o "mejor que" al comparar dos (o más) estímulos al mismo tiempo. Como consecuencia de ello, los datos primarios obtenidos no contienen más información de la que contiene una escala ordinal. Y, sin embargo, mediante ciertas suposiciones, es posible construir una escala intervalo a partir de esos datos.

El método de comparación par es el más utilizado de esta categoría. Para clasificar una serie de estímulos A, B, C, D…, éstos se comparan en pares en todas las n(n – 1) combinaciones AB, BA, CA, BD, etc. A menudo se elige como variable de comparación "preferencia". Luego los participantes indican, después de que se presenta cada par, si prefieren el primer o el segundo estímulo del par. Se deben controlar los errores de posición, tiempo y orden mediante una aleatorización adecuada del orden de presentación.

5.4.7.10 Escalas de puntuación de la Recomendación UIT-T P.835

En esta Recomendación se describe una metodología para evaluar la calidad subjetiva de la voz en presencia de ruido y, en particular, para la evaluación de los algoritmos de compensación de ruido. En la metodología se utilizan escalas de puntuación diferentes para estimar por separado la calidad subjetiva de la señal de voz, el ruido de fondo y la calidad global.

En cada prueba el participante indica tres puntuaciones, una para cada frase o submuestra.

– *Escala de puntuación de la señal*

En una frase de cada prueba se le indica al participante que preste atención únicamente a la señal de voz y valore el grado de distorsión de la señal de voz. Para valorar la señal de voz utilizará la escala de puntuación que se muestra *infra*. Deberá elegir la frase de la lista que en su opinión mejor describe la calidad de la señal de voz únicamente.

Prestando atención únicamente a la señal de voz, seleccionar la categoría que mejor describe la muestra que acaba de escuchar. La señal de voz de esta muestra era:

Sin distorsión	5
Ligeramente distorsionada	4
Algo distorsionada	3
Bastante distorsionada	2
Muy distorsionada	1

– *Escala de puntuación del ruido de fondo*

En otra frase de cada prueba se le indica al participante que preste atención únicamente al ruido de fondo y valore lo evidente o molesto que resulta el ruido de fondo. Para grabar las puntuaciones del ruido de fondo utilizará la escala de puntuación que se muestra abajo. Deberá elegir la frase de la lista que en su opinión mejor describe el ruido de fondo únicamente.

Prestando atención únicamente al ruido de fondo, seleccionar la categoría que mejor describe la muestra que acaba de escuchar. El ruido de fondo de esta muestra era:

Imperceptible	5
Ligeramente perceptible	4
Perceptible aunque no molesto	3
Algo molesto	2
Muy molesto	1

– *Escala de puntuación de la calidad general*

Para la tercera frase de cada prueba se le indicará al participante que preste atención a toda la muestra de sonido (tanto a la señal de voz como al ruido de fondo) y que dé su opinión sobre la **calidad global** de la muestra como si fuera una comunicación diaria corriente.

Seleccione la categoría que mejor describe la muestra que acaba de escuchar como si fuera una comunicación diaria corriente. La muestra de voz global era:

Excelente	5
Buena	4
Regular	3
Mediocre	2
Mala	1

5.4.7.11 Escala de calidad continua

En la Recomendación UIT-T P.880 se describe la metodología denominada evaluación continua de la calidad vocal que varía con el tiempo (CETVSQ) que sirve para evaluar los efectos de las fluctuaciones temporales de la calidad vocal en la calidad instantánea percibida (es decir, percibida en cualquier instante de la secuencia vocal) y en la calidad percibida general (es decir, al final de la secuencia vocal). El método consta de dos partes: en primer lugar, la valoración instantánea que se regula mediante un cursor mientras que se escucha la secuencia vocal en una escala continua, y en segundo lugar, la valoración global en una escala de cinco categorías convencional al final de la secuencia vocal.

Este método no es aplicable para seleccionar los códecs vocales. Sin embargo, puede servir de herramienta de diagnóstico de los efectos de la degradación en la calidad instantánea y global percibida, en particular en el caso de las degradaciones discontinuas que varían con el tiempo (por ejemplo, las debidas a la pérdida de paquetes IP, el traspaso a redes móviles, etc.). También puede servir para desarrollar y validar instrumentos de medición objetiva destinados a predecir la calidad vocal mediante la detección y análisis de diversos tipos de degradaciones en la señal vocal.

Este método se inspiró en el método SSCQE (evaluación continua de la calidad con un solo estímulo) empleado en el campo de vídeo (Recomendación UIT-R BT.500-11) y se ha comprobado su validez para la calidad vocal en diversos estudios previos.

Para la evaluación de la calidad continua debe fabricarse un dispositivo que debe tener las siguientes características:

– mecanismo del cursor sin posición de "reinicio" (es decir, sin que vuelva automáticamente a una posición predefinida);
– deslizamiento lineal de unos 10 cm;
– posición fija o de sobremesa;
– las muestras de la "posición del cursor" se han de tomar dos veces por segundo (velocidad suficiente para capturar con precisión las respuestas de los participantes);
– la "posición del cursor" podrá codificarse desde 0 (valor inferior de la escala) hasta un mínimo de 100 (valor superior de la escala), para lograr una resolución aceptable. La posición inicial del cursor debe ser la correspondiente a la mitad de la escala.

Procedimientos de evaluación continua:

En primer lugar, los participantes tienen que valorar continuamente la calidad vocal de la secuencia desplazando el cursor a lo largo de escala continua de modo que la posición de éste indique su opinión de la calidad en ese instante; podrán deslizar libremente el cursor a lo largo de toda la escala. Para ayudar al participante a ajustar la posición del cursor a unas gamas adecuadas de calidad vocal, se indicarán cinco posiciones en la escala, a saber, excelente, buena, regular, mediocre y mala.

Etiquetas de escala de calidad continua utilizadas para el juicio instantáneo:

Excelente
Buena
Regular
Mediocre
Mala

En segundo lugar, los participantes tienen que valorar la calidad global de cada secuencia al final de la misma, utilizando para ello la siguiente escala de calidad de la escucha de 5 categorías (la misma MOS empleada en la ACR).

Escala de la calidad global (ACR)

Excelente	5
Buena	4
Regular	3
Mediocre	2
Mala	1

5.4.7.12 Escalas multidimensionales

La calidad de sonido general puede considerarse como un atributo multidimensional, es decir constituida por una combinación de dimensiones perceptuales separadas. Por consiguiente, puede resultar útil complementar los juicios sobre la calidad general con juicios sobre las dimensiones perceptuales específicas (Manual sobre Telefonometría, 1992).

Actualmente la CE 12 del UIT-T está estudiando enfoques perceptuales para el análisis multidimensional.

Las hipótesis de prueba de las metodologías de escalas multidimensionales deberían abarcar como mínimo los siguientes tipos de distorsión:

- códecs vocales individuales y en cascada utilizados hoy en día en el entorno de las telecomunicaciones;
- estrategias de ocultación y pérdida de paquetes (conexiones con conmutación de paquetes);
- tramas y bits erróneos (conexiones inalámbricas);
- interrupciones (tales como pérdida de paquetes no ocultos o traspaso en GSM);
- recorte frontal (mutilación temporal);
- recorte de amplitud (sobrecarga, saturación);
- efectos de los sistemas de procesamiento vocal tales como sistemas de reducción del ruido y compensadores de eco en señales vocales limpias;
- efectos de los sistemas de procesamiento vocal tales como sistemas de reducción del ruido (fase de adaptación y estado convergente) y compensadores de eco en señales vocales con ruido;
- efectos de los sistemas de codificación de la voz en las conversaciones con ruido;

- retardo variable (VoIP, videotelefonía)/degradación en el tiempo;
- variaciones de ganancia;
- influencia de las distorsiones lineales (configuración espectral), también variable con el tiempo;
- distorsiones no lineales producidas por el micrófono/transductor en interfaces acústicas;
- sistemas de mejoramiento de las señales vocales en redes y terminales y sus efectos en la calidad de la escucha;
- reverberaciones causadas por instalaciones de prueba manos libres en entornos acústicos definidos.

Actualmente se está estudiando el número y la naturaleza de las escalas de puntuación requeridos para describir con precisión las degradaciones indicadas *supra*.

5.4.8 Instrucciones para los participantes

Los participantes no deben tener ninguna duda sobre los procedimientos de prueba, la grabación de sus respuestas, etc., y a tales efectos se deben preparar e impartir a los participantes instrucciones adecuadas antes de las sesiones de prueba. Por otro lado, no se les debe explicar a los participantes la naturaleza de las condiciones de circuito que se somete a prueba, el tipo de degradación, etc., salvo que están evaluando la calidad de una conexión telefónica. Todas las Recomendaciones UIT-T de la serie P que versan sobre pruebas subjetivas (por ejemplo, UIT-T P.800, UIT-T P.805, UIT-T P.835, UIT-T P.880, etc.) contienen ejemplos de instrucciones para los participantes.

Sección 6

Análisis estadístico y presentación de los resultados

6.0 Introducción

Este capítulo ofrece una panorámica general de los procedimientos y los métodos estadísticos apropiados para analizar datos extraídos de pruebas subjetivas. No pretende ser un manual de análisis estadístico. Más bien, presenta principios y procedimientos generales para elaborar un análisis estadístico específico que permita presentar los resultados de experimentos subjetivos.

6.1 Conceptos básicos de estadística

6.1.1 Principios generales de estadística

En un experimento, la persona que lo realiza manipula una o más variables a fin de determinar el efecto de esta manipulación en otras variables. Las variables independientes son las variables que controla la persona que realiza el experimento, mientras que las variables dependientes reflejan cualquier efecto asociado a la manipulación de las variables independientes. La estadística persigue dos objetivos: extraer unas conclusiones válidas sobre los efectos de las variables independientes en las variables dependientes y formular, a partir de la muestra analizada, generalizaciones válidas para la población (*inferencia estadística*).

La relación entre variables se caracteriza por dos propiedades elementales: la fuerza (o "intensidad") y la fiabilidad. Esta fiabilidad se refiere a la "representatividad" de la relación observada en una muestra específica en comparación con toda la población. Evidentemente, ambas propiedades (fuerza y fiabilidad) no son totalmente independientes, y su relación depende en gran medida del tamaño de la muestra. En una muestra con un tamaño determinado, cuanto más fuerte sea la relación entre variables, más fiable será su relación. Si una relación entre variables es "objetivamente" débil (en la población), será imposible identificar esta relación en un experimento, salvo que la muestra tenga un tamaño considerable. Aun cuando la muestra sea perfectamente representativa, su efecto no será estadísticamente significativo si la muestra es pequeña. Si la relación es objetivamente muy fuerte (en la población), podrá ser muy significativa, incluso en una muestra pequeña.

Significancia estadística y nivel p

La significancia estadística de un resultado es una medición estimada del nivel en el cual dicho resultado es "verdad", es decir, que es representativo de la población. El nivel *p* representa un índice descendente de la fiabilidad de un resultado. Cuanto mayor sea el nivel *p*, menor será la posibilidad de que la relación observada entre las variables de la muestra sea un indicador fiable de la relación entre las respectivas variables en la población. Más exactamente, el nivel *p* representa la probabilidad de error en relación con la aceptación de un resultado observado como válido, es decir, que es representativo de la población. Por ejemplo, un nivel *p* de 0,05 indica que la probabilidad de que la relación entre las variables encontradas en la muestra analizada sea una coincidencia o "se deba al azar" es del 5%. En muchos ámbitos de investigación, un nivel *p* de 0,05 suele considerarse como un "límite" de error aceptable.

Para determinar la significancia de los efectos, se necesita una función que represente la relación entre "fuerza" y "significancia" para un tamaño de muestra determinado. Esta función da la probabilidad de error

para rechazar la idea de que el efecto no existe en la población, es decir para rechazar lo que se conoce como hipótesis nula – H0.

Hipótesis estadísticas H0 y H1

Inicialmente, el objetivo del experimento es evaluar una hipótesis científica. Mediante la lógica deductiva, la hipótesis científica y su negación se expresan como dos hipótesis estadísticas exhaustivas y mutuamente excluyentes que formulan predicciones sobre un parámetro de población. Generalmente, la hipótesis nula asume que la observación o el efecto guardan relación con factores aleatorios debidos a variaciones en la muestra y no con uno o varios factores sistemáticos que explican la observación. En otras palabras, una hipótesis nula es la declaración de que todas las diferencias observadas son el resultado de variaciones relacionadas con la muestra de población, o dependen de la variación aleatoria, y no de otro factor. La hipótesis alternativa asume que la variación aleatoria no puede explicar las diferencias observadas.

Estas hipótesis, expresadas como H0 (hipótesis nula) y H1 (hipótesis alternativa), se basan en la media de la población, su mediana, la varianza, etc. Si, como suele suceder, no es posible observar todos los elementos de la población, se obtendrá una muestra aleatoria. Esta muestra permite estimar el parámetro de población desconocido. El proceso mediante el cual se decide si se rechaza la hipótesis nula se denomina prueba estadística. Existen numerosas pruebas estadísticas, como las que comparan medias o proporciones, las que comparan varianzas, las que estudian la relación entre variables, etc. La mayoría de estas funciones pertenecen a un tipo general de función denominado "normal".

Pruebas estadísticas

La mayoría de pruebas estadísticas se basan directamente en la distribución normal o en distribuciones derivadas de la ley normal, como la *distribución t*, la *distribución F* o la distribución Chi². La representación de la distribución normal es la famosa "campana de Gauss", entre cuyas propiedades está que el 68% de sus observaciones están en un intervalo de ± 1 veces la desviación típica con respecto a la media. Asimismo, un intervalo de 1,96 veces la desviación típica contiene el 95% de las observaciones.

Un ejemplo de prueba estadística es la estadística $z = (\bar{Y} - \mu_0) \times \sqrt{n} / \sigma$, que se puede usar para comprobar una hipótesis relativa a una media de población μ ($\bar{Y}$ es la media de una muestra, μ_0 es el valor hipotético de la media de población según la hipótesis nula $\mu \leq \mu_0$, σ es la desviación típica y *n* es el tamaño empleado para calcular $\bar{Y}$). Si la hipótesis nula es cierta y si, aproximadamente, Y está distribuida normalmente (o si *n* es grande, según el *teorema del límite central*[2]), la distribución de muestreo de esta estadística será la distribución típica normal con una media igual a 0 y una desviación típica igual a 1. Si se elige un nivel de significancia de 0,05 (α o *nivel p*), la hipótesis H0 será rechazada si *z* está en una región crítica definida por el 5% superior de la distribución del muestreo de z.

Una prueba estadística en la que la región crítica, definida por una probabilidad α, se encuentra en el extremo superior o inferior de la distribución de muestreo recibe el nombre de *prueba unilateral*. Si la región crítica se encuentra en el extremo superior y en el extremo inferior de la distribución de muestreo (probabilidad de $\alpha/2$ para ambos), recibe el nombre de prueba *bilateral*. En consecuencia, la persona que realiza el experimento probablemente tendrá menos argumentos para rechazar una hipótesis nula falsa con una prueba bilateral que si usa una prueba unilateral. Se recurre a la prueba unilateral cuando la persona que realiza el experimento elabora una hipótesis direccional sobre el fenómeno estudiado. A menudo, carecemos

---

2 Según este teorema, si se toman muestras aleatorias de una población con una media μ y una desviación típica finita σ, conforme el tamaño de la muestra aumenta *n*, la distribución de $\bar{Y}$ se aproxima a una distribución normal con una media μ y una desviación típica $\sqrt{n} / \sigma$. Para la mayoría de poblaciones encontradas en las ciencias del comportamiento y la educación, un tamaño de muestra de 100 es suficiente para obtener una distribución de la muestra prácticamente normal de $\bar{Y}$.

de información suficiente para elaborar una hipótesis direccional sobre un parámetro de población; simplemente creemos que el parámetro no es igual al valor especificado por la hipótesis nula. Ante tal situación, es necesario recurrir a una prueba bilateral, o no direccional. En este caso, en el ejemplo anterior, la hipótesis estadística para una prueba bilateral tendrá la siguiente forma: H_0: $\mu = \mu_0$ y H_1: $\mu \neq \mu_0$.

Por lo general, estas pruebas necesitan variables normalmente distribuidas, es decir, variables que cumplan la "hipótesis de normalidad". Algunas pruebas, como la de Kolmogorov-Smirnov o la de Shapiro-Wilk, permiten comprobar la normalidad de la distribución de las variables (véase: Marsaglia, G., Tsang, W.W., Wang, J. (2003) *Evaluating Kolmogorov's Distribution*, Journal of Statistical Software, 8 (18), 1-4).

No obstante, parece que las consecuencias de la violación de esta ley son menos importantes de lo previsto (véase: Boneau, C. Alan (1960), *The effects of violations of assumptions underlying the t test*. Psychological Bulletin 57 (1): pp. 49–64); y Edgell, Stephen E., & Noon, Sheila M (1984), *Effect of violation of normality on the t test of the correlation coefficient*. Psychological Bulletin 95 (3): pp. 576–583.).

6.1.2 Definiciones

A continuación se muestran las ecuaciones para la Media, la Desviación Estándar y el Coeficiente de Correlación, las estadísticas descriptivas más comúnmente utilizadas. Para cada ecuación, n es el número de valores de la muestra y X_i y Y_i son los valores individuales de una variable donde *i* varía de *1* a *n*. Todos los sumatorios (Σ) en las ecuaciones están por encima de la gama de i = 1 a *n*.

- Medida de tendencia central – *Media*

$$Media = \overline{X} = \sum_{i=1}^{n} X_i \Big/ n$$

La **media** es una medida especialmente informativa de la "tendencia central" de una variable, a condición de que se proporcionen sus intervalos de confianza. En general, cuanto mayor sea el tamaño de la muestra, más fiable será la media; cuanto mayor sea la dispersión, menos fiable será la media. En sentido general:

$$Media = (\Sigma x_i)/n$$

donde *n* es el tamaño de la muestra y x_i el valor de la variable dependiente para cada observación.

- Medida de variabilidad – *Desviación típica*

$$Desv.Tip = \sigma_x = \sqrt{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \Big/ n \over n-1}$$

La **Desviación Típica** es una medida comúnmente utilizada de dispersión de la **población**. Se calcula de la manera siguiente:

$$\sigma = [\Sigma(x_i - \mu)^2/N]^{1/2}$$

donde μ es la media real de la población y N el tamaño real de la población.

La **estimación de la Desviación Típica** de la población se calcula de la manera siguiente:

$$s = [\Sigma(x_i - \overline{X})^2/n\text{-}1]^{1/2}$$

donde $\overline{X}$ es la media de la muestra y *n* es el tamaño de la **muestra**.

- Medida de la relación entre variables – *Coeficiente de correlación*

$$Correl.(x,y) = r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i - \left[\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) \Big/ n\right]}{\sqrt{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \Big/ n}\sqrt{\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2 \Big/ n}}$$

Virtualmente, todas las operaciones estadísticas que se utilizan en pruebas subjetivas derivan de estas tres medidas de la tendencia central (Media), la variabilidad (Desviación Típica), y las relaciones entre variables (Correlación). Estas estadísticas asumen que la variable está distribuida normalmente, es decir, que la población de los valores forma una distribución "Normal" o Gaussiana.

Intervalos de confianza:

El **intervalo de confianza** CI especifica el intervalo alrededor de la media en el que puede esperarse la media "real" (de la población), con un nivel de certidumbre o nivel de confianza de 1-α. Para una distribución normal,

$$\text{CI} = \text{media} \pm (\text{valor z para el intervalo de confianza } 1-\alpha) \times (\sigma / \sqrt{n})$$

con un **nivel de confianza de 95%,** el intervalo de confianza CI se determina:

$$\text{CI} = \text{media} \pm 1.96 \times (\sigma / \sqrt{n})$$

con un **nivel de confianza de 99%,** el intervalo de confianza CI se determina:

$$\text{CI} = \text{media} \pm 2.58 \times (\sigma / \sqrt{n})$$

$$Lower\ 95\%\ C.I. = \bar{X} - 1.96\left(\sigma_X \Big/ \sqrt{n}\right) \qquad Upper\ 95\%\ C.I. = \bar{X} + 1.96\left(\sigma_X \Big/ \sqrt{n}\right)$$

Prueba t (prueba t de Student)

William Sealy Gosset, un químico conocido con el sobrenombre de "Student" y que trabajaba para la cervecera Guinness en Dublín, Irlanda, definió en 1908 la estadística t.

Hipótesis

La mayoría de estadísticas t tienen la forma T = Z/s, donde Z y s son funciones de los datos. Generalmente, Z está pensada para ser sensible a la hipótesis alternativa (es decir, su magnitud tiende a ser mayor cuando la hipótesis alternativa es cierta), mientras que s es un parámetro progresivo que permite determinar la distribución de T.

Por ejemplo, en una prueba t de una muestra

$$Z = \sqrt{n}\ \bar{X} / \sigma$$

donde $\bar{X}$ es la media de los datos de la muestra, n es el tamaño de la muestra y σ es la desviación típica de la población; en la prueba t de una muestra, s es σ/σ, donde σ es la desviación típica de la muestra.

Las hipótesis subyacentes en una prueba t son que

- Z sigue una distribución típica normal con arreglo a la hipótesis nula;
- ps^2 sigue una distribución χ^2 con p grados de libertad con arreglo a la hipótesis nula, donde p es una constante positiva;
- Z y s son independientes.

En un tipo específico de prueba t, estas condiciones son consecuencia de la población objeto de estudio y la manera como se haga el muestreo de los datos. Por ejemplo, en la prueba t que compara las medias de dos muestras independientes, deberían cumplirse las siguientes hipótesis:

- Cada una de las dos poblaciones comparadas debería seguir una distribución normal (se puede comprobar mediante una prueba de normalidad, como la prueba de Shapiro-Wilk o la de Kolmogorov-Smirnov, o evaluarlo mediante una gráfica de percentil normal).
- Si se utiliza la definición original de Student de la prueba t, las dos poblaciones comparadas deberían tener la misma varianza (se puede comprobar mediante la prueba de Levene, la de Bartlett o la de Brown-Forsythe, o evaluarlo mediante una gráfica de percentil normal). Si los tamaños de la muestra de los dos grupos comparados es más o menos igual, la prueba t de Student original es altamente resistente a la presencia de varianzas desiguales. La prueba t de Welch no es sensible a la igualdad de las varianzas, con independencia de que los tamaños de la muestra sean similares.
- Los datos empleados para llevar a cabo la prueba deberían muestrearse independientemente de las dos poblaciones comparadas. Por lo general, esta prueba no puede llevarse a cabo a partir de los datos; sin embargo, si se sabe que ha habido dependencia en el muestreo de los datos (es decir, si fueron muestreados en grupos), las pruebas t clásicas aquí expuestas podrían ofrecer resultados erróneos.

6.1.3 Aplicación en pruebas subjetivas

Este subapartado podría ofrecer información sobre el nivel p típico, el número de oyentes e iniciar la discusión sobre la comprobación de los oyentes.

En una prueba de opinión sobre la escucha, si un número L de oyentes examina una condición C para T hablantes/pares de frases distintos (T_m hombres y T_f mujeres), y se indican las calificaciones mediante Xc,l,t (l=1..L, t=1..T), puede obtenerse la puntuación de la opinión media (MOS) o la degradación de MOS (DMOS) para cada hablante a través de la condición C (Yc,t) utilizando:

$$Y_{c,\tau} = \frac{1}{L}\sum_{i=1}^{L} X_{c,I,t}$$

El MOS/DMOS desglosado por género de los hablantes del mismo género para la condición C (Yc,m and Yc,f) puede obtenerse a partir de:

$$Y_{c,m} = \frac{1}{T_m}\sum_{\substack{\text{hablantes}\\ \text{masculinos}}} Y_{c,t} \quad \text{para hablantes masculinos} \qquad Y_{c,f} = \frac{1}{T_f}\sum_{\substack{\text{hablantes}\\ \text{femeninos}}} Y_{c,t} \quad \text{para hablantes femeninos}$$

El MOS/DMOS (global) para los hablantes, sin distinción de género, para la condición C (Yc) puede obtenerse a partir de:

$$Y_c = \frac{1}{T}\sum_{t=1}^{T} Y_{c,t}$$

La desviación típica (S) para la condición C, expresada como Sc, puede calcularse como:

$$S_c = \sqrt{\frac{1}{L \times T - 1} \sum_{t=1}^{T} \sum_{l=1}^{L} (X_{c,l,t} - Y_c)^2}$$

En las pruebas estadísticas (véanse los párrafos siguientes) que se emplean para analizar datos recabados en pruebas subjetivas, el *nivel p* considerado suele ser de 0,05, y de 0,01 en aquellos pocos casos en los que la persona que lleva a cabo el experimento quiere reducir el riesgo de error.

Nótese que, por lo general, no se plantea la comprobación a posteriori de los sujetos, de modo que es necesario contar con un número suficiente de sujetos para garantizar que la medición no es demasiado sensible a comportamientos atípicos de estos.

6.2 Pruebas t de Student

6.2.1 Ventajas e inconvenientes

Debe ponerse mucho cuidado en la utilización sin limitaciones de pruebas t múltiples para estudiar las diferencias entre varias medias extraídas de una prueba subjetiva. Con este fin, se recomienda recurrir a pruebas de comparación post-hoc concretas. No obstante, conviene señalar que no se ha demostrado que la realización de múltiples pruebas t introduzca un sesgo sustantivo en tales resultados. Cuando sea necesario llevar a cabo comparaciones entre distintas medias, se recomienda recurrir a pruebas de comparación post-hoc específicas (véase el apartado 6.3, Análisis de la varianza). Nótese que, en caso de comparación de distintas medias, las diferencias entre algunas medias pueden ser considerables si se emplea una prueba t de Student, no así con un método de comparación post-hoc más adecuado, como el método de Tukey de la diferencia honestamente significativa. Así, la prueba t discrimina mucho más que el método de Tukey. Una consecuencia de este hecho es, por ejemplo, que, en la comprobación de requisitos para la fase de caracterización, resulta más sencillo verificar, mediante una prueba t, el requisito "mejor que" que mediante el método de Tukey, teóricamente apropiado, mientras que verificar los requisitos "no peor que" o "equivalente" resulta más complicado.

6.2.2 Definiciones

La prueba t de Student es la prueba más extendida para evaluar las diferencias en la media entre dos grupos, tanto entre grupos independientes como entre grupos apareados (es decir, dependientes). En ambos casos, se parte de la hipótesis de que ambos grupos tienen una distribución normal y unas varianzas equivalentes.

A fin de comparar dos grupos independientes, la *prueba t* calcula la *t* estadística:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{s_0 \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \text{ donde } s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

y donde $\overline{Y}_1$ y $\overline{Y}_2$ son las medias para los grupos 1 y 2 respectivamente, s_1 y s_2 sus desviaciones típicas y n_1 y n_2 sus tamaños.

Entonces se compara la *t* estadística con la distribución *t* teórica de Student (*bilateral* o *unilateral*, según la prueba), para un nivel de significancia determinado. En la hipótesis nula, H_0, no hay diferencia entre ambas medias; en la hipótesis alternativa, H_1, la diferencia entre ambas medias es significativa. La hipótesis nula se rechaza y se acepta la hipótesis alternativa si, para un nivel de probabilidad determinado (por lo general, ≤ 5%), la estadística *t* calculada es mayor que su valor teórico.

La distribución *t* es una función del número de grados de libertad (*df*) donde, para grupos independientes, $df = n_1 + n_2 - 2$. Nótese que, para tamaños de muestra grandes, la distribución *t* se aproxima a la distribución *z* o normal.

Tanto para grupos independientes como apareados, la prueba t se calcula a partir de la diferencia entre *n* observaciones apareadas y la estadística *t* se calcula así:

$$t = \frac{(\sum d_i)/n}{s_0\sqrt{\frac{1}{n}}} \text{ donde } s_0^2 = \frac{\sum (d_i - (\sum d_i)/n)^2}{(n-1)}$$

y donde *d* es la diferencia entre las observaciones apareadas y *n* es el número de resultados apareados.

La estadística t resultante se compara con la distribución *t* teórica de Student con grados de libertad: $df = n - 1$.

6.3 Análisis de la varianza

6.3.1 Finalidad e hipótesis subyacentes

En general, la finalidad del Análisis de la Varianza (ANOVA) es averiguar los efectos de factores experimentales (definidos en el capítulo 3). Amplía el principio de comparación de dos medias a varias medias: si solamente se comparan dos medias, ANOVA dará los mismos resultados que la *prueba t*. Este análisis trabaja con varias hipótesis. En primer lugar, se asume que las variables dependientes se miden al menos en una escala de intervalo (véase el capítulo 3 para las definiciones de escala). En segundo lugar, la variable dependiente está distribuida normalmente en el interior de los grupos. En tercer lugar, la distribución de los resultados se basa en la hipótesis de *homogeneidad de las varianzas*. Para esta última hipótesis, las varianzas en los distintos grupos deben ser equivalentes. No obstante, se ha señalado (Kirk, 1982) que ANOVA es un método bastante sólido con respecto a las desviaciones de la distribución normal y la homogeneidad de las varianzas.

6.3.2 Análisis

La forma más sencilla de ANOVA es el ANOVA de dos factores, en el que un factor es la variable de interés de la prueba (por ejemplo, códecs) y el otro, las respuestas de los sujetos. El factor-prueba representa un efecto fijo y el factor-sujeto, un efecto aleatorio. ANOVA calcula Sumas de cuadrados (SS) a partir de los datos primarios a fin de estimar las varianzas del factor. El SS total es la suma del SS Intergrupos (SS_{BG}, es decir, la varianza entre niveles del factor fijo) y el SS Intragrupos (SS_{WG}, es decir, la varianza entre niveles del factor aleatorio). La relación entre SS_{BG} y SS_{WG} se distribuye como la distribución F con parámetros df_{BG} y df_{WG}. La fórmula para representar la relación entre las distintas fuentes de varianza en el ANOVA de dos factores es:

$$SS_{Total} = SS_{BG} + SS_{WG}$$

$$F = (SS_{BG}/df_{BG}) / (SS_{WG}/df_{WG})$$

La estadística F calculada se evalúa para conocer la significancia con respecto a la distribución F teórica con df_{BG} y df_{WG}. Un valor resultante de F superior al valor teórico de F (para un nivel adecuado de significancia, por ejemplo $p<0,05$) indica la existencia de diferencias significativas entre las medias para el factor prueba, es decir Intergrupos.

Para ANOVA, la distinción entre grupos dependientes e independientes es importante. En el caso de los grupos dependientes, el factor aleatorio es un *factor con medidas repetidas* o un *factor intrasujetos*. En el caso de los grupos independientes (es decir, se comparan distintos grupos de sujetos), el factor se considera como un *factor intergrupos*. Si bien la lógica y la interpretación de ANOVA es idéntica entre estos dos tipos de diseño experimental, el cálculo difiere notablemente. En muchos casos, los experimentos requieren la inclusión *tanto* del factor Intergrupos *como* del factor con medidas repetidas.

En el cuadro 6.3.2-1 se muestra un ejemplo sencillo de la partición de la varianza. Las medias de los tres grupos (2, 3 y 4) difieren ostensiblemente. La suma de los cuadrados dentro de cada grupo es igual a 4. Si prescindimos de la pertenencia al grupo, la suma total de cuadrados basada en la media global (3) equivale a 24.

Cuadro 6.3.2-1 – Ejemplo de partición de la varianza

Datos primarios	**Grupo 1**	**Grupo 2**	**Grupo 3**
Observación 1	1	4	2
Observación 2	2	3	2
Observación 3	1	3	4
Observación 4	3	4	3
Observación 5	2	5	4
Observación 6	3	5	3

Estadísticas	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Total**
Media	2	4	3	3
Sumas de cuadrados (SS)	4	4	4	24

La variación total o SS Total (24) puede dividirse en variación Intragrupos (4+4+4 =12) y variación Intergrupos (24−(4+4+4) =12). A continuación, pueden calcularse las varianzas Intragrupos e Intergrupos:

Varianza intragrupos = $\frac{12}{3\times(6-1)}$ y Varianza intergrupos = $\frac{12}{3-1}$

Por lo tanto, un ANOVA a partir de estos datos y que tenga en cuenta grupos independientes daría los siguientes resultados (cuadro 6.3.2-2):

Cuadro 6.3.2-2 – Ejemplo de resultados de ANOVA en el caso de grupos independientes

Origen de la variación	SS	df	MS	F	Prob.
Intergrupos	12	2	6,00	7,50	0,0055
Intragrupos	12	15	0,80		
Total	24	17			

(SS = Sumas de cuadrados, df = Grado de libertad, MS = Media cuadrática o Varianza.)

La varianza Intragrupos suele conocerse como varianza del Error (o Cuadrado Medio del Error), porque el diseño actual no la puede prever fácilmente. La varianza Intergrupos se conoce como varianza del Efecto (o Cuadrado Medio del Efecto), y puede explicarse a partir de las diferencias en las medias entre los grupos.

En el caso de medidas repetidas, la varianza Intragrupos se corrige suprimiendo la variación debida a la variabilidad de los sujetos. A continuación, se calcula la estadística F como la relación entre la varianza Intragrupos y la varianza Residual (sin la varianza Intergrupos):

Cuadro 6.3.2-3 – Ejemplo de resultados de ANOVA en el caso de grupos dependientes

Origen de la variación	SS	df	MS	F	Prob.
Intergrupos	12	2	6,00	10,00	0,0041
Intragrupos/Entre sujetos	6	5	1,20		
Residual	6	10	0,60		
Total	24	17			

(SS = Sumas de cuadrados, df = Grado de libertad, MS = Media cuadrática o Varianza.)

En los ejemplos básicos anteriormente mencionados, solo hay un factor, pero normalmente se tiene en cuenta más de un factor. Otra ventaja de ANOVA con respecto a las pruebas t simples es que evalúa los efectos de interacción entre factores. Por lo general, existe interacción cuando un efecto modifica otro. Las interacciones bilaterales afectan a las interacciones entre dos factores principales. Existe interacción trilateral cuando un tercer factor modifica una interacción bilateral (hay varios tipos de interacciones en los distintos niveles del tercer factor); existe una interacción cuatrilateral cuando un cuarto factor modifica una interacción trilateral; y así sucesivamente.

Cuando en el experimento se tiene en cuenta más de una variable dependiente y cuando se quiere comprobar la hipótesis de que uno o varios factores afectan a todas las variables dependientes, puede recurrirse al Análisis Multivariante de la Varianza (MANOVA), asignando un valor *F* multivariado (*lambda* de Wilks) en lugar de un valor *F* univariado. La *lambda* de Wilks tiene en cuenta la correlación probable entre las variables dependientes para comprobar si el efecto es significativo.

6.3.3 Comparación post hoc de medias

Con ANOVA solamente se comprueba la hipótesis nula, H_0, a saber que "no existe diferencia entre las medias del factor de la prueba". Un valor F significativo solamente nos revela la existencia de una variación significativa entre las medias. Sin embargo, es imposible saber con certeza cuáles son las medias que difieren significativamente entre sí. Para averiguarlo, pueden emplearse pruebas post hoc. No obstante, las pruebas post hoc solamente se pueden utilizar cuando ANOVA ha puesto de manifiesto un efecto significativo. Si el valor F de un factor no es significativo, no se puede usar una prueba post hoc. Así se evita un uso demasiado liberal de las pruebas post hoc, es decir, el *archivo de datos*.

Las pruebas de rango múltiple, como la prueba de Duncan o la de Newman-Keuls, identifican subconjuntos homogéneos de medias que no difieren entre sí, mientras que las pruebas de comparaciones múltiples por pares permiten estudiar la diferencia entre cada par de medias y trazar una matriz donde los asteriscos indican unas medias notablemente distintas a un nivel alfa de 0,05.

Algunas de las pruebas de comparaciones múltiples por pares son:

- La prueba de Fisher de la Diferencia Menos Significativa, o LSD, que es la prueba más liberal (discriminatoria) de todas las pruebas post hoc. No obstante, no es adecuada para comparar más de tres medias.
- El método de Tukey de la Diferencia Honestamente Significativa, o HSD, que es la prueba post hoc más ampliamente utilizada. Para esta prueba, el número de observaciones de cada media debe ser idéntico.
- La prueba de Scheffé, que es uno de los métodos más flexibles y conservadores. Esta prueba puede utilizarse con un número desigual de observaciones.
- La prueba de Dunett, que se emplea para comparar un grupo de control con múltiples grupos de la prueba sin comparar los grupos de la prueba entre sí.

Sección 7

Diseño experimental de las evaluaciones de codificadores de voz (control de las fuentes de error)

7.0 Introducción

El objetivo de los métodos de escucha para evaluar códecs de voz es:

– controlar las variables del experimento (condiciones experimentales de los códecs, propiedades de las señales de entrada, elección de hablantes y material vocal);

– obtener respuesta de los sujetos (oyentes) sobre sus reacciones subjetivas relativas al códec de voz sometido a prueba (CuT).

Para lograr esos objetivos, es necesario:

– que la selección de las condiciones de los códecs cumpla una serie de requisitos (establecidos por la CE 16 del UIT-T);

– tomar precauciones en vista de que en las pruebas de escucha los oyentes son especialmente propensos a lo que se conoce como el efecto de «contexto»; es decir, sus criterios pueden estar muy influidos por la gama de calidad y nivel de escucha que se produce en las mismas pruebas. Por tanto, las pruebas deben abarcar una gama suficiente de condiciones que deben incluir las mejores encontradas en circunstancias favorables y las peores que pudieran encontrarse en circunstancias desfavorables. En general, la peor condición probada no sería admisible en la práctica. En consecuencia, es importante que las condiciones de circuito elegidas no comprendan algunas demasiado malas (es decir, condiciones que dan una nota de opinión de escucha mediocre incluso con los mejores niveles de escucha);

– tener en cuenta además que como los oyentes sólo realizan una tarea, por ejemplo, escuchar el contenido de una serie de frases cortas, es esencial que se minimice el factor de fatiga. Si el experimento rebasa la máxima duración recomendada de la prueba, deberá incluir más de una sesión.

Véanse las Recomendaciones UIT-T P.800 y UIT-T P.830.

7.1 Factores relativos al material fuente (procedimientos de grabación, factores de los hablantes)

Todos los ejemplos mencionados a continuación denotan el modo de controlar los procedimientos de grabación a fin de reducir al mínimo las variaciones debidas a las diferencias del material fuente producido en distintos laboratorios.

El material fuente deberá muestrearse al menos con una resolución de 16 bits, y se escogerá una frecuencia de muestreo (por ejemplo 8, 16 ó 32 kHz) adecuada para la anchura de la banda de frecuencias necesaria para la prueba; dicha anchura repercute notablemente en la calidad de los archivos fuente grabados.

Antes de aplicar el plan de procesamiento previsto, cada archivo fuente grabado (muestras del hablador y/o de música) debería delimitarse mediante una ventana de suavizado/subida y bajada progresivas de 100 ms.

Para extraer las secuencias de ruido deberá utilizarse una ventana de subida y bajada progresivas de 100 ms.

Se añaden más preámbulos de silencio para compensar el retardo del códec y alinear los límites de la entrada y los archivos procesados.

Las características del hablador (sexo, voz masculina, femenina o infantil, etc.) podrían repercutir en el rendimiento del códec y constituir un factor de variación de los resultados con arreglo al laboratorio de que se trate. Por lo tanto, los aspectos como el nivel promedio de la señal vocal (bajo, medio o alto), el tono de cada hablador, la entonación o el acento, entre otros, se consideran posibles factores que repercuten en la evaluación del sujeto.

7.2 Factores de laboratorio

Deben evitarse errores sistemáticos debidos a la falta de experiencia y conocimientos especializados adecuados sobre el laboratorio de escucha. El conjunto de mediciones debe realizarse con arreglo a un diseño de experimento debidamente planificado. Sus requisitos comprenden la aleatorización y atención adecuadas con respecto a la secuencia y agrupación correctas de las mediciones. En el supuesto de que esto se haya hecho así, los resultados deben tabularse sistemáticamente según los diversos factores y las combinaciones de los mismos.

Un aspecto importante es la reproductividad que cabe prever si el experimento completo se repitiese muchas veces con los mismos sujetos. Sin embargo, la composición del grupo de sujetos rara vez sigue siendo la misma por mucho tiempo, y será en todo caso distinta si las mediciones se repiten en otro laboratorio. Es preciso aceptar que todo análisis estadístico puede ser engañoso y que la conformidad o disconformidad entre los resultados obtenidos en laboratorios diferentes sólo puede verificarse mediante mediciones en varios laboratorios según un conjunto común de condiciones de prueba. Naturalmente, esas condiciones de prueba deberán estar definidas convenientemente y ser cuidadosamente respetadas en todos los laboratorios.

Se ha estimado conveniente comunicar información que pudiera utilizarse en apoyo de conclusiones o para explicar resultados inesperados de las pruebas.

La lista de elementos es la siguiente:

a) Característica sensibilidad/frecuencia del sistema transmisor.

b) Característica sensibilidad/frecuencia del sistema receptor.

c) Característica sensibilidad/frecuencia de otros componentes de la conexión, incluidos los códecs objeto de ensayo.

d) Medida del nivel de ruido y el espectro de frecuencia en la sala de grabación.

e) Medida del nivel de ruido y el espectro de frecuencia en la sala de escucha.

f) Descripción del grupo de sujetos que incluya edad, sexo, experiencia anterior y umbral audiométrico.

7.3 Factores idiomáticos

La voz se puede caracterizar por una amplia variedad de parámetros que normalmente se subdividen en características a largo plazo y a corto plazo. Las características a largo plazo no dependen, por lo general, del idioma, no así las características a corto plazo. Además, ambas características dependen generalmente del sexo del hablante.

Los requisitos lingüísticos que deben reunirse a fin de obtener unas características similares a las de la palabra natural son los siguientes:

– ha de existir una correspondencia suficiente entre las combinaciones de sonidos utilizadas en las listas de frases y la estructura de la palabra natural en un determinado lenguaje;

– ha de existir una correspondencia suficiente entre, por un lado, el espectro de frecuencias y la distribución de amplitudes del material de las listas, y, por otro, las correspondientes características de la palabra en un determinado idioma;

– no se admite que se distorsionen las características fonéticas inherentes a un determinado idioma.

Los idiomas tonales (como por ejemplo el chino y el coreano) y los no tonales (la mayoría de idiomas europeos), así como los hablantes masculinos y femeninos, han de incluirse en el plan de pruebas.

7.4 Factores relativos a los oyentes (selección y muestreo, instrucciones y capacitación)

Los sujetos deberán elegirse de entre una muestra representativa de usuarios del teléfono (hombres y mujeres) y deberán poseer una facultad auditiva normal. Es conveniente que no hayan participado en pruebas subjetivas similares al menos durante los seis meses precedentes. El número de sujetos empleados dependerá del grado de precisión deseado (30 sujetos puede ser un número aceptable, si bien ello depende del diseño del experimento).

Por lo general, se capacita a los oyentes (sujetos) para que utilicen las instalaciones y se les proporcionan instrucciones oportunas sobre el procedimiento de las pruebas. Las instrucciones deberán ser claras e inequívocas, y abarcar todos los aspectos de la prueba, incluidos los que se refieren al establecimiento de la conexión, el temario y el método de indicar la opinión. Antes de comenzar la prueba propiamente dicha los sujetos deberán estar plenamente informados sobre su procedimiento y la grabación de sus respuestas, entre otros aspectos. En cambio, no se les comunicará ni la naturaleza de las condiciones del circuito que se va a probar ni el tipo de degradación, etc., sino únicamente que van a evaluar la calidad de una conexión telefónica.

La fiabilidad intraindividuo ("en el mismo sujeto") se refiere a la concordancia entre calificaciones repetidas de un determinado sujeto con la misma condición de prueba.

La fiabilidad entre individuos ("entre sujetos") se refiere a la concordancia entre calificaciones de diferentes sujetos con la misma condición de prueba.

Con frecuencia se consideran tres factores relativos a los habladores, los oyentes y la interacción entre habladores y oyentes. Cabe suponer que esos factores son independientes de todos los demás.

7.5 Factores temporales/de orden, factores de equilibrio (generación de la aleatorización), contexto de las pruebas

Deben evitarse las anomalías en la secuencia temporal (condiciones aplicadas fuera de orden, mucha práctica o efectos de fatiga, o algo que haya afectado a la concentración de los sujetos, como los efectos relativos al momento o al orden de la presentación); el objetivo de utilizar distintos órdenes de aleatorización para cada grupo de sujetos es reducir al mínimo o equilibrar dichos efectos temporales/de orden.

Los resultados de las pruebas de significación se utilizan para determinar si hay características anómalas de los datos que pudieran presentar cualquier duda razonable de su validez; por ejemplo, un factor de Condiciones no significativo en un contexto de pruebas entrañaría casi ciertamente un mal funcionamiento de los aparatos o una selección extremadamente inadecuada de las condiciones del circuito.

Además de los efectos tanto sobre habladores como sobre oyentes existen muchos factores de degradación que han de tenerse en cuenta en la conexión global, incluido el códec sometido a prueba. Los principales son:

1) Pérdida de transmisión

2) Ruido del circuito

3) Ruido ambiente

4) Distorsión de atenuación/frecuencia

5) Efecto local

6) Distorsiones no lineales

7) Eco

8) Tiempo de propagación

9) Distorsión de retardo

10) Mutilación por conmutación de la voz

En la Recomendación UIT-T P.11 (1989) se proporciona información pormenorizada sobre esas degradaciones y sus efectos.

7.6 Duración de las secuencias/sesiones

La duración de cada sesión de pruebas deberá limitarse a 10-15 minutos (20 minutos como máximo) con objeto de que los oyentes no se cansen ni se fatiguen. Pueden presentarse más muestras de escucha después de un periodo de descanso adecuado. Se recomiendan al menos dos repeticiones (de presentaciones idénticas) a fin de verificar la coherencia de la fiabilidad relativa a los datos en bruto/al oyente.

En total, los sujetos que efectúen un experimento no deberán trabajar más de 2-3 horas, incluido el tiempo empleado en las explicaciones y la práctica/capacitación respecto de las condiciones previas.

7.7 Presentación de los resultados

Es importante reconocer que los resultados obtenidos en una prueba subjetiva pueden estar muy influidos por los criterios que apliquen los sujetos al evaluar cada una de las condiciones experimentales. Es, pues, necesario elegir cuidadosamente una escala de opiniones adecuada, y especialmente el número y descripción de las categorías individuales. Debe señalarse también que puede resultar difícil traducir los nombres de dichas categorías a otros idiomas sin dejar de conservar la misma relación entre categorías que en el idioma original. No es posible, por lo tanto, ofrecer una orientación de carácter universal sobre la elección de las escalas de opinión.

En las Recomendaciones UIT-T P.800 y UIT-T P.830 (anexo B) se indica que los criterios comúnmente empleados para las evaluaciones subjetivas son la calidad de funcionamiento, las preferencias personales y el esfuerzo de escucha, y se proporcionan las escalas de opinión asociadas a dichos criterios.

Una de las finalidades del análisis es determinar si se cumplen los requisitos/objetivos establecidos para los códecs sometidos a prueba, para los que han de proporcionarse cuadros y figuras pertinentes para la presentación de los resultados, incluidas las notas promedio y todos los datos estadísticos necesarios para la planificación de las pruebas.

7.7.1 Herramientas de análisis

Se ha de poner cuidado en el análisis de los datos relativos a un experimento subjetivo. En la Recomendación UIT-T P.800 se describe el análisis estadístico y las herramientas necesarias para llevar a cabo ese análisis de forma adecuada, dependiendo del método escogido (pruebas de detectabilidad de respuesta cuantificada, evaluación por categorías absolutas, índices por categorías de degradación, evaluación por categorías de comparación o método del umbral para comparar sistemas de transmisión con un sistema de referencia). Por ejemplo, las sensibilidades pueden cuantificarse mediante pruebas estadísticas de comparación múltiple. Si es preciso efectuar una comparación *a posteriori* de los circuitos, puede aplicarse eficazmente la prueba de Tukey de la diferencia honradamente significativa (HSD). La prueba HSD está diseñada para efectuar todas las comparaciones por pares entre las medias y determinar el grado de significación de las diferencias en los valores medios.

También ha de ponerse cuidado en el análisis de los datos procedentes de un experimento CCR. Dado que la mitad de las pruebas, para cualquier condición de las mismas, se presenta en orden (no procesado, procesado) y la otra mitad se presenta en orden opuesto, un sencillo promedio de las notas numéricas debe proporcionar un CMOS de aproximadamente 0 para todas las condiciones. Es necesario recodificar los datos primarios. Si el orden de presentación es (procesado, no procesado), entonces el signo de la nota numérica debe invertirse. Las notas recodificadas pueden utilizarse para calcular la CMOS, las desviaciones típicas, etc.

Los resultados se presentan, pues, en el orden de (no procesado, procesado). Puede también realizarse un análisis de varianza apropiado u otras pruebas estadísticas sobre las notas recodificadas. No obstante, no cabe deducir que las notas de opinión sobre las comparaciones representen una escala de intervalos lineal. Por consiguiente, tal vez necesiten aplicarse en su lugar estadísticas de escalas ordinales.

7.7.2 Recopilación de datos

Deberán preverse formularios de respuesta para que los sujetos inscriban sus opiniones sobre cada condición probada, así como las respuestas a cualesquiera otras preguntas que se les formulen. Con el fin de minimizar los efectos de aprendizaje o asociación, no deberá consignarse más de una respuesta en cada formulario. Otro procedimiento para recabar los datos en bruto consiste en utilizar técnicas de registro automático de datos mediante un dispositivo de pulsador provisto de una pantalla y conectado a un computador.

Los laboratorios de pruebas subjetivas cualificados han desarrollado sofisticados programas informáticos para realizar complejos experimentos mejorando la presentación del material procesado y la obtención de respuestas relativas a cada presentación, con objeto de facilitar el análisis posterior de los datos en bruto y la presentación de los resultados.

7.7.3 GAL

En las rondas de pruebas subjetivas coordinadas a nivel internacional, cada laboratorio de pruebas podrá remitir los datos en bruto a un Laboratorio de Análisis Mundial designado (GAL) a tal efecto.

Los resultados serán comunicados por cada Laboratorio de Escucha y/o el Laboratorio de Análisis Mundial, que se encargará del análisis del conjunto general de los datos en bruto de la manera detallada a continuación:

1) calcular y tabular las notas medias de opinión, las desviaciones típicas y los intervalos de confianza;

2) llevar a cabo las pruebas estadísticas pertinentes e informar si se han cumplido los requisitos y objetivos.

Sección 8

Ejemplos de planes de pruebas/resultados sobre codificadores

En la página electrónica <http://ftp3.itu.int/av-arch/audio-site/tpref/> figuran ejemplos pormenorizados sobre:

- el Plan de calidad y el Análisis global;
- el Plan de procesamiento;
- la Fase de caracterización.

El conjunto de archivos disponible en <http://ftp3.itu.int/av-arch/audio-site/tpref/> contiene el historial de este ejercicio del UIT-T, basado en la Cuestión 7/12 y en la Cuestión 10/16, a saber:

- Plan de pruebas de evaluación de la calidad del módulo de procesamiento previo o posterior del actual códec UIT-T G.711.
- Plan de procesamiento del UIT-T G.711-Plus (versión definitiva que contiene el plan de procesamiento utilizado durante las pruebas de escucha para evaluar la calidad del módulo de procesamiento posterior del UIT-T G.711).
- Plan de pruebas de evaluación de la calidad de la fase de caracterización del UIT-T G.711.1.
- Plan de procesamiento de la fase de caracterización del UIT-T G.711.1.
- Plan de procesamiento revisado de la fase de optimización/caracterización del UIT-T G.711.1.
- "Archivos de aleatorización" utilizados para presentar estímulos a los oyentes en orden exacto y controlado.
- Algunos informes de las reuniones del Grupo de Relator del UIT-T Q10/16 (la Cuestión Q10/16 del UIT-T abordaba el ejercicio sobre ese códec).
- La declaración de coordinación entre la CE 12 y la CE 16 que figura en los anexos, incluidos los resultados de pruebas subjetivas y el correspondiente análisis de datos en bruto.

Los "ejemplos de planes de pruebas/resultados sobre codificadores" mencionados anteriormente ilustran un procedimiento definitivo que ha utilizado el UIT-T durante muchos años para evaluar, escoger y caracterizar nuevos algoritmos de codificación.

Anexo A

Principios de procesamiento (Biblioteca de herramientas de soporte lógico)

A.1 Ámbito de aplicación

En el presente Anexo se describen los principios de procesamiento basados en la Biblioteca de herramientas de soporte lógico (STL) del UIT-T y su uso práctico.

A.2 Antecedentes

En julio de 1990, la Comisión de Estudio XV del entonces CCITT (el predecesor del UIT-T) decidió crear, en colaboración con el Grupo de Expertos sobre la calidad de las señales vocales (SQEG) de la CE XII, un grupo denominado "Grupo de usuarios de herramientas de soporte lógico" (UGST). Dicho Grupo se instauró, junto con la definición de un conjunto de herramientas de soporte lógico, con el fin de elaborar normas de codificación de señales vocales que permitirían no sólo contar con un conjunto común de herramientas que funcionarían de manera idéntica o similar en distintas plataformas informáticas, sino también resolver varios problemas relacionados con la interpretación exacta de algunas Recomendaciones pertinentes de la UIT.

El primer conjunto de herramientas producido por el Grupo, la Biblioteca de herramientas de soporte lógico (STL) del UIT-T, fue publicado por el UIT-T (el entonces CCITT) en 1992 en el marco de la Recomendación UIT-T G.191. Tras esa primera publicación, la Comisión de Estudio 15 del UIT-T aprobó otra publicación en mayo de 1996, que se denominó UIT-T 96. La STL96 permitió introducir mejoras sustantivas y nuevos elementos en el UIT-T 92. En noviembre de 2000, la Comisión de Estudio 16 del UIT-T aprobó una versión actualizada de la STL, la STL2000. En julio de 2005, se publicó otra versión actualizada de la STL, la STL2005. Esa última versión permitió corregir errores de programación, introducir cambios (tales como ponderaciones de operadores básicos de acumulador de 32 bits) y añadir herramientas nuevas (conjunto alternativo de operadores básicos en aritmética de 40 bits, nuevos filtros FIR para el procesamiento de señales en banda superancha, una herramienta de reverberación, una herramienta de medición de la respuesta en frecuencia y una herramienta de truncamiento del tren de bits). La última publicación es la STL2009, que fue aprobada en marzo de 2010 y contiene varias herramientas nuevas y actualizadas que se desarrollaron con arreglo a las últimas actividades de normalización del códec de señales vocales y de audio y, en particular, para el audio en toda la banda y las señales estereofónicas. También se añadieron herramientas destinadas a resolver el problema de la inserción de errores en los trenes de bits estructurados en capas de los codificadores jerarquizados, así como herramientas asociadas con los operadores básicos destinadas a evaluar la complejidad del ROM de programa para el código de coma fija y otra destinada a evaluar la complejidad (incluido el ROM de programa) de las implementaciones de coma flotante. Por último, y sin ser menos importante, en la STL se ha incluido una implementación de referencia, esperada desde hace mucho tiempo, del código fuente C UIT-T G.728.

Con el tiempo, la STL se ha utilizado para definir normas de codificación de audio y voz en el marco del UIT-T y otras organizaciones de normalización (SDO), tales como el ETSI, el 3GPP y el 3GPP2.

En la Recomendación UIT-T G.191 figuran los términos y condiciones relativos a la utilización de la STL del UIT-T.

Cabe señalar que la STL del UIT-T es un conjunto de herramientas dinámicas que evolucionan con el tiempo. Cualquier información detallada en el presente documento puede no ser válida como consecuencia de las modificaciones introducidas en él, en cuyo caso debería consultarse el manual adjunto al paquete.

A.3 Motivación

La Biblioteca de herramientas de soporte lógico del UIT-T se creó con el fin de elaborar un marco de herramientas de soporte lógico relacionadas con la codificación de señales vocales que permitiese a las distintas organizaciones realizar sus actividades en un terreno común.

La creación de la STL tenía dos objetivos operacionales básicos, a saber: desarrollar módulos de herramientas (en contraposición de los programas independientes) y un elevado grado de portabilidad. La decisión de aplicar modelos de biblioteca en lugar de una colección de programas independientes tenía por finalidad ofrecer a los usuarios la flexibilidad necesaria para que pudieran ya sea utilizar los programas de demostración (o "demo") proporcionados o escribir su propio código para adaptarlo a una aplicación en concreto. La utilización de un subconjunto del lenguaje ANSI-C y varios convenios sobre interfaces de módulo permitió lograr la portabilidad deseada. Por consiguiente, el código de la STL compila y puede utilizarse correctamente con una intervención mínima del usuario (en su caso) al menos en las plataformas Microsoft Windows, varios sistemas similares a Unix (incluido Cygwin, un puerto Windows del entorno similar a Unix) y MS-DOS.

Otro de los objetivos de la creación de la STL era aplicar los módulos ya sea para emular sistemas de soporte físico ya existentes o para dar cabida a aplicaciones de soporte físico y lógico idénticas. El procedimiento de emulación, que está diseñado para reflejar las condiciones de utilización futuras, se describe en los planes de procesamiento (véase la Sección A.4).

Otra finalidad de la STL era contribuir a realizar evaluaciones de la calidad de manera cuidadosa y subjetiva proporcionando herramientas con el fin de procesar referencias y/o soportes en el plan de prueba de la calidad (véase la Sección A.5).

Por otra parte, existen herramientas que permiten realizar una especificación exacta de bits de las normas de codificación de audio y voz del UIT-T así como una evaluación objetiva del funcionamiento de las mismas (véase la Sección A.6). En la Sección A.7 figura un ejemplo de marco que puede utilizarse para implantar funciones de laboratorio de acogida a través de la STL.

A.4 Herramientas de soporte lógico destinadas a emular cadenas de transmisión

En esta sección se presenta una visión de conjunto de las herramientas que se utilizan para simular las condiciones de utilización reales. A fin de reflejar las condiciones de utilización futuras, la cadena de comunicación se emula tal y como se describe en los planes de procesamiento. En esos planes de procesamiento se establece la manera en que se debe preparar el material de audio (voz y otros sonidos, como música, canciones publicitarias, ruidos de fondo, etc.) para simular las condiciones de utilización reales. En dichos planes, se describen todas las fases de procesamiento de una cadena de comunicación, desde la preparación del material de audio original (discurso, ruidos de fondo,…) hasta la preparación de la señal de entrada y el dispositivo de escucha. En ellos también se especifica qué herramientas de soporte lógico deben utilizarse y cómo deben utilizarse, y se indica, en forma de diagramas, las fases de procesamiento que se necesitan. En las Secciones A.4.1 y A.4.4 se expone la manera en que las señales de audio se condicionan en los lados de emisión y de recepción, respectivamente, mientras que en la Sección A.4.3 se indica la manera en que se simulan las condiciones de transmisión. Además, en la Sección A.4.1 se enumeran algunas herramientas que se utilizan para manipular las señales de entrada.

A.4.1 Manipulación de las señales de entrada

Existen varias herramientas que sirven para manipular las señales de entrada, las cuales se describen en el Apéndice A del manual de la STL; por ejemplo, "astrip" divide un segmento en palabras de 16 bits con una posible disposición en ventana para redondear los bordes, "sb" desplaza bytes de ficheros orientados a palabras, lo que puede resultar necesario para utilizar ficheros creados en sistemas con diferentes órdenes de almacenamiento de bytes, "concat" concatena ficheros en modo binario, etc.

A.4.2 Condicionamiento de las señales de entrada

El condicionamiento de las señales de entrada abarca el ajuste de la frecuencia de ejemplo y/o nivel correctos, la adición de ruidos de fondo y/o reverberación, y el filtraje encaminado a reproducir las características del terminal emisor. En una primera etapa, las herramientas se desarrollaron para las señales de un solo canal (monofónicas). En la STL2009, se incluyó la herramienta de procesamiento estereofónico STL para permitir el acondicionamiento de las señales estereofónicas con esas herramientas.

A.4.2.1 Muestreo ascendente y descendente

Las señales de audio originales suelen grabarse en una frecuencia de muestreo de mayor resolución. Los filtros de alta calidad permiten modificar la velocidad de muestreo y minimizar a la vez la distorsión de fase y de amplitud de la señal de entrada. En el Capítulo 10 "RATE-CHANGE: Up-and down-sampling module" del manual de la STL figura documentación detallada sobre esta herramienta de muestreo ascendente y descendente.

A.4.2.2 Características del terminal emisor

Ponderación en la banda estrecha

En lo que respecta a las comunicaciones vocales en banda estrecha, la ponderación del sistema intermedio de referencia (SIR) corresponde a una característica de filtrado en paso de banda cuya plantilla se menciona en la Recomendación UIT-T P.48. La ponderación del SIR señalada en la Recomendación UIT-T P.48 se basa en los microteléfonos medios de banda estrecha que se utilizan en la RTPC analógica. El SIR modificado (SIRM) es más representativo de las conexiones a través de servicios digitales.

Las velocidades de muestreo disponibles para el filtro normal del SIR en el lado transmisión en virtud de la Recomendación UIT-T P.48 son de 8 y 16 kHz. El filtro del MSIR en el lado transmisión está disponible para muestreo a 16 kHz y 48 kHz. La ponderación de la banda estrecha se describe en el punto 10.1.2 del manual de la STL.

Ponderación en la banda ancha

En el caso de las comunicaciones vocales en banda ancha (50-7 000 Hz), las especificaciones para los lados de emisión y de recepción se describen en la Recomendación del UIT-T P.341. En la STL se ha incorporado una aplicación de la plantilla para el lado de emisión, así como un filtro que estimula la característica de respuesta de entrada de determinados terminales móviles para los datos muestreados a 16 kHz. La ponderación de la banda ancha se describe en el apartado 10.1.3 del manual de la STL.

Ponderación en la banda superancha

En cuanto a los datos muestreados a 32 kHz, en la publicación STL2005 se ha incluido un filtro en paso de banda [50 Hz – 14 kHz] basado en la característica de respuesta de entrada de determinados terminales de videoconferencia en banda superancha. La plantilla se ha establecido como continuación de la

Recomendación UIT-T P.341. La ponderación de la banda superancha se describe en el apartado 10.1.4 del manual de la STL.

Ponderación en toda la banda

En lo que respecta a los datos muestreados a 48 kHz, se ha diseñado e incluido en la STL2009 un filtro FIR plano de banda de paso [20 Hz – 20 kHz]. La ponderación de la banda superancha también se describe en el apartado 10.1.4 del manual de la STL.

Cabe señalar que, a diferencia de los casos de banda estrecha y banda ancha, actualmente ninguna recomendación del UIT-T se refiere a las características entrada/salida para los terminales en banda superancha y en toda la banda. Está previsto abordar esas especificaciones en el marco de la Cuestión 3/12 ("Características de transmisión de terminales vocales para redes fijas con conmutación de circuitos, redes móviles y redes (IP) con conmutación de paquetes") en el periodo de estudios 2009-2012.

A.4.2.3 Ponderación del ruido

Se dispone de los dos filtros de ponderación del ruido siguientes: los filtros sofométricos y los filtros de ponderación Δ_{SM}. Los filtros sofométricos fueron diseñados para las transmisiones de voz muestreadas a 8 kHz, y los filtros Δ_{SM} para las muestreadas a 16 kHz. La ponderación del ruido también se describe en el apartado 10.1.5 del manual de la STL.

A.4.2.4 Ajuste de nivel

En la Recomendación UIT-T P.56 se describe la especificación para la medición del nivel activo de las señales de voz. Además de esa Recomendación, también figura información adicional en el apartado "Measurement of Speech" que figura en el Manual sobre Telefonometría del UIT-T. En el algoritmo señalado en la Recomendación UIT-T P.56 se toman muestras de una señal en la anchura de banda vocal y se calcula su nivel vocal activo. La herramienta de ajuste de nivel se describe en el capítulo 13 "SVP56: The Speech Voltmeter" del manual de la STL.

A.4.2.5 Reverberación

En algunas aplicaciones manos libres (por ejemplo, de videoconferencia o audioconferencia), el sonido que se recibe suele estar compuesto de sonido directo del interlocutor y sus componentes reverberados. Ese efecto de reverberación equivale a la modificación de la señal vocal como consecuencia de la respuesta acústica del recinto. Por consiguiente, ese efecto puede simularse mediante la convolución de una respuesta de impulso en una sala real con señales anecoicas. La herramienta de reverberación de la STL permite realizar esa convolución mediante respuestas de impulso en salas de conferencia típicas. Dicha herramienta se describe en el capítulo 14 "ITU-T Reverberation tool" del manual de la STL.

A.4.2.6 Acondicionamiento de las señales estereofónicas

Para lograr el desarrollo y la caracterización de algoritmos estereofónicos, se necesitan señales estereofónicas adecuadamente acondicionadas. A fin de reutilizar el conjunto vigente de herramientas de procesamiento de un solo canal de la STL para acondicionar las entradas y las salidas en una cadena de procesamiento estereofónico, las herramientas de procesamiento estereofónico de la STL proporciona las funciones básicas de intercalado de ficheros de un solo canal, división de ficheros estereofónicos en ficheros de un solo canal y creación de señales específicas de mezclado descendente de un solo canal a partir de una determinada entrada estereofónica. La herramienta de procesamiento estereofónico se describe en el capítulo 17 "ITU-T Stereo processing tool" del manual de la STL.

A.4.3 Condiciones de transmisión

Se pueden simular diversas condiciones de transmisión, a saber: borrado de tramas, errores en los bits, conexiones en cascada y truncamientos de tren de bits para estructuras integradas.

A.4.3.1 Errores de canal

Para estudiar el comportamiento de los sistemas de transmisión digital y los equipos en condiciones de error, se generan patrones de errores y se insertan errores en el tren de bits. Esto obliga a crear modelos para los canales de transmisión y algoritmos conexos de generación de errores. Se cuenta con bits erróneos y con generadores de borrado de tramas. Para cada tipo de error, se generan errores aleatorios y a ráfagas. En la STL2009, se han desarrollado modelos de borrado de tramas para planes EV.

También existen herramientas destinadas a insertar errores en los trenes de bits utilizando los patrones de errores generados, con miras a estimular los sistemas de transmisión y los equipos en condiciones de error. Todas las herramientas relacionadas con los errores de canal se describen en el capítulo 11 "EID: Error Insertion Device" del manual de la STL.

A.4.3.2 Conexiones en cascada

Las conexiones de códecs asíncronas en cascada se simulan con un filtrado no lineal de fase (retardo de grupo no constante).

A.4.3.3 Truncamiento de trenes de bits

Los códecs escalonables (o integrados) constituyen una técnica de codificación extremadamente flexible que se caracteriza por un tren de bits multicapa con una capa de núcleo que ofrece la calidad mínima y capas superiores que mejoran la calidad aumentando la velocidad binaria. Cualquier componente de la cadena de comunicación puede ajustar la velocidad binaria entre los valores mínimo y máximo. Para hacer frente a la congestión de red, la velocidad binaria puede ajustarse trama por trama. La herramienta que sirve para realizar el ajuste de la velocidad binaria efectúa un truncamiento directo del tren de bits a la velocidad correcta dividiendo bits. La herramienta de truncamiento del tren de bits se describe en el capítulo 15 "ITU-T Bitstream truncation tool" del manual de la STL.

A.4.4 Acondicionamiento de la señal de salida

El acondicionamiento de la señal de salida abarca el ajuste al nivel de escucha y/o la frecuencia de muestreo correctos, y el filtrado para reproducir las características del terminal receptor. No cabe duda de que la herramienta "astrip" descrita en el apartado A.4.1 resulta útil para modificar la longitud del fichero de la señal vocal de salida.

A.4.4.1 Ajuste de nivel

Véase el apartado A.4.2.4.

A.4.4.2 Muestreo ascendente y descendente

Las señales de audio suelen escucharse a una frecuencia de muestreo con una resolución diferente de la frecuencia de muestreo de la salida de procesamiento. Los filtros de alta calidad permiten modificar la velocidad de muestreo a la vez que minimizar la distorsión de fase y de amplitud de la señal de audio. En el capítulo 10 "RATE-CHANGE: Up- and down-sampling module" del manual de la STL figura documentación detallada sobre la herramienta de muestreo ascendente y descendente.

A.4.4.3 Características del terminal receptor

El filtro del SIR modificado en el lado de recepción está disponible para las señales muestreadas a 8 kHz y 16 kHz. Para las señales audio de mayor anchura de banda, pueden utilizarse filtros en paso de banda. Las ponderaciones para las señales a 8 kHz y 16 kHz se describen en los apartados 10.1.2 y 10.1.3 del manual de la STL, respectivamente.

A.4.5 Manipulación de las señales de salida

Véase el apartado A.4.1.

A.5 Herramientas de soporte lógico para el tratamiento de referencias para pruebas subjetivas

La calidad de funcionamiento suele evaluarse a través de pruebas subjetivas de escucha reglamentarias. En esas pruebas, la calidad del codificador en las condiciones de prueba mencionadas en el mandato se evalúa con el fin de compararla con la calidad de los codificadores de referencia. Si bien se pueden utilizar ejecutables para tratar esas condiciones de referencia, suelen preferirse los códigos de fuente C, que proporcionan una especificación binaria exacta de coma fija. La STL incluye los códigos C de todas las normas del UIT-T que no están disponibles como parte normativa de las normas (véase el apartado A.5.1). Las pruebas subjetivas también requieren soportes de prueba destinados a calibrar la prueba y abarcar la gama de notas. En el apartado A.5.2 se reseñan las herramientas de la STL utilizadas para generar dichos soportes.

A.5.1 Codificadores de referencia

Desde 1995, y la normalización de las Recomendaciones UIT-T G.729 y UIT-T G.723.1, el código ANSI-C constituye una parte integrante de las recomendaciones del UIT-T relativas a la codificación de audio y voz, a diferencia de lo que ocurría con las normas anteriores. En el caso de las antiguas normas, relativas a las Recomendaciones UIT-T G.711, UIT-T G.726, UIT-T G.727, UIT-T G.722 y UIT-T G.728, sus códigos C se encuentran en la biblioteca de herramientas de soporte lógico. También se incluye el algoritmo RPE-LTP de 13 kbit/s del GSM de velocidad completa (GSM Rec. 06.10).

Igualmente, desde mediados del decenio de 1990, el procedimiento de ocultación de la pérdida de paquete (*Packet Loss Concealment*, PLC) se ha requerido y normalizado con respecto a la Recomendación principal. En cuanto a las normas del UIT-T más antiguas, la PLC fue añadida ulteriormente, cuando comenzaron a utilizarse los códecs en las aplicaciones VoIP ante las condiciones de borrados de tramas. El código de fuente C de la PLC de ejemplo para la Recomendación UIT-T G.711, que se describe en el presente Apéndice I, se proporciona en la STL. Asimismo, también se facilita en la STL el código de fuente C de la PLC correspondiente a la Recomendación UIT-T G.728, como se define en el Anexo I a la Recomendación UIT-T G.728. En relación con los procedimientos de PLC que figuran en la Recomendación UIT-T G.722, se han añadido funcionalidades básicas de referencia de la ocultación de la pérdida de paquete en la herramienta de soporte lógico de la Recomendación UIT-T G.722 revisada, que se publicó en la STL2009.

A.5.2 Anclajes de pruebas subjetivas

A.5.2.1 Aparato de referencia para ruido modulado

El aparato de referencia para ruido modulado (MNRU), que en un primer momento tenía por objeto evaluar la calidad de los sistemas de codificación de forma de onda MIC log, se ha utilizado en las pruebas de evaluación de la calidad subjetiva en numerosos procesos de normalización de códecs de voz de banda

estrecha y de banda ancha en el marco del UIT-T y otras SDO (por ejemplo, el ETSI y el 3GPP). Los MNRU se emplean fundamentalmente en las pruebas diseñadas por medio de las metodologías de prueba previstas en la Recomendación UIT-T P.800 (por ejemplo, ACR y DCR). Las herramientas MNRU se describen en el capítulo 12 "Duo-MNRU: The Dual-mode Modulated Noise Reference Unit" del manual de la STL.

A.5.2.2 Filtrado de paso bajo y en paso de banda

En el caso de las señales de audio de mayor anchura de banda, se utilizan metodologías de prueba del UIT-R (previstas por ejemplo en las Recomendaciones UIT-R BS.1116, UIT-R BS.1285 y UIT-R BS.1534), en lugar de adaptar las que figuran en la Recomendación UIT-T P.800. En esas pruebas, la gama de calidades se abarca con varios filtrados de paso bajo de las señales de audio originales: la calidad aumenta con la frecuencia de corte. También se utilizan filtros en paso de banda. La STL proporciona un conjunto grande de esos filtrados de paso bajo con distintas frecuencias de corte (1,5 kHz, 3,5 kHz, 7 kHz y 10 kHz, 12 kHz, 14 kHz y 20 kHz (para una frecuencia de muestreo de 48 kHz)). También existe un filtro FIR de banda de paso plano de 20-20 kHz. Los filtros que limitan la banda se describen en el apartado 10.2.1 del manual de la STL.

A.6 Herramienta de soporte lógico para la especificación de los códecs de voz y audio del UIT-T y evaluación del funcionamiento objetivo

A.6.1 Especificación de los códecs de voz y audio del UIT-T

Desde mediados del decenio de 1990, en la especificación de los codificadores de voz y audio del UIT-T se utiliza el código C con exactitud de bits y coma fija tanto del codificador como del decodificador como método normativo de descripción del algoritmo, con el fin de asegurar la calidad de las implementaciones y evitar ambigüedades en las mismas. En la descripción de coma fija se utiliza un conjunto de operadores básicos que representa el conjunto de instrucciones habitualmente disponible en los procesadores de señales digitales. Además de los operadores de 16 y 32 bits y los operadores de flujo de control, existe otro conjunto de operadores de 40 bits. Los operadores básicos se describen detalladamente en el capítulo 18 "`BASOP' ITU-T Basic Operators" del manual de la STL.

A.6.2 Evaluación de la complejidad

Los operadores básicos y los operadores de flujo de control que se utilizan para especificar las recomendaciones del UIT-T, tienen ponderaciones de complejidad que reflejan la evolución de las capacidades del procesador. La especificación de las normas con dichos operadores ponderados permite evaluar la complejidad de cálculo del codificador. El número total de instrucciones necesarias por trama se calcula sumando el número total de operaciones ponderadas a fin de obtener la complejidad en millones ponderados de operaciones por segundo (WMOPS). Si bien los códecs de voz y audio del UIT-T deben especificarse en términos de coma fija, siempre resulta útil calcular la complejidad de una implementación de coma flotante. Para desarrollar un códec se puede utilizar una herramienta de cómputo de los operadores de coma flotante. Dicha herramienta se describe en el apartado 18.7 del manual de la STL.

A.6.3 Evaluación de la ROM de programa

Para calcular el tamaño de la ROM de programa existe una herramienta llamada "basop_cnt", que cuenta el número de operadores básicos y llamadas de función en un fichero de fuente C. Esa herramienta también permite contar el número de llamadas a funciones definidas por el usuario. La suma de esas dos cifras da una estimación del PROM necesario para ese fichero de fuente C. Cabe señalar que la RAM y la ROM de datos

deben calcularse por otros medios. Esa herramienta se describe en detalle en el apartado 18.6 "Program ROM estimation tool for fixed-point C Code".

A.6.4 Evaluación de la respuesta en frecuencia

Para poder evaluar la anchura de banda efectiva del códec, se creó una herramienta de medición de respuesta en frecuencia para la STL2005, que se actualizó en la STL2009. En el capítulo 16 "ITU-T frequency response measurement tool" del manual de la STL también se ofrecen pautas de orientación sobre las señales de entrada de prueba adecuadas para evaluar la respuesta en frecuencia.

A.7 Utilización de la STL para implementar el tratamiento asociado al diseño de pruebas

En el presente apartado se da un ejemplo de cómo se puede utilizar la STL, de forma modular, para implementar el tratamiento de ficheros basándose en el diseño de pruebas subjetivo. Cabe señalar que el procedimiento que se describe a continuación, si bien no constituye la única manera de implementar las funciones de procesamiento de laboratorios centrales, puede considerarse un ejemplo práctico para los usuarios que intentan implementar una partiendo de cero.

Ese procedimiento modular, que ya se ha utilizado con éxito para implementar varias funciones de laboratorios centrales del UIT-T, se basa en los tres niveles de programas siguientes: programas de alto nivel (derivados directamente de un diseño experimental y puestos en marcha en una plataforma común); programas de nivel medio (que cuentan, a nivel de usuario, con una interfaz común para las distintas funciones de procesamiento que deben realizarse, funcionan como un "soporte intermedio" o una capa de adaptación, y se ejecutan en una plataforma común), y los programas de bajo nivel (que implementan el procesamiento real del material vocal y se ejecutan en plataformas específicas).

En los apartados que figuran a continuación se describe un ejemplo simplificado de ese procedimiento, para facilitar la lectura. En la dirección http://ftp3.itu.ch/av-arch/audio-site/tpref/processing-framework-example.zip se puede descargar una versión completa y en funcionamiento de ese ejemplo, y experimentar con él.

A.7.1 Diseño experimental

Se implementará el siguiente diseño de prueba ficticia. En el plan de prueba subjetiva se especifican dos experimentos, cada uno de los cuales se pondrá en marcha en distintos laboratorios de escucha: C y D. En el experimento 1 se pone a prueba el códec X de prueba, 32 kbit/s UIT-T G.726, directo (a saber, sin procesamiento) y una condición MNRU (18 dB) para una señal vocal de entrada ponderada en plano. El códec X se pone a prueba para borrados de tramas de un 3% ("pérdida de paquetes"), niveles reducidos libres de errores y nominales, y 2 en cascada, y la Recomendación UIT-T G.726 se pone a prueba a nivel nominal para 4 cascadas asíncronas. En el experimento 2 se pondrá a prueba el códec X para ruido de fondo como murmullo según el SIR modificado y se contará con una referencia MNRU (15dB) y 2 en cascada 24 kbit/s del UIT-T G.726. En ambos experimentos, se utilizará sólo una muestra de voz para cada uno de los 4 hablantes. En el plan de procesamiento se especifican tres fases en los trabajos de los laboratorios centrales, a saber: preprocesamiento, procesamiento y postprocesamiento. En el preprocesamiento, los ficheros recibidos del laboratorio de escucha se concatenan con un intervalo inicial de silencio de 5 segundos (para permitir la convergencia del códec), y en el experimento 2, se añade murmullo. A continuación, los ficheros de voz se submuestrean de 16 kHz a 8 kHz. En la fase de procesamiento, se procesan las condiciones de prueba específicas a través de los circuitos definidos en el plan de prueba de procesamiento. Por último, en la fase de postprocesamiento, los ficheros procesados se sobremuestrean a 16 kHz, se filtran a través de un filtro de recepción del SIR modificado y se vuelven a separar en ficheros de pares de frases con la señal de voz procesada. En el ejemplo que presentamos, todos los códecs, a excepción del Códec "X", se ejecutan en una

plataforma Windows; el Códec X se ejecuta en Linux (a efectos de ilustrar el tratamiento en múltiples plataformas).

A.7.2 Programas de alto nivel

Conviene que los programas de alto nivel sean lo más genéricos y legibles por seres humanos posible. Esto permitirá implementar y depurar rápidamente un conjunto de condiciones experimentales. En este caso ficticio, los programas de alto nivel se llaman proc-exp1a y proc-exp2a, y las porciones pertinentes de los mismos deberían tener el aspecto siguiente:

```
# Experiment 1
…
if test $PLAT = "nt"
then
 mnru prefile prcfile.c01 18
 direct prefile prcfile.c02
 g726 prefile prcfile.c03 32 0
 tandem prefile prcfile.c04 g726 / g726 / g726 / g726
elif test $PLAT = "linux"
then
 cutx prefile prcfile.c05 0
 cutx prefile prcfile.c06 -10
 tandem prefile prcfile.c07 cutx / cutx
 cutx prefile prcfile.c08 RFER3%

fi…
```

```
# Experiment 2
…
if test $PLAT = "nt"
then
 mnru prefile prcfile.c01 15
 g726 prefile prcfile.c02 24 0
 tandem prefile prcfile.c03 g726 24 / g726 24
elif test $PLAT = "linux"
then
 cutx prefile prcfile.c04 0
fi...
```

A.7.3 Programas de nivel medio

Los programas de nivel medio invocados en los programas de alto nivel son MNRU, directos, g726, en cascada y cutx. Los programas MNRU y directos son muy sencillos, ya que cuentan con un número de parámetros muy limitado. Los programas g726 y cutx constituyen en sí una llamada de capa superior para las funciones de núcleo UIT-T G.726 y Códec X implementadas en g726-core y cutx-core. La implementación se realiza de esta manera porque los diagramas de procesamiento tienen diferentes bloques para las conexiones en cascada y las que no están en cascada. El programa en cascada es un programa recurrente que llama a otros programas de nivel medio "*-core", según sea necesario. Debe reconocer distintos tipos de conexión en cascada e implementarlos según los bloques constitutivos especificados. A continuación figura un ejemplo de programa de nivel medio para el procesamiento en cascada y UIT-T G.726 para bash en Linux.

A.7.3.1 Procesamiento UIT-T G.726

```
#!/bin/bash
# Processing by the G.726 codec
#
# Usage:
# ${0##*/} InFile OutFile Bitrate
# Parameters:
# InFile: Input sample file
# OutFile:    Output bitstream file
# Bitrate:    40, 32, 24, 16 kbit/s

# Default variables
input=$1 output=$2

# If not script generation, force eval of shell functionality
if test -z $x
then
 export x=eval
fi

# File names without paths
fi=$tmp${input##*/}
fo=$tmp${output##*/}

# Get Experiment number
expno=${input##*/}
expno=${expno%%[mf]*}

# Parse options
while test -n "$3"
do
case $3 in
 16) enc=g726demo a load 16 ; dec=g726demo a adlo 16 ;
 24) enc=g726demo a load 24 ; dec=g726demo a adlo 24 ;
 32) enc=g726demo a load 32 ; dec=g726demo a adlo 32 ;
 40) enc=g726demo a load 40 ; dec=g726demo a adlo 40 ;
 *) echo ERROR! Unrecognized option \"$3\"
 exit ;;
esac
shift
done

# Check if bitrate was specified
if test -z "$enc"
then
 echo ERROR! Bitrate not specified
 exit
fi

# Processing
echo

# ... Encoding
$x $enc   $fi.gi    $fi.bs    \> /dev/null

# ... Decoding
$x $dec $fi.bs     $fo.go \> /dev/null
```

A.7.3.2 Procesamiento en cascada

```
#!/bin/bash
# Tandem processing of codecs
# ${0##*/} InFile OutFile codec [parms] / codec [parms] …]
# Parameters:
# InFile: Input file
# OutFile:    Output file
# codecx: amr fr efr hr g728 g729
# parmsx: valid parameters for codecd
# / is the delimiter between options for codec d and 2
#

# Skip processing if the asked file does not exist
if test ! –f $1 –a –z "$test"
then
 echo Rem File $1 does not exist – skipped ${0##*/} process
 exit
fi

# Default variables
input=$1
output=$2
codec=
parms=
tdm=1

# File names without paths
fi=$tmp${input##*/}
fo=$tmp${output##*/}


# Parse options for codecs

while :
do
 while test "$3" != "/" –a –n "$3"
 do
 case $3 in
 direct|DIRECT) codec=direct ;;
 mnru|MNRU) codec=mnru ;;
 g726|G726) codec=g726 ;;
 codec[a-d]|CODEC[a-d]) codec=$3 ;;
 *) parms="$parms $3" ;;
 esac
 shift
 done

# Skip this "/"
 shift

 # Processing this part of the tandem
 # The experiment number *may* be needed by some of the
 # underlying scripts
 export expno=`echo $input|gawk '{print
tolower(substr($1,match(tolower($1),/exp/)+3,1))}'`

 if test –z "$3"
 then
 # No more tandems – process codec and exit
 $codec   $input    $output   $parms
 break
 else
```

```
 # More parameters in command line - not last tandem
 $codec $input ${fo%.*}.$tdm $parms
 delay ${fo%.*}.$tdm ${fo%.*}.${tdm}d
 input=${fo%.*}.${tdm}d
 let 'tdm = tdm + 1'
 fi
done
```

A.7.4 Programas de bajo nivel

Al igual que estos dos programas están procesados por el programa rector gen-lls, se llama a los programas de nivel medio y el documento impreso se graba en cuatro programas de bajo nivel (dos experimentos por dos plataformas), que en nuestro caso se denominarán gen-e#a-nt.sh y gen-e#a-linux.sh (#=1 & 2).

A.7.4.1 Ejemplo de programas generados

```
# Processing on linux for Experiment 01
# Generated on Fri Feb 9 18:37:46 2001 - CYGWIN_NT-4.0 BAREIL
export PATH=$PATH:../bin
ndate
rm -f ../tmp/*
echo === Files left in ../tmp/: ===
ls ../tmp/

scaldemo -q -dB -gain 10 ../src/exp01/cat-flat.src ../tmp/cat-flat.src.gi
g726demo a load 16 ../tmp/cat-flat.src.gi ../tmp/cat-flat.7h7.bsi
g726demo a adlo 16 ../tmp/cat-flat.7h7.bsi ../tmp/cat-flat.7h7.go
scaldemo -q -dB -gain -10 ../tmp/cat-flat.7h7.go ../prc/exp01/cat-flat.7h7

scaldemo -q -dB -gain 10 ../src/exp01/cat-flat.src ../tmp/cat-flat.src.gi
g726demo a load 23 ../tmp/cat-flat.src.gi ../tmp/cat-flat.7h7.bsi
g726demo a adlo 23 ../tmp/cat-flat.7h7.bsi ../tmp/cat-flat.7h7.go
scaldemo -q -dB -gain -10 ../tmp/cat-flat.7h7.go ../prc/exp01/cat-flat.7h7

…
```

Anexo B

Relación entre métodos subjetivos y objetivos de prueba

B.1 Recomendación UIT-T P.862

B.1.1 Introducción

Los modelos de evaluación objetiva de la calidad vocal analizan muestras de habla y predicen la nota media de opinión (MOS) que se obtendría para esa muestra en un experimento subjetivo bien diseñado y debidamente equilibrado. En comparación con los experimentos subjetivos, las mediciones objetivas son rápidas, económicas y su grado de repetición es alto.

El objetivo principal de este capítulo es presentar el modelo de evaluación de la calidad vocal por percepción, normalizado en la serie de Recomendaciones UIT-T P.862 (UIT-T P.862, UIT-T P.862.1, UIT-T P.862.2 y UIT-T P.862.3). Los conceptos descritos en este capítulo pueden ampliarse fácilmente a la utilización de otros modelos de evaluación objetiva de la calidad vocal de referencia completa. Este capítulo también se propone ofrecer una introducción más general a los distintos tipos de modelo de evaluación objetiva de la calidad y cómo se pueden emplear en distintas configuraciones de pruebas.

Aquí se tratarán las siguientes cuestiones:

- Una introducción a la evaluación objetiva de la calidad vocal
- Una visión general del funcionamiento de la Recomendación UIT-T P.862
- Una guía para utilizar la Recomendación UIT-T P.862
- La importancia de las señales de prueba
- Cómo interpretar de manera objetiva los valores de la nota media de opinión
- Cómo fijar la precisión de un modelo objetivo
- Pautas futuras

B.1.2 Antecedentes

Por lo general, los modelos de evaluación objetiva de la calidad vocal están pensados para procesar muestras de habla de una duración de entre 4 y 30 segundos, en gran medida porque se trata de modelos diseñados para predecir los resultados de experimentos subjetivos. Esto significa que estos modelos son idóneos para procesar conjuntos de datos similares a los generados para experimentos de evaluación subjetiva de la calidad de escucha. La diferencia estriba en que un modelo objetivo solamente necesita procesar una vez cada muestra de habla y puede funcionar a una velocidad muy superior al tiempo real. Por ejemplo, la Recomendación UIT-T P.862 puede procesar un archivo de señales vocales de 8 segundos en aproximadamente 150 ms en un procesador de 3 GHz. Esto significa que el material vocal preparado para un experimento subjetivo ACR típico puede analizarse en menos de un minuto; además, el costo del análisis será sustancialmente inferior al costo que supondría realizar el experimento.

El otro beneficio principal derivado de la utilización de modelos objetivos es la capacidad de repetición. Son muchos los factores que pueden incidir en el valor exacto de la nota media de opinión que una muestra de habla produce en un experimento subjetivo, entre ellos la calidad del resto de muestras del experimento. Esto significa que una muestra de habla determinada podrá no dar exactamente la misma nota media de opinión cuando se evalúa en dos experimentos subjetivos distintos, y que es muy difícil comparar notas medias de

opinión procedentes de experimentos subjetivos distintos, salvo que los experimentos se diseñaran pensando específicamente en esa comparación y se normalicen los datos entre los distintos experimentos mediante técnicas estadísticas avanzadas. Por su parte, en un modelo objetivo la misma muestra de habla siempre dará el mismo resultado. Esto significa que se pueden hacer directamente miles de mediciones en momentos y lugares distintos y compararlas directamente, un detalle que puede ser un requisito fundamental en muchas aplicaciones de evaluación de la calidad.

Asimismo, los modelos de evaluación objetiva de la calidad vocal son idóneos para realizar un gran número de mediciones en redes reales.

B.1.3 Limitaciones

Llegados a este punto, el lector puede estar preguntándose por qué se siguen utilizando los experimentos subjetivos, pero lo cierto es que las pruebas subjetivas siguen desempeñando un papel fundamental en el desarrollo de sistemas de transmisión vocal.

En un mundo ideal, los modelos objetivos reproducirían totalmente los aspectos cognitivos de la evaluación humana de la calidad, así como los aspectos fisiológicos de los sentidos. No obstante, aunque tenemos un conocimiento amplio de los aspectos fisiológicos del sistema de audición humano, y por lo tanto podemos replicarlos, los procesos cognitivos resultan algo más desconocidos. Esto significa que, mientras que es posible diseñar un modelo objetivo para predecir con exactitud el resultado de los experimentos subjetivos existentes en el momento en que se desarrolla un modelo, siempre cabe el riesgo de que el modelo no prediga con exactitud el efecto de nuevas tecnologías de tratamiento de señales vocales que no estaban representadas en los datos de optimización del modelo.

Por lo tanto, debe procederse con cautela cuando se usa un modelo objetivo con nuevas tecnologías de codificación del habla, y debería seguir recurriéndose a los experimentos subjetivos para verificar si un modelo objetivo puede predecir con exactitud las distorsiones introducidas por una nueva tecnología antes de utilizar ambos conjuntamente.

En definitiva, si bien los modelos objetivos son una poderosa herramienta para analizar las tecnologías de codificación del habla ya existentes, los experimentos subjetivos son una solución más adecuada para comparar la calidad de funcionamiento de nuevas tecnologías de codificación y de tecnologías ya existentes, al menos hasta que se haya podido describir el comportamiento del modelo con la nueva tecnología de codificación. Por ejemplo, el UIT-T ha empleado la Recomendación UIT-T P.862 para verificar la calidad de funcionamiento de las aplicaciones fijas y en coma flotante del mismo códec y para entender cómo cambia el comportamiento de un códec de velocidad variable determinado en función de la velocidad binaria, pero las fases de cualificación y selección del códec por parte del UIT-T siguen basándose en datos procedentes de pruebas subjetivas.

B.1.4 Tipos de modelos de evaluación objetiva de la calidad vocal y configuraciones de pruebas

El UIT-T ha definido una serie de términos para señalar la diferencia entre las notas medias de opinión relativas a la calidad de escucha y la calidad de conversación y la diferencia entre las notas medias de opinión procedentes de experimentos subjetivos y las predicciones de la nota media de opinión elaboradas por modelos objetivos (Recomendación UIT-T P.800.1). A continuación se muestra la notación UIT-T. Asimismo, en ocasiones se añaden los sufijos N y W a la notación para indicar si el experimento subjetivo, o el experimento predicho, se llevó a cabo en un contexto de banda estrecha (300–3 400 Hz) o de banda ancha (50-7 000 Hz). Por ejemplo, la calidad de escucha evaluada mediante un experimento subjetivo en un contexto de banda estrecha se indicaría como MOS-LQSN.

Notación de la nota media de opinión del UIT-T

Descripción	Notación
Evaluación subjetiva de la calidad de la escucha	MOS-LQS
Evaluación subjetiva de la calidad de conversación y de la escucha	MOS-CQS
Evaluación objetiva de la calidad de la escucha	MOS-LQO
Evaluación objetiva de la calidad de conversación y de la escucha	MOS-CQO

Tipos de modelos objetivos

En la Enmienda 2 a la Recomendación UIT-T P.10 (2006), el UIT-T ha facilitado una lista de definiciones de términos que se emplean regularmente, aunque a menudo de manera ambigua, relacionados con la clasificación de modelos objetivos según sus funciones y aplicaciones. El texto que se presenta a continuación abunda un poco más en estos conceptos.

Por lo general, los modelos objetivos pertenecen a una de estas tres categorías: de referencia completa, de referencia reducida o sin referencia. Los modelos paramétricos predicen el efecto de errores de red en la calidad de transmisión sin hacer referencia a la forma de onda real de las señales vocales.

a) Modelos de referencia completa

Los modelos de evaluación objetiva de referencia completa miden el impacto de la calidad vocal percibida en uno o más elementos de red comparando dos versiones de una señal de prueba. La primera señal es una copia de la señal original introducida en el sistema sometido a prueba; la segunda es la señal recibida, típicamente degradada. Los modelos de referencia completa se pueden emplear para medir la calidad de un único elemento de red, por ejemplo un códec vocal, o de todo un enlace, como un radioenlace móvil. Por lo general, estos modelos son el tipo de modelo más preciso ya que pueden basar su análisis en una comparación entre las formas original y degradada de la señal de prueba. La Recomendación UIT-T P.862 es un ejemplo de modelo de evaluación objetiva de la calidad vocal de referencia completa que predice la nota media de opinión de la calidad de escucha, o MOS LQO (Recomendación UIT-T P.862).

b) Modelos de referencia reducida

Los modelos de referencia reducida se asemejan a los modelos de referencia completa pero emplean un conjunto reducido de información sobre la señal de prueba original en lugar de una copia completa. Normalmente, los modelos de referencia reducida se usan en aplicaciones en las que la señal de referencia no es una señal de prueba predeterminada sino un tráfico real. Esto significa que la información sobre la señal de referencia debe transmitirse hasta el punto de evaluación a fin de establecer una comparación entre ambas. Aunque el UIT-T no ha trabajado en la normalización de un modelo vocal de referencia reducida, sí que ha normalizado un modelo de vídeo de referencia reducida para complementar sus modelos de vídeo de referencia completa.

c) Modelos sin referencia

Los modelos sin referencia, o de un solo extremo, basan la predicción de la nota media de opinión en una única señal de entrada y se emplean normalmente para comprobar la calidad del tráfico real. Estimar la calidad de una señal sin referencia a la versión no degradada de la señal es un problema complejo, y por lo tanto los modelos sin referencia suelen ser menos precisos que las técnicas equivalentes de referencia completa o de referencia reducida. La Recomendación UIT-T P.563 es un ejemplo de modelo de evaluación objetiva de la calidad vocal sin referencia que predice la nota media de opinión de la calidad de escucha, o MOS-LQO (UIT-T P.563); la Recomendación UIT-T P.562 es un ejemplo de modelo sin referencia que predice la nota media de opinión de la calidad de conversación, o MOS-CQO (UIT-T P.562).

Modelos paramétricos

Los modelos paramétricos están ideados para predecir el impacto de las degradaciones de la transmisión en la calidad vocal percibida por el usuario final de un sistema. Los modelos paramétricos son modelos sin referencia, pero no analizan la forma de onda de la señal vocal real sino parámetros relacionados con el sistema de transmisión subyacente. Por ejemplo, en el caso de un sistema de voz por IP (VoIP), los parámetros principales procederán de la pérdida de paquetes y de las características de fluctuación del enlace. Los modelos paramétricos de evaluación de la calidad vocal basan sus predicciones en la hipótesis de que la carga útil contiene una señal vocal típica y bien condicionada.

Encontramos un ejemplo de modelo paramétrico en la Recomendación UIT-T P.564, que se ocupa de un tipo de modelos de predicción de MOS-LQO que solamente analizan el encabezamiento IP/UDP/RTP de los paquetes de VoIP. Este enfoque presenta varias ventajas, entre otras baja complejidad de funcionamiento y resistencia a la encriptación de la carga útil. Nótese que la Recomendación UIT-T P.564 no especifica un modelo paramétrico único, sino los criterios mínimos de funcionamiento que deben satisfacer los modelos para ajustarse a la norma.

Configuración de las pruebas

Las técnicas objetivas de prueba pueden dividirse en activas o pasivas.

a) Pruebas activas

En una configuración de prueba activa, se introduce una señal de prueba en el sistema sometido a prueba y, a continuación, se captura y se evalúa en un punto posterior. Las pruebas activas pueden emplearse durante el desarrollo, la puesta en servicio y la comprobación técnica rutinaria de un servicio de comunicaciones. Además, las pruebas activas se consideran "intrusivas" si la interrupción en el servicio o el usuario se deberá, o podrá deberse, a la prueba.

Generalmente, las pruebas activas se llevan a cabo utilizando un modelo de referencia completa, porque se conoce de antemano la señal de prueba y se puede almacenar una copia en el punto de evaluación. No obstante, en principio no existen motivos para no utilizar un modelo de referencia reducida o sin referencia para medir la calidad de la señal en el punto de evaluación.

Cuando se emplean pruebas activas para medir la calidad de funcionamiento de una red real, normalmente se introduce y se captura la señal de prueba en los extremos de la red. Por ejemplo, en las pruebas en movimiento en redes móviles, se usan aparatos portátiles para introducir y capturar la señal. No obstante, conviene señalar que los puntos de prueba no tienen que ser equipo del cliente, y que se puede utilizar la misma señal de prueba para múltiples mediciones intrusivas.

Las configuraciones de prueba activa están especialmente indicadas para pruebas en laboratorio, por su facilidad para aislar componentes de red e introducir y capturar señales de prueba.

b) Prueba pasiva

En las pruebas activas, es necesario utilizar la capacidad de red, y puede ser difícil introducir una señal de prueba en un gran número de puntos de acceso a la red distintos. En las pruebas pasivas no se necesita una señal de prueba. En su lugar, se comprueba el tráfico real del cliente para determinar su calidad en distintos puntos de una red. Las mediciones del tráfico del cliente permiten que no se pierda la capacidad de red y que el proveedor del servicio pueda medir la calidad vocal de señales que atraviesan un gran número de rutas distintas, lo que permite realizar un número mayor de mediciones a un coste menor. Las pruebas pasivas se califican como "no intrusivas" porque no crean interrupciones en el servicio.

Normalmente, en las pruebas pasivas se emplea un modelo sin referencia, aunque también puede utilizarse un modelo de referencia reducida, como el que se ha descrito anteriormente. Sea como fuere, es fundamental

que el modelo pueda funcionar con una gama de señales muy amplia, ya que la señal de entrada no se limita a una señal de prueba cuidadosamente controlada, tal y como sucede con las pruebas activas. En efecto, uno de los retos que plantea el diseño de un modelo adecuado para realizar pruebas pasivas es que debe ser capaz de determinar de una manera fiable qué partes de la señal habría que analizar y cuáles habría que ignorar. Por ejemplo, en todas las conversaciones telefónicas suele haber toses, risas y tonos de red, elementos que sin embargo deben excluirse del proceso de medición de la calidad vocal.

B.1.5 Recomendación UIT-T P.862: Evaluación de la calidad vocal por percepción

La Recomendación UIT-T P.862 normalizó, en 2001, la evaluación de la calidad vocal por percepción, que ha sido aplicada con posterioridad por distintos suministradores de equipos de prueba y medición.

Algoritmos para la evaluación objetiva de la calidad vocal por percepción como el de la Recomendación UIT-T P.862 emplean modelos psicoacústicos para transformar las señales de entrada en una forma que represente cómo percibe un humano las señales. Su introductor fue, en 1985, Karjalainen, con el algoritmo de distancia del espectro auditivo (ASD) (Karjalainen). Otros autores, entre ellos Wang, Hollier y Beerends ((Wang 1992), (Hollier 1994), (Beerends 1994)), siguieron desarrollando modelos de evaluación de la calidad vocal por percepción.

El primer modelo de evaluación de la calidad vocal de referencia completa normalizado por el UIT-T fue la medida perceptual de la calidad vocal, desarrollado por Beerends et al en KPN Research, en los Países Bajos, y normalizado en 1996 en la Recomendación UIT-T P.861.

El alcance de la Recomendación UIT-T P.861 se limitó a la evaluación de artefactos de codificación y no abarcaba los efectos de los errores de transmisión. En consecuencia, el UIT-T invitó a que se presentaran contribuciones para un nuevo modelo de evaluación de la calidad vocal de referencia completa con un alcance más amplio. Distintas organizaciones propusieron varios modelos, y el modelo normalizado en la Recomendación UIT-T P.862, de 2001, se basó en una combinación de dos de estos: el modelo desarrollado por el equipo de KPN y el modelo desarrollado por Hollier et al en BT, en el Reino Unido ((Rix 2000), (Rix 2002), (Beerends 2002)).

En los experimentos de comparación que llevó a cabo el UIT-T, la Recomendación UIT-T P.862 alcanzó un coeficiente de correlación media del 0,93 en comparación con los valores de la nota media de opinión por condición obtenidos en 22 experimentos subjetivos, que representan más de 5 000 archivos de prueba con un amplio abanico de condiciones de red. Más adelante se aborda la cuantificación de la calidad de funcionamiento del modelo objetivo.

Recomendación UIT-T P.862.1: Correspondencia MOS-LQO

El resultado del algoritmo original de la Recomendación UIT-T P.862 no refleja la escala de la nota media de opinión 1 a 5, pero el valor resultante se encuentra entre -0,5 y 4,5. La Recomendación UIT-T P.862.1 especifica una función que traslada el resultado bruto de UIT-T P.862 a la escala de MOS-LQO. De este modo, los resultados de MOS-LQO para códecs de referencia como UIT-T G.711 y UIT-T G.729 son indicativos de los resultados que cabría esperar en un experimento subjetivo bien diseñado y equilibrado. Más adelante se aborda con más detalle la cuestión de las correspondencias de la nota media de opinión.

Recomendación UIT-T P.862.2

La Recomendación UIT-T P.862.2 amplía el funcionamiento del algoritmo original de la Recomendación UIT-T P.862 a señales con una anchura de banda de la señal sonora de hasta 8 kHz (es decir, una velocidad de muestreo de 16 kHz). El algoritmo de la Recomendación UIT-T P.862.2 posee una correspondencia integral de MOS LQO y, de este modo, genera directamente resultados en la escala de la nota media de opinión 1 a 5.

Visión general del algoritmo de la Recomendación UIT-T P.862

En la figura B-1 se describen los distintos pasos del procesamiento realizado por la Recomendación UIT-T P.862.

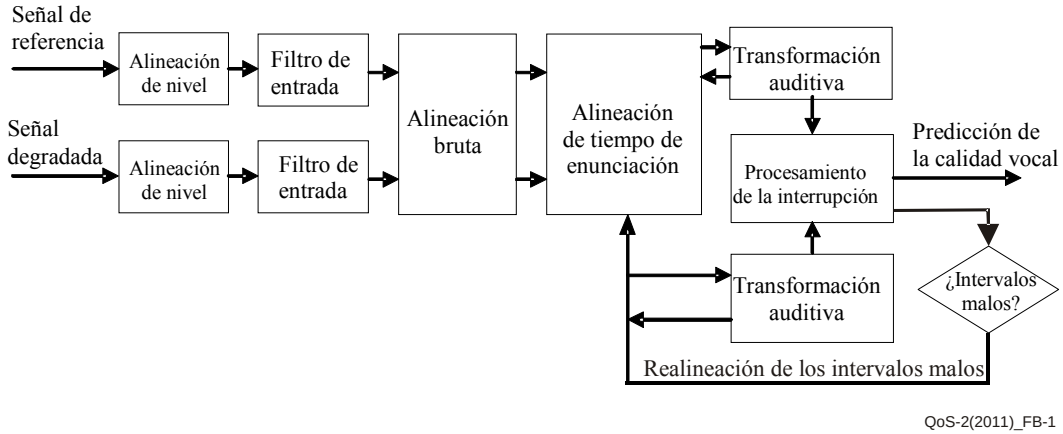


Figura B-1 – Etapas de procesamiento de la Recomendación UIT-T P.862

Alineación de nivel

La Recomendación UIT-T P.862 asume que las señales de entrada llegan al sujeto a un nivel de escucha subjetivo de 79dB SPL en el punto de referencia oído. En consecuencia, se aplica una ganancia tanto a la señal de referencia como a la señal degradada para que alcancen este nivel en el modelo auditivo. Esta alineación de nivel debe aplicarse porque no es necesario que la señal de referencia esté a un nivel determinado y porque la ganancia del sistema sometido a prueba se desconoce antes de la prueba.

Filtrado de entrada

La Recomendación UIT-T P.862 emplea un filtro de entrada para simular el trayecto de recepción del microteléfono que tiene en cuenta el efecto de los componentes eléctricos y acústicos de un microteléfono típico. En el funcionamiento en banda estrecha, el filtro de entrada que se emplea se asemeja al receptor SIR característico especificado en la Recomendación UIT-T P.48; en el funcionamiento en banda ancha, se utiliza un filtro simple de paso alto de 100 Hz. Dado que la respuesta de recepción efectiva de los microteléfonos varía considerablemente, la Recomendación UIT-T P.862 es relativamente sensible a las características del filtro de entrada.

Alineación de tiempo

La señal en el archivo degradado a menudo presenta un desplazamiento temporal con respecto a la señal del archivo de referencia. Estos desplazamientos pueden deberse al retardo introducido por el sistema sometido a prueba o a errores en la sincronización de la captura del archivo degradado con respecto a la inserción del archivo de referencia. En consecuencia, es necesario alinear la señal de referencia y la señal degradada antes de poder compararlas.

Asimismo, algunos sistemas de transmisión pueden introducir un retardo variable, por ejemplo cuando se reinicia una memoria intermedia de fluctuación en un dispositivo de VoIP. La Recomendación UIT-T P.862 aborda este problema utilizando un detector de actividad vocal para dividir la señal de referencia y la señal degradada en fragmentos de habla o "enunciaciones". La Recomendación UIT-T P.862 compensa los cambios en el retardo alineando independientemente cada fragmento de habla en la señal degradada con su enunciación correspondiente en la señal de referencia.

Inicialmente, se considera enunciación todo fragmento de habla sin pausas superior a 200 ms; por lo general, cada enunciación representa una frase de habla. No obstante, si bien la alineación basada en esta clasificación inicial resolvería los cambios en el retardo entre enunciaciones, sería menos efectiva para los cambios en el retardo que se producen durante una enunciación. Por lo tanto, el algoritmo de alineación de tiempo analiza el retardo en cada mitad de cada enunciación, y divide en dos la enunciación si estos retardos difieren en más de 4 ms. Este proceso se repite una y otra vez hasta que cada enunciación tiene un retardo casi constante, o hasta que una nueva división dé como resultado una enunciación de menos de 300 ms.

Esta descomposición recurrente de la señal en enunciaciones con un retardo casi constante supone que el tiempo que se emplea para procesar un par de archivos con un retardo relativo variable puede ser significativamente mayor que el tiempo que se emplea para un par de archivos con un retardo relativo fijo.

Tras la transformación auditiva se lleva a cabo una última fase de la alineación de tiempo, en la que se realinean aquellas secciones de señales vocales con una perturbación muy grande. Este paso mejora la exactitud del modelo con un número reducido de archivos en los que el proceso inicial de alineación de tiempo no identificó correctamente los cambios en el retardo.

Transformación auditiva

Para comparar la señal de referencia y la señal degradada basándose en su posible recepción en el oyente, se somete cada una de estas señales a una transformación auditiva que reproduce algunas propiedades importantes del sistema auditivo humano, entre ellas la frecuencia de percepción (Bark) y la sonoridad (Sone).

La transformación auditiva que figura en la Recomendación UIT-T P.862 se basa en una transformada rápida de Fourier (FFT) a corto plazo de la señal de entrada. La transformada agrupa distintos conjuntos de grupos de frecuencias adyacentes en bandas críticas aproximadas separadas entre sí 0,5 Bark (Zwicker, 1961). La estimación de la sonoridad se basa en el cálculo de la "sonoridad específica" de Zwicker, que transforma la excitación sonora de una banda crítica determinada en sonoridad en sone (Zwicker, 1990). Esto da una representación tiempo-frecuencia de la sonoridad percibida de la señal que se conoce como sensación de superficie.

Ecualización

Una parte de la transformación auditiva mitiga algunos factores de escasa importancia subjetiva. En primer lugar, se estima la función de transferencia del sistema sometido a prueba y se utiliza el resultado para ecualizar la señal de referencia con respecto a la señal degradada en el ámbito de la transformación auditiva. Esto tiene en cuenta cualquier elemento de red que filtre la señal. En segundo lugar, se estima la ganancia de amplitud por tramas del sistema y se utiliza el resultado para ecualizar la transformación auditiva del archivo degradado con respecto a la referencia. En ambos casos, se limita la ecualización y no se cancelan los valores elevados de filtración o de variación de la ganancia, sino que se tratan como diferencias audibles entre ambas señales.

Procesamiento de la perturbación

La diferencia entre las superficies de sensación del archivo de referencia y del archivo degradado se conoce como superficie de error y refleja toda diferencia audible introducida por el sistema sometido a prueba. La superficie de error se analiza mediante un proceso que tiene en cuenta un efecto en virtud del cual las pequeñas distorsiones en una señal no son audibles en presencia de señales potentes. En psicoacústica, este efecto se conoce como "enmascaramiento".

Dos parámetros de perturbación se calculan como medias no lineales sobre áreas específicas de la superficie de error. Estos parámetros de perturbación son:

- la perturbación absoluta (simétrica) – una medición del error audible absoluto;
- la perturbación aditiva (asimétrica) – una medición de los errores audibles significativamente más altos que la referencia.

Los dos parámetros representan la cantidad total de cada tipo de error audible.

Por ultimo, los parámetros de error se convierten en una nota de calidad, que es una combinación lineal del valor medio de la perturbación simétrica y del valor medio de la perturbación asimétrica.

B.1.6 Uso de la Recomendación UIT-T P.862

En 2005, el UIT-T publicó una Guía de Aplicación para la Recomendación UIT-T P.862. La Recomendación UIT-T P.862.3 resultante ofrece directrices sobre las mejores prácticas cuando se utiliza el algoritmo de la Recomendación UIT-T P.862 y el algoritmo de la Recomendación UIT-T P.862.2.

En esta sección, se abordan algunos de los factores más pertinentes para las pruebas en laboratorio y se ofrecen algunas pautas adicionales con respecto a la utilización de la Recomendación UIT-T P.862 que no figuraban en la Recomendación UIT-T P.862.3. Se recomienda encarecidamente a los lectores que deseen aprovechar de manera óptima el modelo que lean también la Recomendación UIT-T P.862.3.

La Recomendación UIT-T P.862 se elaboró para predecir los experimentos de evaluación de la calidad de escucha a partir de la escala de evaluación por categorías absolutas (ACR). Sin embargo, el método de evaluación de la calidad vocal de referencia completa se asemeja mucho más a los experimentos en los que se ha empleado la escala de evaluación por categorías de degradación (DCR), en la que se pide a los sujetos que comparen señales degradadas con una versión no degradada de la misma señal.

Esta situación puede provocar de vez en cuando discrepancias. Por ejemplo, puede aparecer una discrepancia significativa debida a que el material vocal de distintos hablantes puede dar lugar a valores estadísticamente distintos de la nota media de opinión en un experimento subjetivo ACR, aun cuando no se haya distorsionado o procesado la señal vocal. Esto obedece simplemente a que los sujetos parecen preferir unos hablantes a otros. La Recomendación UIT-T P.862, por su parte, siempre dará un resultado de MOS-LQO de 4,56 si ambas señales de entrada son idénticas, con independencia del material vocal.

La capacidad de la Recomendación UIT-T P.862 para facilitar una predicción de la calidad vocal rápida y reproducible sobre un conjunto muy grande de material vocal puede resultar sumamente útil al desarrollar códecs.

Cuando el UIT-T normalizó la Recomendación UIT-T P.862, se validó su calidad de funcionamiento a partir de 22 experimentos ACR. El alcance de estos experimentos incluía:

- una serie de códecs de velocidad binaria media a alta;
- los efectos de los errores binarios, la pérdida de tramas y los algoritmos de ocultación de errores;
- códecs en tándem;
- los efectos de los niveles de entrada de distintos códecs;
- codificación a una velocidad binaria variable.

Los códecs específicos incluyen el UIT-T G.711, UIT-T G.726, UIT-T G.727, UIT-T G.728, UIT-T G.729, UIT-T G.723.1, GSM FR, HR, EFR y AMR, VSELP y p.i.r.e-A. No obstante, conviene decir que se observa un rendimiento sistemático por debajo de lo normal en términos de predicción en la familia de codificadores p.i.r.e, entre ellos p.i.r.e.-A, p.i.r.e.-B y p.i.r.e.-WB, por parte del UIT-T P.862.1 y del UIT-T P.862 en

comparación con las pruebas subjetivas, y más si cabe con respecto a los codificadores del tipo AMR. Este rendimiento sistemático por debajo de lo normal en términos de predicción puede observarse en una amplia gama de velocidades binarias y patrones de error.

En la Recomendación UIT-T P.862 se describe todo el alcance, pero debería señalarse que la Recomendación UIT-T P.862 no es idónea para probar las tecnologías de codificación que se suelen emplear en códecs vocales con una velocidad binaria muy baja, como IMBE, AMBE, PWI y STC, ya que no se incluyeron en el alcance original de la Recomendación UIT-T P.862 y, en consecuencia, el algoritmo no se diseñó para predecir su efecto sobre la calidad vocal.

También conviene señalar que, aunque la Recomendación UIT-T P.862 incluye en su alcance la transposición temporal, pruebas más recientes sugieren que la Recomendación UIT-T P.862 no gestiona de una manera fiable los efectos de los errores de transposición temporal y de velocidad de muestreo entre la señal de referencia y la señal degradada.

Los desarrolladores de códecs no deberían utilizar la Recomendación UIT-T P.862 aisladamente. Se sigue alentando a los desarrolladores a que utilicen en su ciclo de desarrollo experimentos subjetivos, y es especialmente importante recurrir a experimentos subjetivos para comprobar que la Recomendación UIT-T P.862 funciona de manera fiable con nuevas tecnologías de codificación.

El recurso a la Recomendación UIT-T P.862 para llevar a cabo pruebas intrusivas en sistemas reales plantea una serie de desafíos adicionales relacionados con la función de transferencia y los silencios al principio y al final en la señal degradada que aquí no se abordan.

B.1.7 Elección del material de referencia

La base de datos utilizada para validar la Recomendación UIT-T P.862 incluía experimentos subjetivos en inglés norteamericano y británico, francés, alemán, italiano, sueco, neerlandés y japonés. El abanico concreto de idiomas utilizados para las pruebas de laboratorio de la Recomendación UIT-T P.862 dependerá de la aplicación.

El anexo B a la Recomendación UIT-T P.501 incluye distintas muestras de habla en diversas lenguas que pueden emplearse con la Recomendación UIT-T P.862. No obstante, debe procederse con cautela porque no todas las muestras reúnen los requisitos mínimos en lo relativo al umbral de ruido, la actividad vocal y los silencios al principio y al fin.

Por lo general, se recomienda obtener para las pruebas material que no haya sido filtrado y, a continuación, utilizar la biblioteca de herramientas de soporte lógico (STL) del UIT-T definida en la Recomendación UIT-T G.191 para realizar un filtrado previo del material vocal de referencia antes de introducirlo en el códec/sistema sometido a prueba.

Los autores no recomiendan utilizar señales vocales artificiales, definidas en la Recomendación UIT-T P.50, con el algoritmo de la Recomendación UIT-T P.862.

Toda vez que el material para las pruebas debería estar mayoritariamente libre de ruido de fondo, la Recomendación UIT-T P.862 puede ser hipersensible a unos niveles bajos de ruido en la señal degradada si en las pausas de las señales de referencia el silencio es absoluto. Por lo tanto, se recomienda leer y cumplir la sección sobre Fondo de Ruido de la Recomendación UIT-T P.862.3. Los usuarios de la Recomendación UIT-T P.862 también deberían prestar atención a las recomendaciones sobre actividad vocal y pausas al principio y al final del material de referencia.

Una cuestión que hay que tener en cuenta al utilizar la Recomendación UIT-T P.862 es que esta Recomendación pasará por alto las distorsiones introducidas en la señal de referencia que estén antes de la

señal vocal correspondiente a la primera enunciación en la señal de referencia o después de la señal vocal correspondiente a la última enunciación en la señal de referencia.

Existen buenos motivos para utilizar muestras de habla de 8 a 12 segundos. En primer lugar, la mayoría de experimentos subjetivos ACR como el que se predice en la Recomendación UIT-T P.862 utilizan muestras de habla de 8 segundos que constan de dos frases cortas. La validación de la Recomendación UIT-T P.862 se llevó a cabo principalmente utilizando estos "pares de frases", de manera que es lógico emplear material similar cuando se use el algoritmo de la Recomendación UIT-T P.862 para llevar a cabo mediciones. En efecto, a menudo el material vocal de los experimentos subjetivos está pensado para que sea fonéticamente equilibrado, y por lo tanto es idóneo para la Recomendación UIT-T P.862.

El segundo motivo para evitar muestras de habla más largas son las dudas acerca de la calidad de funcionamiento de la Recomendación UIT-T P.862 cuando entre el inicio del archivo de referencia y el final de la última enunciación transcurren más de 16 segundos. En esta situación, se desacentúan las degradaciones al principio del archivo con respecto a las del final del archivo, a fin de reproducir lo que se conoce como efecto "reciente", en virtud del cual la opinión de un oyente puede decantarse por causa de la calidad de la señal al final de una señal de prueba larga; no obstante, en muchas situaciones, este efecto puede no ser deseable, sobre todo si, para la prueba, se ha elegido un material con la intención de que sea fonéticamente equilibrado.

El apéndice I de la Recomendación UIT-T P.862.3 ofrece una serie de directrices sobre la utilización de STL para procesar material vocal de referencia a fin de simular distintas hipótesis operativas, entre ellas los filtros de entrada y la introducción de fases de codificación adicionales, como la Recomendación UIT-T G.711.

B.1.8 Interpretación de los resultados objetivos de la nota media de opinión

Al interpretar los valores de la nota media de opinión que han predicho los modelos de evaluación objetiva de la calidad, se asume que el "contexto" es un experimento bien diseñado y bien equilibrado, es decir, que incluye una amplia gama de distorsiones que están presentes de manera equilibrada.

Véanse, por ejemplo, los resultados de procesar la Recomendación UIT-T P.862 con la correspondencia de MOS-LQO de la Recomendación UIT-T P.862.1 que figuran en el apéndice I a la Recomendación ITU-T P.862.3.

¿Qué grado de exactitud poseen los modelos objetivos? La calidad de funcionamiento de un modelo objetivo suele determinarse procesando con dicho modelo material de habla empleado en distintos experimentos subjetivos y comparando posteriormente las predicciones con los valores de la nota media de opinión obtenidos de los sujetos. Normalmente, esta comparación se calcula como el coeficiente de correlación de Pearson o como el error cuadrático medio (ECM).

Pese a que estas mediciones pueden parecer útiles, plantean el problema de cómo compensar el hecho de que un mismo archivo de señales vocales pueda, a causa del "efecto contexto" anteriormente mencionado, tener valores de MOS-LQO muy distintos en dos experimentos distintos. La solución más habitual consiste en aplicar una función de correspondencia a los datos de la prueba objetiva o subjetiva antes de calcular el coeficiente de correlación o el valor del ECM.

B.1.9 Orientaciones futuras

En el momento de redactar este manual, el UIT-T ha elaborado un nuevo modelo de evaluación objetiva de la calidad vocal de referencia completa, publicado como Recomendación UIT-T P.863. El modelo funcionará para dos velocidades de muestra: 8 kHz y 48 kHz. El modo correspondiente a una velocidad de 8 kHz se empleará para evaluar las señales vocales de banda estrecha (anchura de banda 300-3 400 Hz),

mientras que el modo correspondiente a una velocidad de 48 kHz se utilizará para evaluar las señales vocales de banda ancha (anchura de banda de audio 50-7 000 Hz) y superancha (50-14 000 Hz). Además de ampliar el funcionamiento a anchuras de banda superiores, el modelo se validará con tipos de distorsión que no predominaban cuando se normalizó la Recomendación UIT-T P.862, por ejemplo la transposición temporal, la introducción de errores en la velocidad de muestreo, los efectos de la supresión del ruido y los efectos del nivel de escucha (que, en la Recomendación UIT-T P.862 se consideraba que tenía un valor fijo de 79 dB SPL).

El problema de crear modelos para la evaluación de la calidad vocal sin referencia sigue planteando un desafío considerable, y la Recomendación UIT-T P.563 es menos precisa que la Recomendación UIT-T P.862. Aunque no parece posible que los modelos sin referencia lleguen a alcanzar la misma calidad de funcionamiento que sus homólogos de referencia completa, se cree que todavía hay margen de mejora más allá de lo que se ofrece en la Recomendación UIT-T P.563. Existen modelos exclusivos que ofrecen el mismo nivel de precisión para cálculos de una complejidad reducida.

Pese a que los modelos paramétricos no ofrecen los beneficios de un análisis de la forma de onda real, sí que pueden ser muy ligeros, hecho que permite analizar muchos miles de trenes reales de VoIP coincidentes. Los modelos que se ajustan a la Recomendación UIT-T P.564 también pueden predecir la calidad de trenes IP encriptados a través del protocolo de transporte en tiempo real seguro (SRTP), porque solamente pueden basarse en la información de la cabecera del protocolo en tiempo real y el tiempo de captura de los paquetes, y no en la carga útil del protocolo en tiempo real. Estos modelos paramétricos ya están integrados en los puntos extremos y en el equipo de comprobación técnica situado en los puntos de demarcación primaria en las redes, por ejemplo en pasarelas, controladores de frontera de sesión y encaminadores de borde que interconectan las redes de área local y las redes de área extensa.

B.1.10 Conclusiones sobre los métodos subjetivos y objetivos de prueba

Hasta que seamos capaces de replicar totalmente los aspectos cognitivos de la evaluación humana de la calidad además de los aspectos fisiológicos de los sentidos, que en la actualidad entendemos mucho mejor, las pruebas subjetivas seguirán siendo, con toda probabilidad, el "patrón-oro" en lo que respecta a la evaluación de la calidad vocal. No obstante, si nos fijamos en su alcance y en sus limitaciones, los modelos objetivos pueden revelarse como una herramienta muy valiosa tanto en el laboratorio como sobre el terreno, por su carácter asequible y reiterativo. En particular, la posibilidad de que los modelos objetivos procesen un gran número de muestras vocales en un corto espacio de tiempo puede ser de suma utilidad para desarrolladores de nuevos códecs.

B.2 Conteo de los errores de clasificación evaluados por percepción

B.2.1 Introducción

En la Recomendación UIT-T G.720.1, el "detector de actividad sonora genérica" (GSAD) es un módulo de procesamiento de extremo delantero independiente que se puede aplicar antes de las aplicaciones de procesamiento de la señal que funcionan en la entrada de banda ancha o de banda estrecha, como los códecs vocales o sonoros. Su función principal es indicar la actividad de la trama de entrada. En el caso de una trama activa, indica además si la trama de entrada es una señal vocal o música; en el caso de una trama inactiva, indica si la trama es una trama de silencio o una trama de ruido audible. El GSAD también puede tener únicamente la función primaria de indicar la actividad de la trama de entrada. Uno de los principales parámetros de calidad de funcionamiento del GASD es la velocidad a la que se clasifica erróneamente una trama vocal activa como inactiva. No obstante, comoquiera que esta velocidad de clasificación errónea no

tiene en cuenta el comportamiento por percepción de la detección de actividad en el GSAD, se ha utilizado un nuevo sistema de medición objetivo, conocido como conteo de los errores de clasificación evaluados por percepción (PWMC). La finalidad que se persigue al utilizar este sistema en lugar de la velocidad de clasificación errónea es velar por que exista una correlación sólida con la calidad vocal subjetiva durante periodos en los que se producen errores de clasificación.

B.2.2 Descripción general del enfoque PWMC

El PWMC se basa en la evaluación de la degradación por percepción resultante de una trama vocal activa erróneamente clasificada como inactiva por el GSAD. Dado que el GSAD no comporta una codificación adicional, se asume que la degradación puede medirse sustituyendo la señal de la trama activa erróneamente clasificada por el ruido de fondo subyacente original, y medir a continuación la diferencia por percepción entre la señal modificada y la señal sin modificar.

El PWMC está pensado para ser utilizado como una medida relativa, es decir, el valor de PWMC del posible algoritmo GSAD se comparará con el valor de PWMC de un algoritmo de referencia. Además, el PWMC está diseñado para ser utilizado con señales vocales, es decir, que no se llevaron a cabo pruebas de validación para utilizar el PWMC con otras señales, como la música.

A continuación se ofrece una descripción más detallada del enfoque de PWMC:

1) Se obtienen las marcas de actividad real del material vocal de prueba utilizando un umbral de energía sobre la señal vocal limpia, antes de añadir el ruido.

2) Se prepara el material vocal de prueba añadiendo ruidos de fondo a la señal vocal limpia en distintas SNR.

3) Se procesa el material vocal de prueba con el posible algoritmo GSAD y con un algoritmo VAD de referencia para generar las marcas de actividad de salida de cada trama, tanto para el posible algoritmo GSAD como para el VAD de referencia.

4) Las marcas de actividad de salida generadas por el posible algoritmo y por el algoritmo de referencia en el paso 3 se comparan con las marcas de actividad real. Cada trama vocal activa erróneamente clasificada se sustituye por el ruido de fondo subyacente original resultante en dos versiones modificadas (degradadas) del material vocal de prueba, una para el posible algoritmo GSAD y la otra para el VAD de referencia, utilizando las ventanas de aumento y disminución gradual de la señal, tal y como se describe en el apéndice 1.

5) Las dos versiones degradadas del material vocal de prueba se comparan con el material vocal de prueba no degradado utilizando una medición objetiva por percepción basada en la Recomendación UIT-T P.862. A continuación, se evalúan los valores de PWMC resultantes para especificar la calidad de funcionamiento relativa del posible algoritmo GSAD con respecto al VAD de referencia.

B.2.3 Descripción de las pruebas subjetivas utilizadas para validar el enfoque PWMC

El objetivo de estas pruebas era proporcionar una base de datos subjetiva con un gran número de muestras de habla y sus versiones degradadas, correspondientes a una amplia gama de distribuciones con errores de clasificación y sus valores correspondientes de nota media de opinión. Esta base de datos puede utilizarse posteriormente para validar la medición objetiva de PWMC. Si los valores objetivos de PWMC son monótonos con respecto al valor real de la nota media de opinión, se valida el PWMC. Para garantizar que la base de datos contiene una gama amplia de distribuciones y de velocidades con errores de clasificación realistas, se utilizaron en la prueba de escucha 16 distribuciones distintas con errores de clasificación (16 patrones con errores de clasificación), a partir de los patrones reales erróneamente clasificados elaborados

por el VAD del anexo B a la Recomendación UIT-T G.729. Los 16 patrones con errores de clasificación se obtuvieron de la manera siguiente:

1) Se ajustó el nivel de todas las muestras de señales vocales limpias de 8 segundos (vectores de prueba) de la base de datos al nivel nominal, y se concatenaron todas las muestras con ajuste de nivel en un archivo de señales vocales largo.

2) Se utilizó el archivo de señales vocales limpias largo para generar 16 versiones largas con ruido; cada una de ellas corresponde a una de las 16 combinaciones posibles de tipo de ruido y nivel SNR. Los tipos de ruido incluían el tráfico, la calle, murmullos y la oficina, y los niveles SNR eran de 6dB, 10dB, 15dB y 20dB.

3) Se obtuvieron los 16 archivos VAD del anexo B a la Recomendación UIT-T G.729 para los 16 archivos de señales vocales largos con ruido bajo la forma de 16 patrones con errores de clasificación.

4) Se descompusieron los 16 archivos de marca largos en los correspondientes archivos de marca VAD de cada sector de prueba de 8 segundos, con arreglo al anexo B a la Recomendación UIT-T G.729. Así, cada vector de prueba tiene 16 patrones con errores de clasificación distintos.

Estos 16 patrones con errores de clasificación se analizaron y abarcan una velocidad global de error de clasificación que va de 0,2% a 50%, una gama amplia de errores de clasificación en el inicio de la señal vocal (2,3% hasta 88%), la continuación de la señal vocal (0% hasta 60%), el final de la señal vocal (0% hasta 72%) y una gama amplia de errores de clasificación repentinos relativos al inicio, la continuación y el fin de la señal vocal, respectivamente. Las muestras de habla con el mismo patrón con errores de clasificación se probaron en cuatro entornos de ruido distintos, ruido de tráfico (15dB), ruido de tráfico (25dB), cafetería (15dB) y cafetería (25dB), a fin de tener en cuenta el efecto de distintas relaciones SNR y de distintos tipos de ruido de fondo en la percepción auditiva humana. De esta comprobación salieron las 64 condiciones de prueba que figuran en el plan de prueba. Véase el cuadro B-1 a continuación:

Cuadro B-1 – Lista de Condiciones

Número	Condiciones de prueba	Ruido	S/N
1	Patrón con errores de clasificación 1	Tráfico	25 dB
2	Patrón con errores de clasificación 2	Tráfico	25 dB
3	Patrón con errores de clasificación 3	Tráfico	25 dB
4	Patrón con errores de clasificación 4	Tráfico	25 dB
5	Patrón con errores de clasificación 5	Tráfico	25 dB
6	Patrón con errores de clasificación 6	Tráfico	25 dB
7	Patrón con errores de clasificación 7	Tráfico	25 dB
8	Patrón con errores de clasificación 8	Tráfico	25 dB
9	Patrón con errores de clasificación 9	Tráfico	25 dB
10	Patrón con errores de clasificación 10	Tráfico	25 dB
11	Patrón con errores de clasificación 11	Tráfico	25 dB
12	Patrón con errores de clasificación 12	Tráfico	25 dB
13	Patrón con errores de clasificación 13	Tráfico	25 dB
14	Patrón con errores de clasificación 14	Tráfico	25 dB
15	Patrón con errores de clasificación 15	Tráfico	25 dB
16	Patrón con errores de clasificación 16	Tráfico	25 dB

Número	Condiciones de prueba	Ruido	S/N
17	Patrón con errores de clasificación 1	Tráfico	15 dB
18	Patrón con errores de clasificación 2	Tráfico	15 dB
19	Patrón con errores de clasificación 3	Tráfico	15 dB
20	Patrón con errores de clasificación 4	Tráfico	15 dB
21	Patrón con errores de clasificación 5	Tráfico	15 dB
22	Patrón con errores de clasificación 6	Tráfico	15 dB
23	Patrón con errores de clasificación 7	Tráfico	15 dB
24	Patrón con errores de clasificación 8	Tráfico	15 dB
25	Patrón con errores de clasificación 9	Tráfico	15 dB
26	Patrón con errores de clasificación 10	Tráfico	15 dB
27	Patrón con errores de clasificación 11	Tráfico	15 dB
28	Patrón con errores de clasificación 12	Tráfico	15 dB
29	Patrón con errores de clasificación 13	Tráfico	15 dB
30	Patrón con errores de clasificación 14	Tráfico	15 dB
31	Patrón con errores de clasificación 15	Tráfico	15 dB
32	Patrón con errores de clasificación 16	Tráfico	15 dB
33	Patrón con errores de clasificación 1	Cafetería	25 dB
34	Patrón con errores de clasificación 2	Cafetería	25 dB
35	Patrón con errores de clasificación 3	Cafetería	25 dB
36	Patrón con errores de clasificación 4	Cafetería	25 dB
37	Patrón con errores de clasificación 5	Cafetería	25 dB
38	Patrón con errores de clasificación 6	Cafetería	25 dB
39	Patrón con errores de clasificación 7	Cafetería	25 dB
40	Patrón con errores de clasificación 8	Cafetería	25 dB
41	Patrón con errores de clasificación 9	Cafetería	25 dB
42	Patrón con errores de clasificación 10	Cafetería	25 dB
43	Patrón con errores de clasificación 11	Cafetería	25 dB
44	Patrón con errores de clasificación 12	Cafetería	25 dB
45	Patrón con errores de clasificación 13	Cafetería	25 dB
46	Patrón con errores de clasificación 14	Cafetería	25 dB
47	Patrón con errores de clasificación 15	Cafetería	25 dB
48	Patrón con errores de clasificación 16	Cafetería	25 dB
49	Patrón con errores de clasificación 1	Cafetería	15 dB
50	Patrón con errores de clasificación 2	Cafetería	15 dB
51	Patrón con errores de clasificación 3	Cafetería	15 dB
52	Patrón con errores de clasificación 4	Cafetería	15 dB
53	Patrón con errores de clasificación 5	Cafetería	15 dB
54	Patrón con errores de clasificación 6	Cafetería	15 dB
55	Patrón con errores de clasificación 7	Cafetería	15 dB

Número	Condiciones de prueba	Ruido	S/N
56	Patrón con errores de clasificación 8	Cafetería	15 dB
57	Patrón con errores de clasificación 9	Cafetería	15 dB
58	Patrón con errores de clasificación 10	Cafetería	15 dB
59	Patrón con errores de clasificación 11	Cafetería	15 dB
60	Patrón con errores de clasificación 12	Cafetería	15 dB
61	Patrón con errores de clasificación 13	Cafetería	15 dB
62	Patrón con errores de clasificación 14	Cafetería	15 dB
63	Patrón con errores de clasificación 15	Cafetería	15 dB
64	Patrón con errores de clasificación 16	Cafetería	15 dB

Los vectores de prueba de 8 segundos que debían escuchar los sujetos se generaron de la manera siguiente:

1) Se obtuvo el archivo de marca real para cada archivo de señales vocales limpias de 8 segundos antes de añadir el ruido.

2) Se concatenaron todos los archivos de señales vocales limpias de 8 segundos con ajuste de nivel nominal en un archivo de señales vocales limpias largo.

3) Se preprocesaron el archivo de señales vocales limpias largo y los archivos de ruido.

4) Se añadieron los archivos de ruido filtrado al archivo de señales vocales filtradas largo para obtener cuatro archivos de señales vocales con ruido largos (archivos de señales vocales con ruido de tráfico (25dB), de cafetería (25dB), de tráfico (15dB) y de cafetería (15dB), respectivamente).

5) Se descompuso cada archivo de señales vocales con ruido largo en archivos con ruido de 8 segundos, manteniendo el contenido original de señal limpia.

6) Para cada archivo de señales vocales con ruido de 8 segundos, se comparó uno de los 16 archivos VAD del anexo B a la Recomendación UIT-T G.729 correspondientes con su archivo de marca real. En cada comparación, se generó un vector de prueba sustituyendo todas las tramas activas erróneamente clasificadas por su correspondiente ruido subyacente. Todos los vectores de prueba resultantes conforman la base de datos para la prueba de validación.

B.2.4 Validación de la Recomendación UIT-T P.862 para su uso en el PWMC

A partir de la base de datos subjetiva proporcionada por la prueba descrita anteriormente, se evaluó el algoritmo de la Recomendación UIT-T P.862 en tanto que medición por percepción a fin de estimar la degradación subjetiva para la condición de prueba correspondiente. El criterio de validación es que, si el resultado de la Recomendación UIT-T P.862 es monótono con respecto al valor real de la nota media de opinión en la base de datos subjetiva, puede considerarse válida la Recomendación UIT-T P.862 para la aplicación de PWMC. En el plan de la prueba se establece la necesidad de realizar el experimento al menos con dos idiomas. Se eligieron el inglés y el francés. Las figuras B-2 – B-5 que se muestran a continuación representan cuatro diagramas de dispersión, cada uno de ellos para un entorno de ruido concreto; el eje X es la puntuación de la Recomendación UIT-T P.862, mientras que el eje Y es el valor de la nota media de opinión para las 16 condiciones de la prueba en inglés.

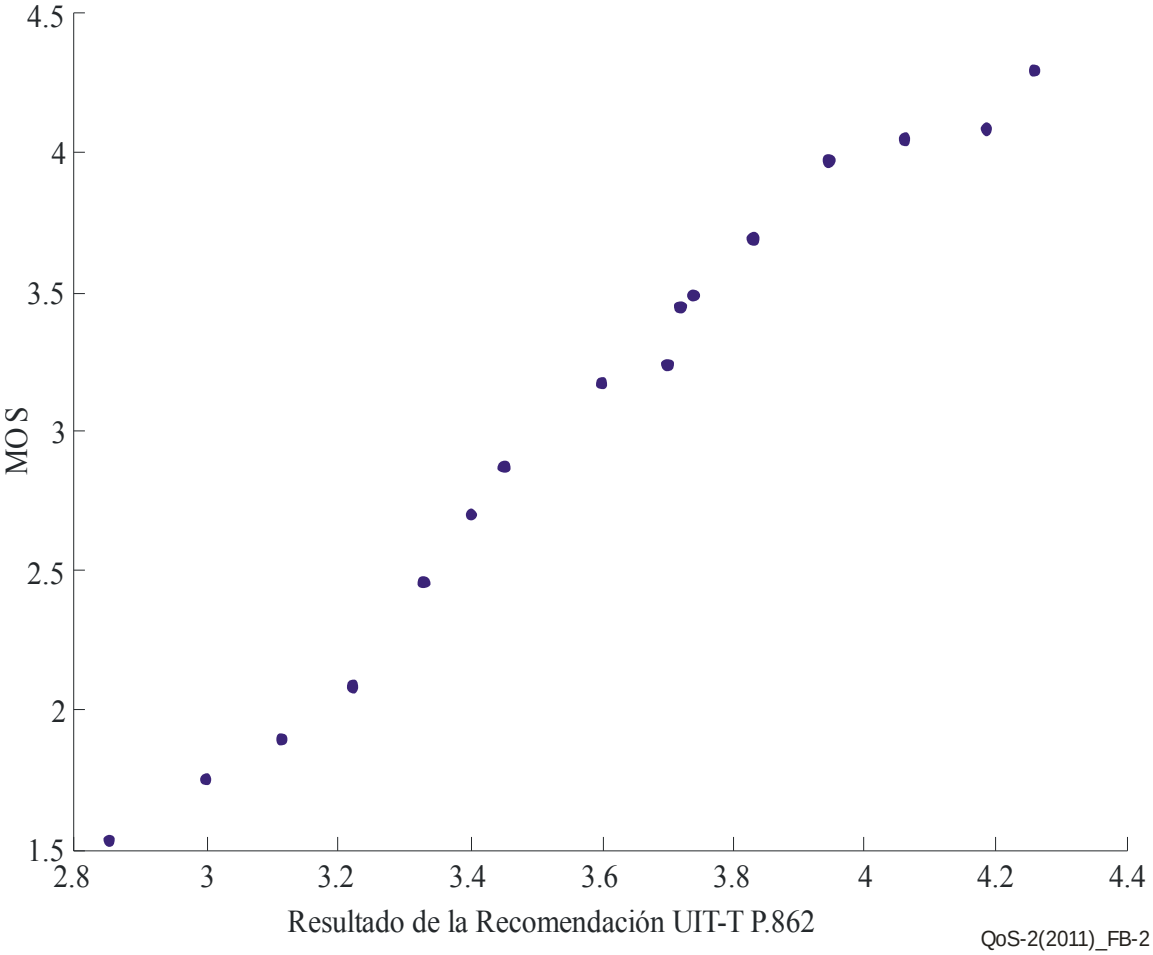


Figura B-2 – Patrones erróneamente clasificados para el entorno de ruido cafetería a 25dB (prueba en inglés)

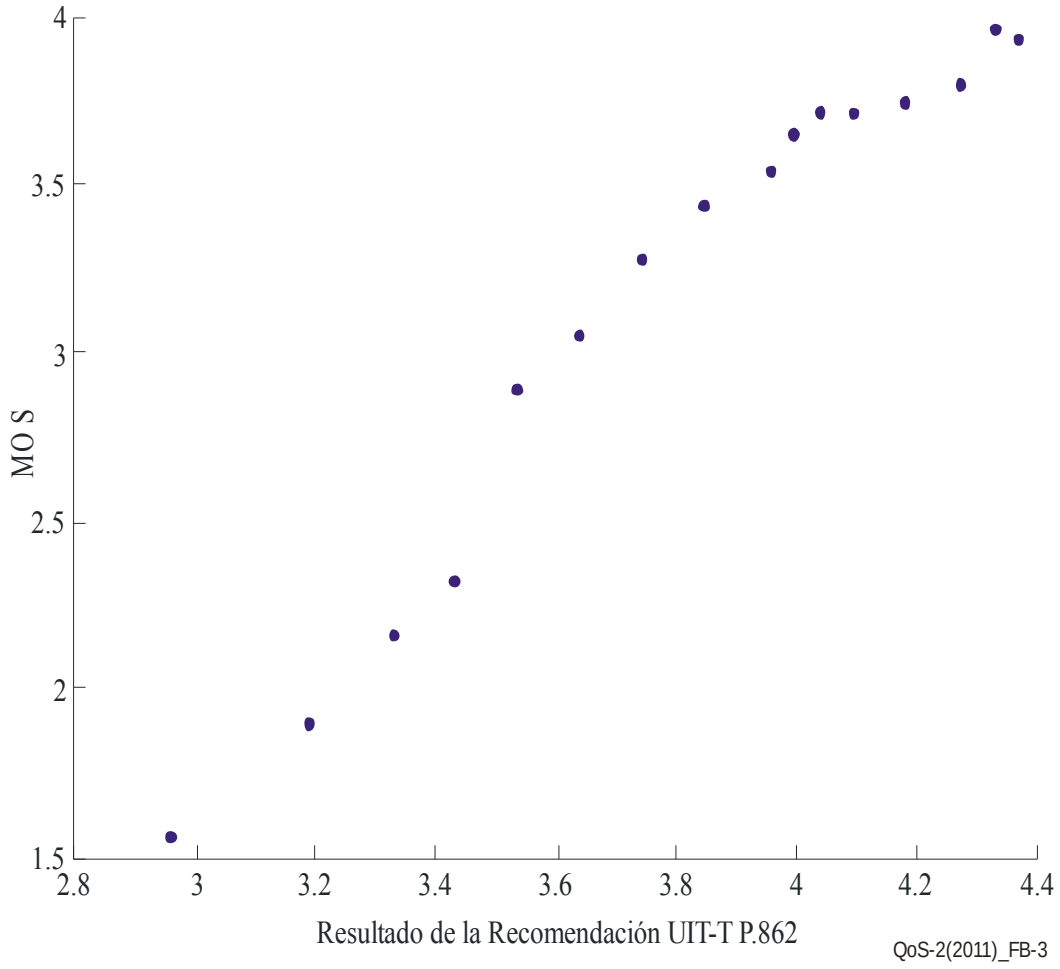


Figura B-3 – Patrones erróneamente clasificados para el entorno de ruido cafetería a 15dB (prueba en inglés)

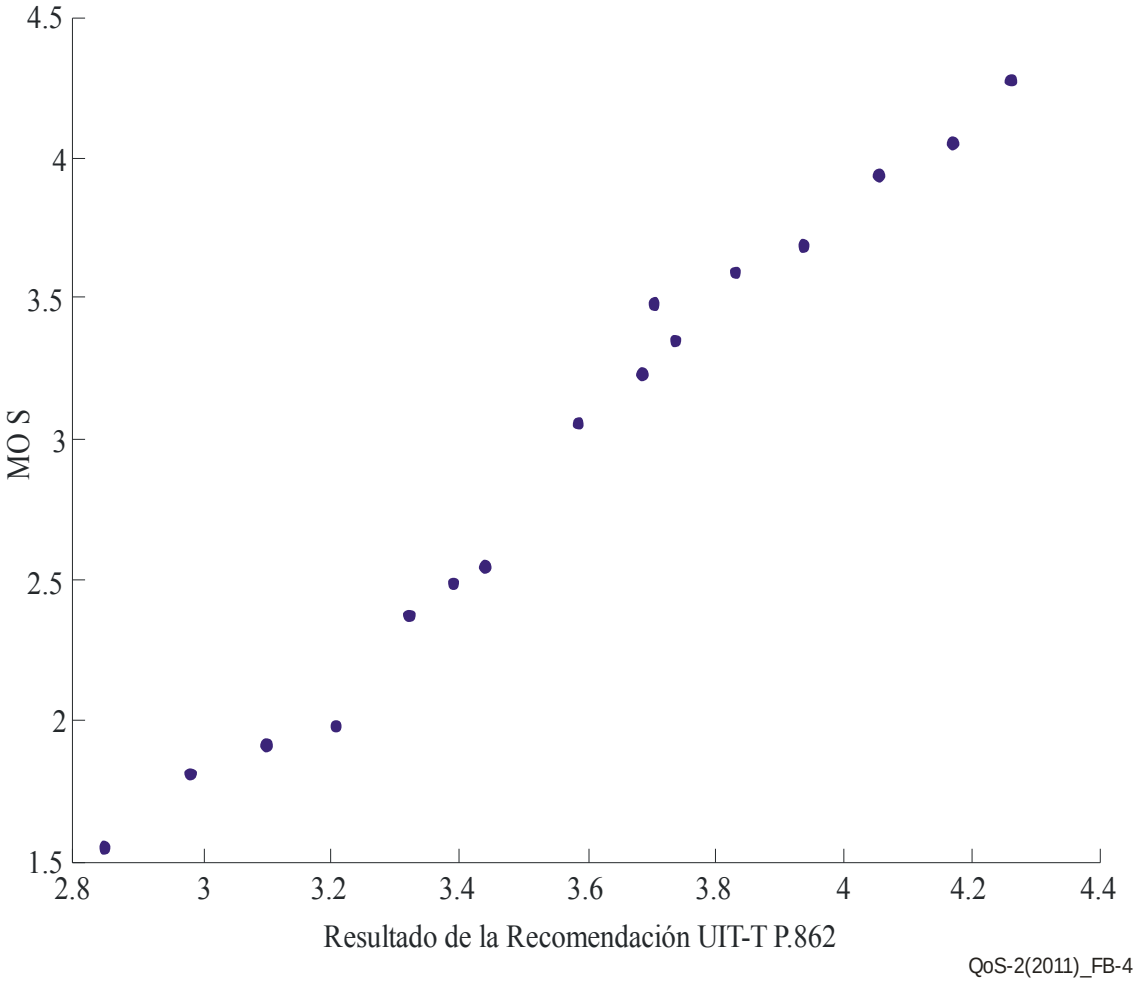


Figura B-4 – Patrones erróneamente clasificados para un entorno de ruido tráfico a 25dB (prueba en inglés)

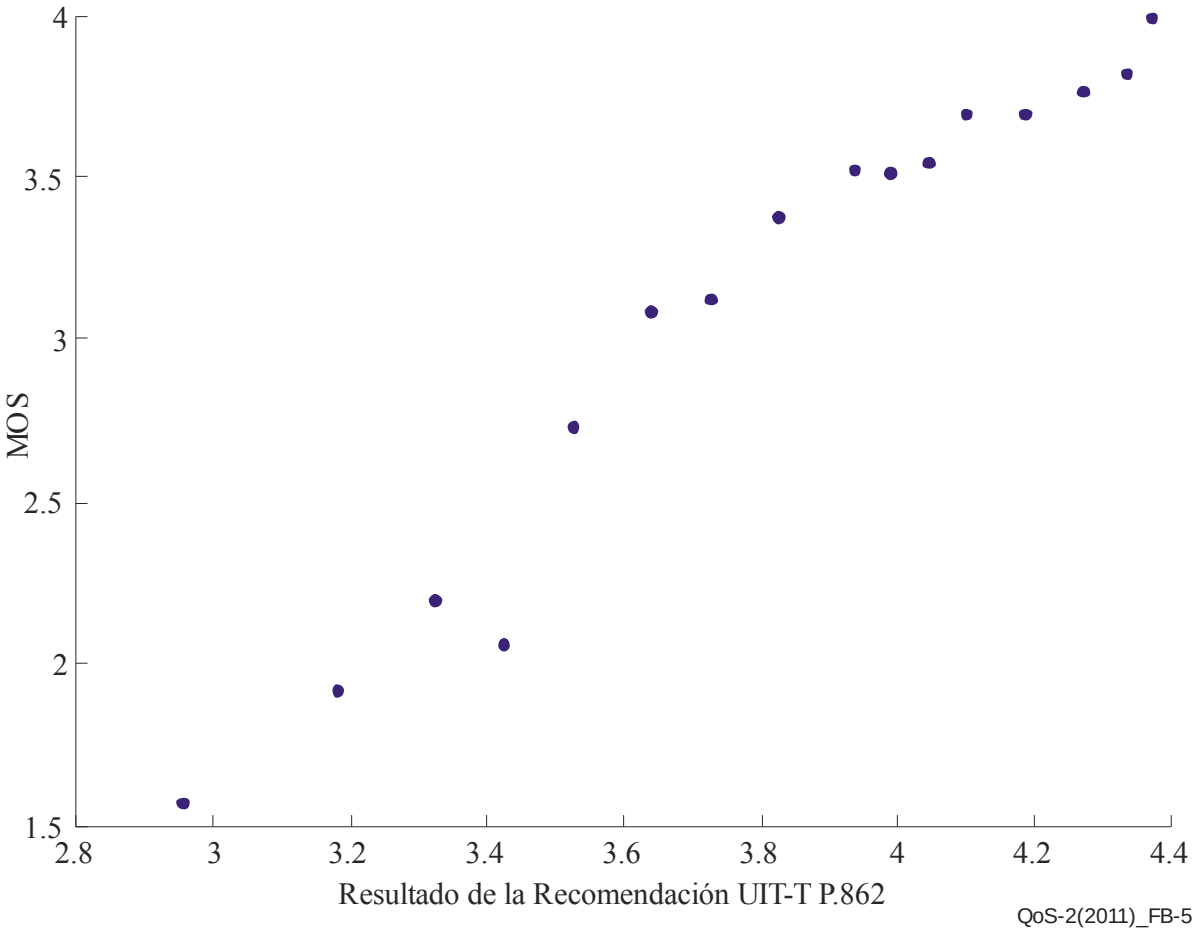


Figura B-5 – Patrones erróneamente clasificados para un entorno de ruido tráfico a 15dB (prueba en inglés)

Las figuras B-6 – B-9 que se muestran a continuación representan cuatro diagramas de dispersión, cada uno de ellos para un entorno de ruido concreto; el eje X es la puntuación de la Recomendación UIT-T P.862 mientras que el eje Y es el valor de la nota media de opinión para las 16 condiciones de la prueba en francés.

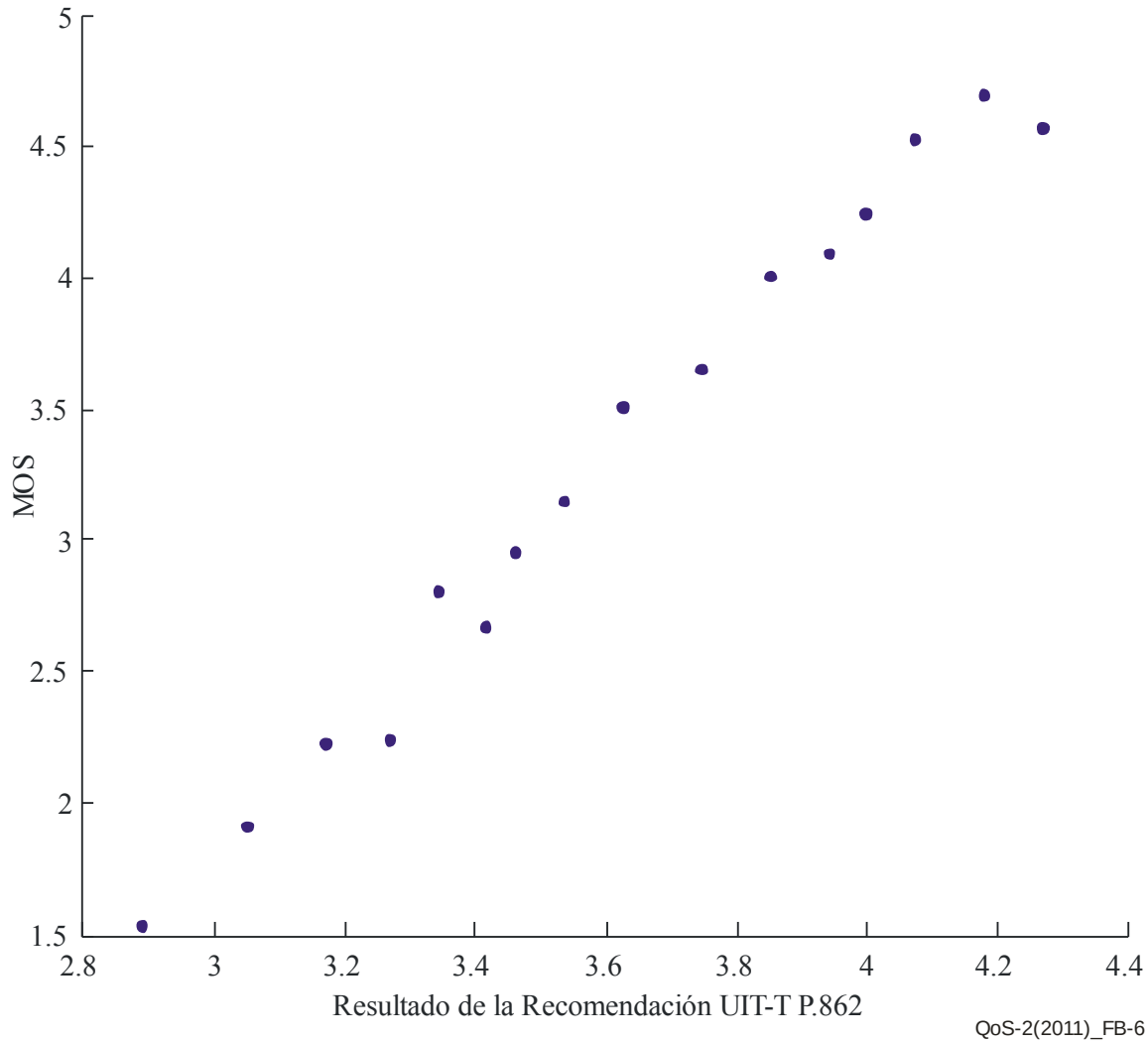


Figura B-6 – Patrones erróneamente clasificados para el entorno de ruido cafetería a 25dB (prueba en francés)

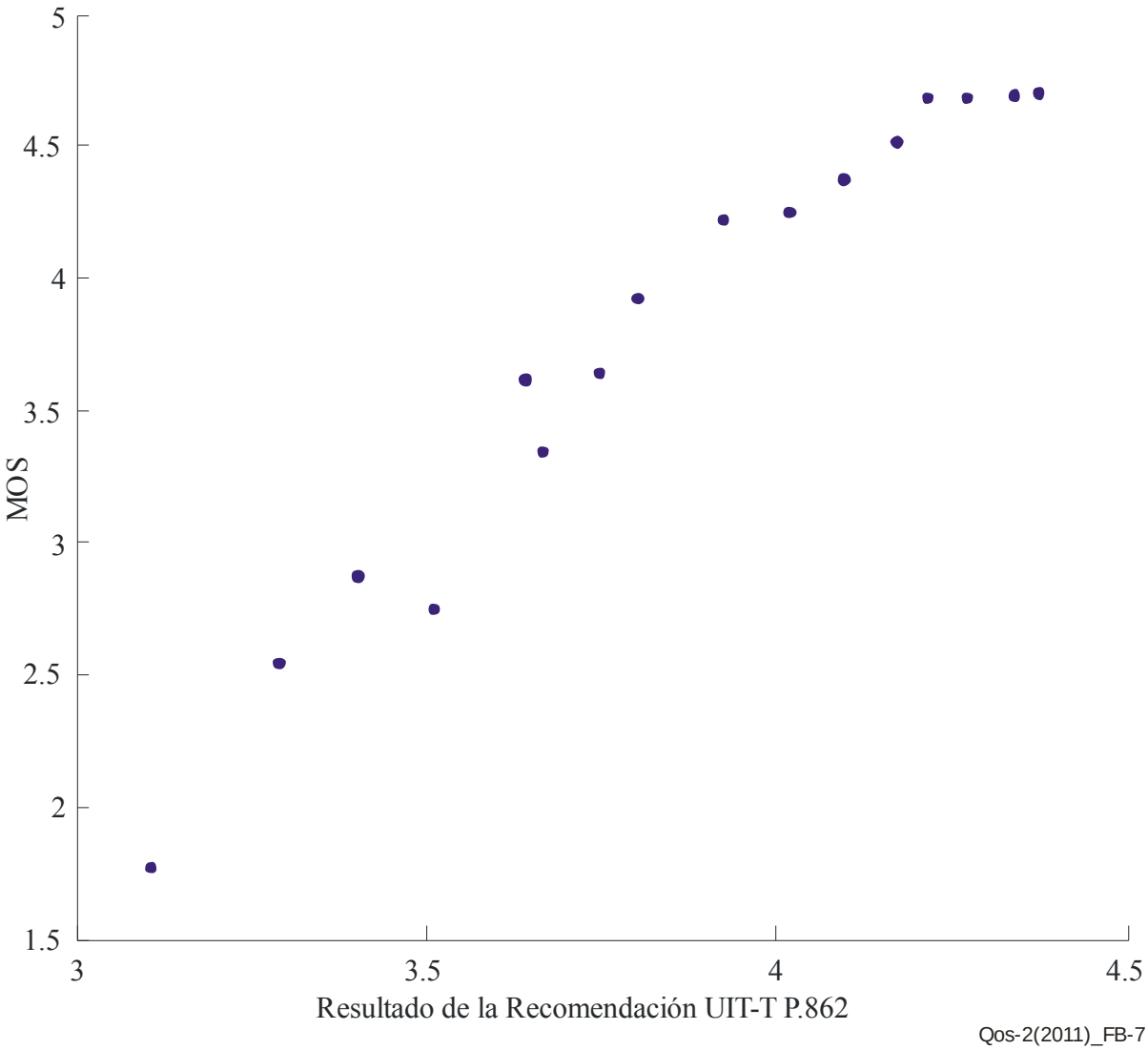


Figura B-7 – Patrones erróneamente clasificados para el entorno de ruido cafetería a a5dB (prueba en francés)

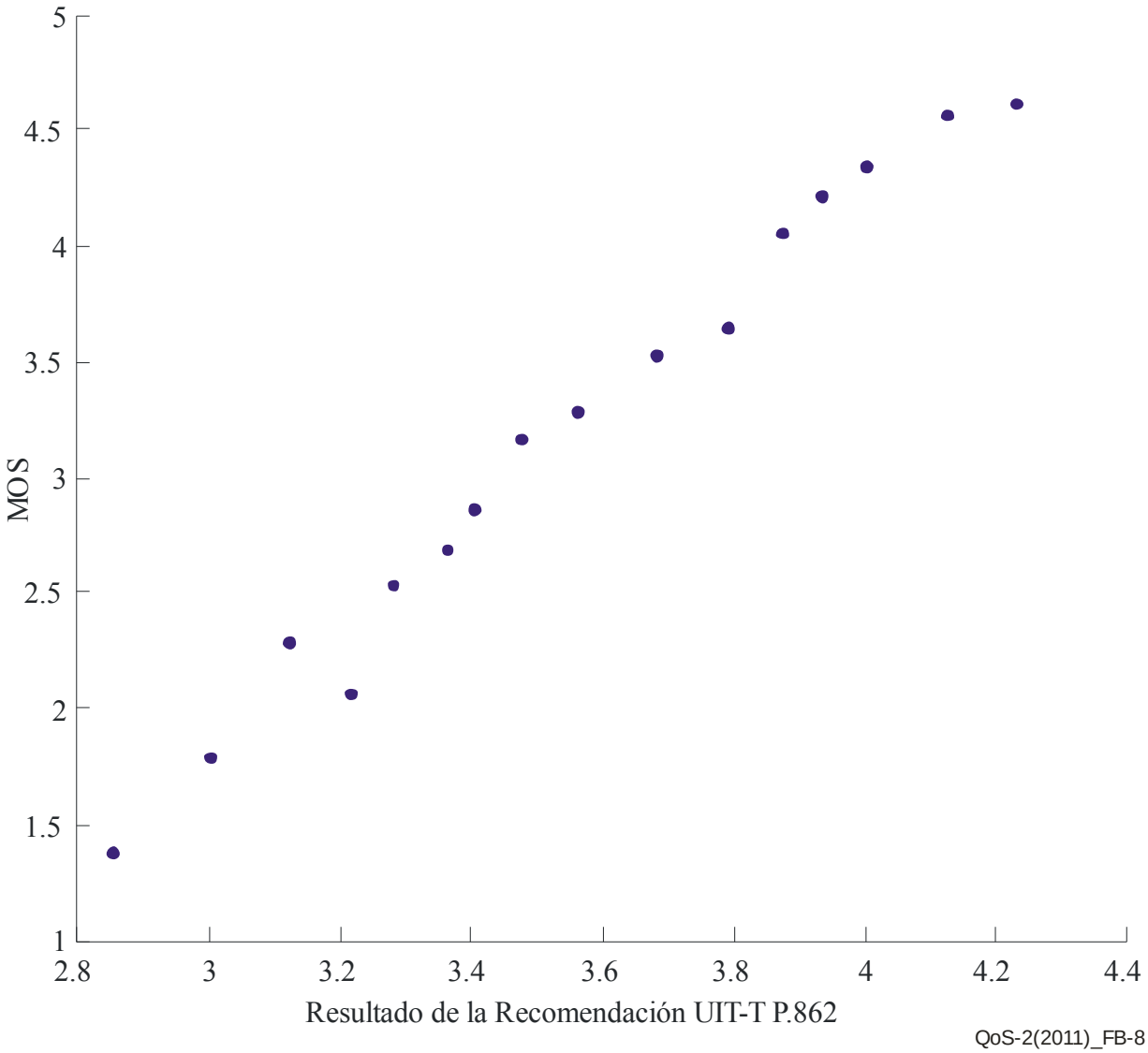


Figura B-8 – Patrones erróneamente clasificados para el entorno de ruido tráfico a 25dB (prueba en francés)

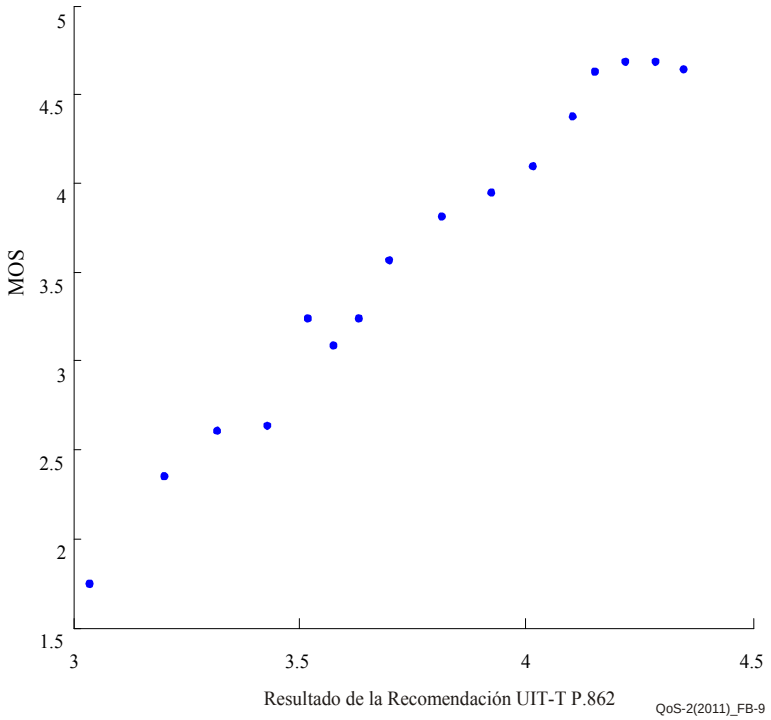


Figura B-9 – Patrones erróneamente clasificados para el entorno de ruido tráfico a 15dB (prueba en francés)

En los cuadros B-2 y B-3 que se muestran a continuación se ofrecen los coeficientes de correlación entre el resultado de la Recomendación UIT-T P.862 y el valor real de la nota media de opinión para las pruebas en inglés y en francés.

Cuadro B-2 – Correlación entre el resultado de la Recomendación UIT-T P.862 y el valor real de la nota media de opinión para la prueba en inglés

Correlación	Tráfico/25dB (MOS)	Tráfico/15dB (MOS)	Cafetería/25dB (MOS)	Cafetería/15dB (MOS)
Tráfico/25dB (UIT-T P.862)	0,991	/	/	/
Tráfico/15dB (UIT-T P.862)	/	0,974	/	/
Cafetería/25dB (UIT-T P.862)	/	/	0,993	/
Cafetería/15dB (UIT-T P.862)	/	/	/	0,946

Cuadro B-3 – Correlación entre el resultado de la Recomendación UIT-T P.862 y el valor real de MOS para la prueba en francés

Correlación	Tráfico/25dB (MOS)	Tráfico/15dB (MOS)	Cafetería/25dB (MOS)	Cafetería/15dB (MOS)
Tráfico/25dB (UIT-T P.862)	0,992	/	/	/
Tráfico/15dB (UIT-T P.862)	/	0,990	/	/
Cafetería/25dB (UIT-T P.862)	/	/	0,991	/
Cafetería/15dB (UIT-T P.862)	/	/	/	0,977

En las figuras B-2 – B-9 y los cuadros B-2 – B-3 se observa que el resultado de la Recomendación UIT-T P.862 suele ser monótono, con un valor real de la nota media de opinión prácticamente por encima de toda la escala de la nota media de opinión, desde los valores de muy baja calidad (nota media de opinión de aproximadamente 1,5) hasta los valores de muy alta calidad (nota media de opinión de aproximadamente 4,4). Solamente algunos puntos parecen lejanos, y la correlación entre el resultado de la Recomendación UIT-T P.862 y el valor real de la nota media de opinión es ciertamente fuerte. Así, puede validarse la Recomendación UIT-T P.862 para la aplicación de PWMC.

Apéndice 1

Procedimiento de sustitución de la trama activa erróneamente clasificada

Los puntos de acceso a la fase de sustitución de la trama activa erróneamente clasificada son: a) el archivo de marca real para el archivo concatenado largo; b) el archivo de marca de GSAD para el archivo de señales vocales con ruido largo; c) el archivo de señales vocales concatenado largo y d) el archivo de ruido procesado con ajuste de nivel. El resultado de esta fase es el archivo de señales vocales con ruido concatenado y modificado largo.

En el procesamiento, se compara el archivo de marca GSAD con el archivo de marca real, lo que permite identificar las ubicaciones en las que se producen errores de clasificación de tramas activas e inactivas. Las tramas de las ubicaciones anteriormente mencionadas del archivo de señales vocales concatenado largo se modifican por multiplicación mediante el establecimiento de una ventana de aumento y disminución gradual de la intensidad de la señal, WINs. Las tramas de las ubicaciones anteriormente mencionadas del archivo de ruido procesado con ajuste de nivel se modifican por multiplicación mediante el establecimiento de otra ventana de aumento y disminución gradual de la intensidad de la señal, WINn. Las tramas en otras ubicaciones, tanto en el archivo de señales vocales como en el archivo de ruido, no se modifican. El archivo de señales vocales concatenado y modificado largo y el archivo de ruido procesado modificado se suman aritméticamente, hecho que da como resultado el archivo de señales vocales con ruido concatenado y modificado largo que se introducirá en la fase de evaluación de la Recomendación UIT-T P.862. A continuación se proporcionan los dos conjuntos de ventanas anteriormente mencionadas:

1) Si no se modifican ni las tramas anteriores ni las posteriores de la trama activa erróneamente clasificada real,

WINs(i) (para señales de entrada de 8kHz) = {1, 0.9823, 0.8763, 0.6845, 0.4258, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0.4258,0.6845,0.8763,0.9823,1}

WINs(i) (para señales de entrada de 16kHz) = {1,0.9921,0.9823,0.9409,0.8763,0.7902,0.6845,0.5621,
0.4258,0.279,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.279, 0.4258,0.5621, 0.6845,
0.7902, 0.8763, 0.9409, 0.9823, 0.9921,1}

WINn(i) (para señales de entrada de 8kHz) = 1-WINs(i), para i=0,1,2,…79

WINn(i) (para señales de entrada de 16kHz) = 1-WINs(i), para i=0,1,2,…159

2) Si se modifica la trama anterior de la trama activa erróneamente clasificada real pero no se modifica la trama posterior de la trama activa erróneamente clasificada real,

WINs(i) (para señales de entrada de 8kHz) =
{0,0,00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.4258,0.6845,0.8763,0.9823,1}

WINs(i) (para señales de entrada de 16kHz) =
{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.279,0.4258,0.5621, 0.6845, 0.7902, 0.8763,
0.9409, 0.9823, 0.9921,1}

WINn(i) (para señales de entrada de 8kHz) = 1-WINs(i), para i=0,1,2,…79

WINn(i) (para señales de entrada de 16kHz) = 1-WINs(i), para i=0,1,2,…159

3) Si no se modifica la trama anterior de la trama activa erróneamente clasificada real pero se modifica la trama posterior,

WINs(i) (para señales de entrada de 8kHz) = {1,0.9823,0.8763,0.6845,0.4258,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}

WINs(i) (para señales de entrada de 16kHz) = {1, 0.9921, 0.9823, 0.9409, 0.8763, 0.7902, 0.6845, 0.5621, 0.4258,0.279,,0}

WINn(i) (para señales de entrada de 8kHz) = 1-WINs(i), para i=0,1,2,…79

WINn(i) (para señales de entrada de 16kHz) = 1-WINs(i), para i=0,1,2,…159

4) Si se modifican la trama anterior y la posterior de la trama activa erróneamente clasificada real,

WINs(i) (para señales de entrada de 8kHz) = {0}, para i=0, 1,2…79

WINs(i) (para señales de entrada de 16kHz) = {0}, para i=0, 1,2…159

WINn(i) (para señales de entrada de 8kHz) = {1}, para i=0, 1,2…79

WINn(i) (para señales de entrada de 16kHz) = {1}, para i=0, 1,2…159

Bibliografía

Recomendación UIT-T G.191 (2010), *Herramientas de soporte lógico para la normalización de la codificación de señales vocales y de audio*.

Recomendación UIT-T G.720.1 (2010), *Detector genérico de actividad sonora*.

Recomendaciones UIT-T Serie P: *Terminales y métodos de evaluación subjetivos y objetivos*.

Recomendación UIT-T P.10 (1993), *Vocabulario de términos sobre calidad de transmisión telefónica y aparatos telefónicos*.

Recomendación UIT-T P.11 (1993), *Efectos de las degradaciones de la transmisión*.

Recomendación UIT-T P.48 (1988), *Especificación de un sistema intermedio de referencia*.

Recomendación UIT-T P.50 (1993), *Voces artificiales*.

Recomendación UIT-T P.56 (1993), *Medición objetiva del nivel vocal activo*.

Recomendación UIT-T P.70 (1984) Suplemento Nº 14, *Evaluación subjetiva de la calidad de funcionamiento de procesos digitales utilizando la unidad de referencia de ruido modulado (MNRU).*

Recomendación UIT-T P.74 (1984) Suplemento Nº 5, *Método de prueba subjetiva SIBYL*.

Recomendación UIT-T P.78 (1993), *Método de prueba subjetivo para determinar índices de sonoridad de acuerdo con la Recomendación P.76*.

Recomendación UIT-T P.80 (1993), *Métodos de determinación subjetiva de la calidad de transmisión*.

Recomendación UIT-T P.81 (1989), *Aparato de referencia para ruido modulado (MNRU)*.

Recomendación UIT-T P.82 (1984), *Método para la evaluación del servicio desde el punto de vista de calidad de transmisión de la palabra*.

Recomendación UIT-T P.85 (1994), *Método para la evaluación subjetiva de la calidad vocal de los dispositivos generadores de voz*.

Recomendación UIT-T P.800 (1996), *Métodos de determinación subjetiva de la calidad de transmisión*.

Recomendación UIT-T P.805 (2007), *Evaluación subjetiva de la calidad de la conversación*.

Recomendación UIT-T P.810 (1996), *Aparato de referencia para ruido modulado (MNRU)*.

Recomendación UIT-T P.830 (1996), *Evaluación subjetiva de la calidad de funcionamiento de los códecs digitales de banda telefónica y de banda ancha*.

Recomendación UIT-T P.832 (2000), *Evaluación subjetiva de la calidad de funcionamiento de los terminales manos libres*.

Recomendación UIT-T P.835 (2003), *Metodología de prueba subjetiva para evaluar los sistemas de comunicación vocal que utilizan un algoritmo de cancelación de ruido*.

Recomendación UIT-T P.840 (2003), *Método de prueba de escucha subjetiva para la evaluación de equipos de multiplicación de circuitos*.

Recomendación UIT-T P.851 (2003), *Evaluación de la calidad subjetiva de los servicios telefónicos basados en sistemas conversacionales*.

Recomendación UIT-T P.862 (2001), *Evaluación de la calidad vocal por percepción: Un método objetivo para la evaluación de la calidad vocal de extremo a extremo de redes telefónicas de banda estrecha y códecs vocales*.

Recomendación UIT-T P.880 (2004), *Evaluación continúa de la calidad vocal que varía con el tiempo*.

Cygwin, *artículo de Wikipedia sobre Cygwin*. <http://en.wikipedia.org/wiki/Cygwin> (visitado el 9 de abril de 2010).

Fay, M.P.; Proschan, M.A. (2010), *Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules.* Statistics Surveys 4: pp. 1-39.

Kirk, R.E. (1982), *Experimental Design: Procedures for the Behavioral Sciences*, 2nd edition. Brooks/Cole Publishing Company.

Mankiewicz, Richard (2004), *The Story of Mathematics*, Princeton University Press, p. 158.

Unión Internacional de Telecomunicaciones
Place des Nations
CH-1211 Ginebra 20
Suiza

www.itu.int

36624

Impreso en Suiza
Ginebra, 2012
ISBN 92-61-13793-8