UNIÓN INTERNACIONAL DE TELECOMUNICACIONES

UIT-T P.861

(02/98)

SECTOR DE NORMALIZACIÓN
DE LAS TELECOMUNICACIONES
DE LA UIT

SERIE P: CALIDAD DE TRANSMISIÓN TELEFÓNICA, INSTALACIONES TELEFÓNICAS Y REDES LOCALES

Métodos de evaluación objetiva y subjetiva de la calidad

Medición objetiva de la calidad de los códecs vocales de banda telefónica (300-3400 Hz)

Recomendación UIT-T P.861

(Anteriormente Recomendación del CCITT)

RECOMENDACIONES DE LA SERIE P DEL UIT-T

CALIDAD DE TRANSMISIÓN TELEFÓNICA, INSTALACIONES TELEFÓNICAS Y REDES LOCALES

Vocabulario y efectos de los parámetros de transmisión sobre la opinión de los clientes	Serie	P.10
Líneas y aparatos de abonado	Serie	P.30
		P.300
Patrones de transmisión	Serie	P.40
Aparatos para mediciones objetivas	Serie	P.50
		P.500
Medidas electroacústicas objetivas	Serie	P.60
Medidas relativas a la sonoridad vocal	Serie	P.70
Métodos de evaluación objetiva y subjetiva de la calidad	**Serie**	**P.80**
		P.800
Calidad audiovisual en servicios multimedios	Serie	P.900

Para más información, véase la Lista de Recomendaciones del UIT-T.

RECOMENDACIÓN UIT-T P.861

MEDICIÓN OBJETIVA DE LA CALIDAD DE LOS CÓDECS VOCALES DE BANDA TELEFÓNICA (300-3400 Hz)

Resumen

La presente Recomendación describe un método objetivo para la estimación subjetiva de la calidad de los códecs vocales de banda telefónica (300-3400 Hz).

La presente Recomendación especifica la producción de señales vocales fuente para la medición objetiva de la calidad, las condiciones del códec y de referencia para las cuales se ha demostrado que el método de medición objetiva de la calidad proporciona resultados válidos, el cálculo objetivo de la calidad basado en la medida objetiva de calidad denominada medición perceptual de la calidad vocal (PSQM, *perceptual speech quality measure*), la estimación subjetiva de la calidad a partir de los resultados de la medición objetiva, y un análisis de los resultados.

La presente Recomendación se puede aplicar cuando se evalúan por métodos subjetivos los efectos de los niveles de entrada de la señal vocal, los participantes, las velocidades binarias y las transcodificaciones sobre la calidad.

Orígenes

La Recomendación UIT-T P.861, ha sido revisada por la Comisión de Estudio 12 (1997-2000) del UIT-T y fue aprobada por el procedimiento de la Resolución N.° 1 de la CMNT el 27 de febrero de 1998.

Palabras clave

Evaluación subjetiva de la calidad vocal, medición objetiva de la calidad vocal.

PREFACIO

La UIT (Unión Internacional de Telecomunicaciones) es el organismo especializado de las Naciones Unidas en el campo de las telecomunicaciones. El UIT-T (Sector de Normalización de las Telecomunicaciones de la UIT) 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.

La Conferencia Mundial de Normalización de las Telecomunicaciones (CMNT), que se celebra cada cuatro años, establece los temas que han de estudiar las Comisiones de Estudio del UIT-T, que a su vez producen Recomendaciones sobre dichos temas.

La aprobación de Recomendaciones por los Miembros del UIT-T es el objeto del procedimiento establecido en la Resolución N.° 1 de la CMNT.

En ciertos sectores de la tecnología de la información que corresponden a la esfera de competencia del UIT-T, se preparan las normas necesarias en colaboración con la ISO y la CEI.

NOTA

En esta Recomendación, la expresión "Administración" se utiliza para designar, en forma abreviada, tanto una administración de telecomunicaciones como una empresa de explotación reconocida de telecomunicaciones.

PROPIEDAD INTELECTUAL

La UIT señala a la atención la posibilidad de que la utilización o aplicación de la presente Recomendación suponga el empleo de un derecho de propiedad intelectual reivindicado. La UIT no adopta ninguna posición en cuanto a la demostración, validez o aplicabilidad de los derechos de propiedad intelectual reivindicados, ya sea por los miembros de la UIT o por terceros ajenos al proceso de elaboración de Recomendaciones.

En la fecha de aprobación de la presente Recomendación, la UIT ha recibido notificación de propiedad intelectual, protegida por patente, que puede ser necesaria para aplicar esta Recomendación. Sin embargo, debe señalarse a los usuarios que puede que esta información no se encuentre totalmente actualizada al respecto, por lo que se les insta encarecidamente a consultar la base de datos sobre patentes de la TSB.

© UIT 1998

Es propiedad. Ninguna parte de esta publicación puede reproducirse o utilizarse, de ninguna forma o por ningún medio, sea éste electrónico o mecánico, de fotocopia o de microfilm, sin previa autorización escrita por parte de la UIT.

ÍNDICE

		Página
1	Alcance	1
2	Referencias	3
3	Abreviaturas	3
4	Definiciones	3
5	Convenios	4
6	Resumen del procedimiento de medición objetiva	4
7	Preparación del material de la señal vocal fuente	4
7.1	Voces reales	5
7.2	Voces artificiales	5
8	Selección de parámetros experimentales	5
9	Cálculo objetivo de la calidad	6
9.1	Inicializaciones globales	13
9.1.1	Alineación de tiempo	13
9.1.2	Aplicación de escala global	13
9.1.3	Calibración global	14
9.2	Correspondencia tiempo-frecuencia	14
9.2.1	Disposición en ventana	15
9.2.2	Densidad de potencia espectral (SPD, *spectral power density)* muestreada	15
9.3	Transposición y filtrado de frecuencia	15
9.3.1	Densidad de potencia de tono muestreada	15
9.3.2	Aplicación de escala local	16
9.3.3	Filtrado de la banda telefónica	16
9.3.4	Ruido de Hoth	17
9.4	Transposición de intensidad	17
9.5	Modelado cognitivo	18
9.5.1	Escala de sonoridad	18
9.5.2	Densidad de la perturbación debida al ruido muestreada	18
9.5.3	Tratamiento de la asimetría	18
9.5.4	Perturbación debida al ruido que incluye el tratamiento del intervalo de silencio	19
10	Transformación de la escala de calidad objetiva a la escala de calidad subjetiva	20
10.1	Notas medias de opinión	20
10.2	Valores Q equivalentes	20

Página

11 Análisis de resultados 21

Apéndice I – Contenido del disquete que acompaña a la Recomendación P.861 22

I.1 Introducción 22

I.2 Directorio\test 22

Apéndice II – Medición objetiva de la calidad de los códecs vocales de banda telefónica (300-3400 Hz), mediante bloques de normalización de la medición (MNB, *measuring normalizing slocks*) 24

II.1 Introducción 24

II.2.1 Especificaciones de entrada-salida 27

II.2.2 Retardo de tiempo 27

II.2.3 Preparación de la señal 28

II.2.4 Transformación al dominio de la frecuencia 28

II.2.5 Selección de tramas 29

II.2.6 Aproximación de la sonoridad percibida 29

II.2.7 Bloque de normalización de la medición de frecuencias (FMNB) 29

II.2.8 Cálculo de bloques de normalización de la medición del tiempo 29

II.2.9 Combinación lineal de mediciones para la estructura de MNB 31

Introducción

La evaluación subjetiva de la calidad de los códecs vocales es una de las tecnologías clave utilizadas en el diseño de las redes de telecomunicación digitales. En la Recomendación P.830 se definen metodologías de pruebas subjetivas para códecs vocales. Dado que la evaluación subjetiva de la calidad requiere mucho tiempo y no es económica, es deseable desarrollar una metodología objetiva de evaluación de la calidad para la estimación subjetiva de la calidad de los códecs vocales con pruebas menos subjetivas.

La medida objetiva de la calidad vocal más utilizada para demostrar la calidad de funcionamiento de los códecs vocales es la relación señal/ruido (SNR = S/N, *signal to noise ratio*). No obstante, es necesario señalar que la SNR no predice adecuadamente la calidad subjetiva (calidad evaluada subjetivamente) de los componentes de red modernos. Esto es particularmente cierto con respecto a los últimos códecs de baja velocidad binaria. Por lo tanto, se han elaborado una variedad de medidas objetivas de la calidad más especializadas, tales como la medida de la distancia Cepstrum (CD, *Cepstrum distance measure*) de codificación de predicción lineal (LPC, *linear predictive coding*) [1], el índice de información (II, *information index*) [2], la función de coherencia (CHF, *coherence function*) [3], el reconocimiento de diagramas (EPR, *expert pattern recognition*) [4], y la medición perceptual de la calidad vocal (PSQM, *perceptual speech quality measure*) [5]. Desde el decenio de 1980 el UIT-T investiga el comportamiento de dichos sistemas en cuanto a su capacidad para ofrecer evaluaciones objetivas precisas de la calidad.

Tras efectuar cuidadosas comparaciones entre dichas medidas objetivas de la calidad, se llegó a la conclusión de que la PSQM era la que mejor se correlacionaba con la calidad subjetiva de la señal vocal codificada. Por consiguiente, en esta Recomendación se describe la evaluación objetiva de la calidad tomando la PSQM como medida objetiva de la calidad [12].

A fin de ayudar a los lectores de esta Recomendación a desarrollar su propia implementación del PSQM, se ha incluido un disquete con esta Recomendación. Puede verse una descripción del contenido de este disquete en el fichero `README` del disquete y en el apéndice I.

Recomendación P.861

MEDICIÓN OBJETIVA DE LA CALIDAD DE LOS CÓDECS VOCALES DE BANDA TELEFÓNICA (300-3400 Hz)

(revisada en 1998)

1 Alcance

La evaluación subjetiva de la calidad de los códecs vocales se puede efectuar en pruebas de escucha solamente (unidireccionales) o en pruebas de conversación (bidireccionales). La medición objetiva de la calidad descrita en esta Recomendación evalúa la calidad subjetiva en pruebas de escucha solamente.

Para demostrar la calidad subjetiva de funcionamiento de un códec, se deben investigar los efectos de una variedad de factores de calidad (véase la Recomendación P.830). No se ha verificado la precisión de la medición objetiva de la calidad descrita en esta Recomendación para el examen de todos los factores especificados en la Recomendación P.830. El cuadro 1 sirve de orientación para facilitar la determinación de los factores de prueba por parte de los lectores, las tecnologías de codificación y las aplicaciones cubiertas por la presente Recomendación.

Cuadro 1/P.861 – Relación de las tecnologías de codificación, los factores experimentales y las aplicaciones a esta Recomendación

Factores de prueba	Nota
Niveles de entrada de la señal vocal a un códec	1
Niveles de escucha en experimentos subjetivos	2
Factores de dependencia de los participantes	1
Múltiples participantes simultáneos	2
Errores en el canal de transmisión	2
Velocidades binarias si el códec tiene más de un modo de velocidad binaria	1
Transcodificaciones	1
Desadaptación de la velocidad binaria entre codificador y decodificador si el códec tiene más de un modo de velocidad binaria	2
Ruido ambiental en el lado de emisión	2
Señales de información de la red como entradas a un códec	2
Música como entrada a un códec	2
Retardo	3
Transposición temporal a corto plazo de la señal audio	2
Transposición temporal a largo plazo de la señal audio	4
Recorte en tiempo de la señal vocal	2
Recorte en amplitud de la señal vocal	2

Cuadro 1/P.861 – Relación de las tecnologías de codificación, los factores experimentales y las aplicaciones a esta Recomendación *(fin)*

Tecnologías de codificación	Nota
Forma de onda	1
CELP e híbridos ≥ 4 kbit/s	1
CELP e híbridos < 4 kbit/s	2
VOCODER (codificadores de señales vocales)	2
Otros codificadores	2
Aplicaciones	
Optimización del codificador	1
Evaluación del codificador	1
Selección del codificador	2
Planificación de la red	5
Pruebas de red reales	6
Dispositivos de medición que no ocasionan interrupción del servicio	3

NOTA 1 – La medida objetiva ha demostrado una precisión aceptable en presencia de esta variable.

NOTA 2 – Se dispone de información insuficiente sobre la precisión de la medida objetiva con respecto a esta variable.

NOTA 3 – Es sabido que la medida objetiva proporciona predicciones imprecisas cuando se la utiliza junto con esta variable o, en otros casos, cuando no está previsto utilizarla con dicha variable.

NOTA 4 – Es sabido que la medida objetiva proporciona predicciones imprecisas cuando hay una fluctuación lenta de fase apreciable (más del 10% de la longitud de trama). Queda en estudio la aplicabilidad de la medida cuando la fluctuación es pequeña.

NOTA 5 – Con las precauciones necesarias, la medida objetiva podría ser utilizada para ciertos fines relativos a la planificación de la red. Cabe observar al lector que existen importantes factores en la planificación de la red a los que no es aplicable la presente Recomendación (véase la sección "Factores de prueba" de este cuadro).

NOTA 6 – Con las precauciones necesarias, la medida objetiva podría ser utilizada para algunas pruebas de red reales. Cabe observar al lector que pueden existir factores o tecnologías en una conexión de red real a los cuales no es aplicable la presente Recomendación (véanse las secciones "Factores de prueba" y "Tecnologías de codificación" de este cuadro).

Al comparar un códec con otro códec o con una condición de referencia basada en resultados experimentales subjetivos, suelen utilizarse pruebas estadísticas que tienen en cuenta las distribuciones de las apreciaciones subjetivas. Dado que la medición objetiva indicada en la presente Recomendación evalúa únicamente la media de las apreciaciones subjetivas (por ejemplo, MOS, DMOS), dichas pruebas estadísticas no pueden aplicarse a los resultados de la medición objetiva. Se estudia actualmente la predicción del porcentaje de mediocre o peor (%PoW, *per cent poor or worse*) y del porcentaje de bueno o mejor (%GoB, *per cent good or better*).

2 Referencias

Las siguientes Recomendaciones del UIT-T y otras referencias contienen disposiciones que, mediante su referencia en este texto, constituyen disposiciones de la presente Recomendación. Al efectuar esta publicación, estaban en vigor las ediciones indicadas. Todas las Recomendaciones y otras referencias son objeto de revisiones por lo que se preconiza que los usuarios de esta Recomendación investiguen la posibilidad de aplicar las ediciones más recientes de las Recomendaciones y otras referencias citadas a continuación. Se publica periódicamente una lista de las Recomendaciones UIT-T actualmente vigentes.

– Recomendación G.711 del CCITT (1988), *Modulación por impulsos codificados (MIC) de frecuencias vocales*.

– Recomendación G.726 del CCITT (1990), *Modulación por impulsos codificados diferencial adaptativa (MICDA) a 40, 32, 24, 16 kbit/s*.

– Recomendación G.728 del CCITT (1992), *Codificación de señales vocales a 16 kbit/s utilizando predicción lineal con excitación por código de bajo retardo*.

– Recomendación UIT-T G.729 (1996), *Codificación de la voz a 8 kbit/s mediante predicción lineal con excitación por código algebraico de estructura conjugada*.

– Recomendación UIT-T P.50 (1993), *Voz artificial*.

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

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

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

– Suplemento N.º 13 de las Recomendaciones del CCITT de la serie P (1988), *Espectros de ruido*.

3 Abreviaturas

En esta Recomendación se utilizan las siguientes siglas.

ACR Índices por categorías absolutas (*absolute category rating*)

CELP Predicción lineal con excitación por código (c*ode excited linear prediction*)

DCR Índices por categorías de degradación (*degradation category rating*)

DMOS Nota media de opinión sobre las degradaciones (*degradation mean opinion score*)

MOS Nota media de opinión (*mean opinion score*)

PSQM Medición perceptual de la calidad vocal (*perceptual speech quality measure*)

4 Definiciones

En esta Recomendación se define el término siguiente.

4.1 **dBov**: dB relativo al nivel de saturación de un sistema digital.

5 Convenios

La evaluación subjetiva de los códecs vocales se puede realizar utilizando métodos de pruebas subjetivas de escucha solamente o de conversación. En la práctica, las pruebas de escucha solamente constituyen el único método factible de prueba subjetiva durante el desarrollo de códecs vocales, cuando no se dispone de una implementación en tiempo real del códec. En esta Recomendación se examina una técnica de medición objetiva para la evaluación subjetiva de la calidad obtenida en pruebas de escucha solamente.

6 Resumen del procedimiento de medición objetiva

La figura 1 ilustra el procedimiento de medición objetiva.

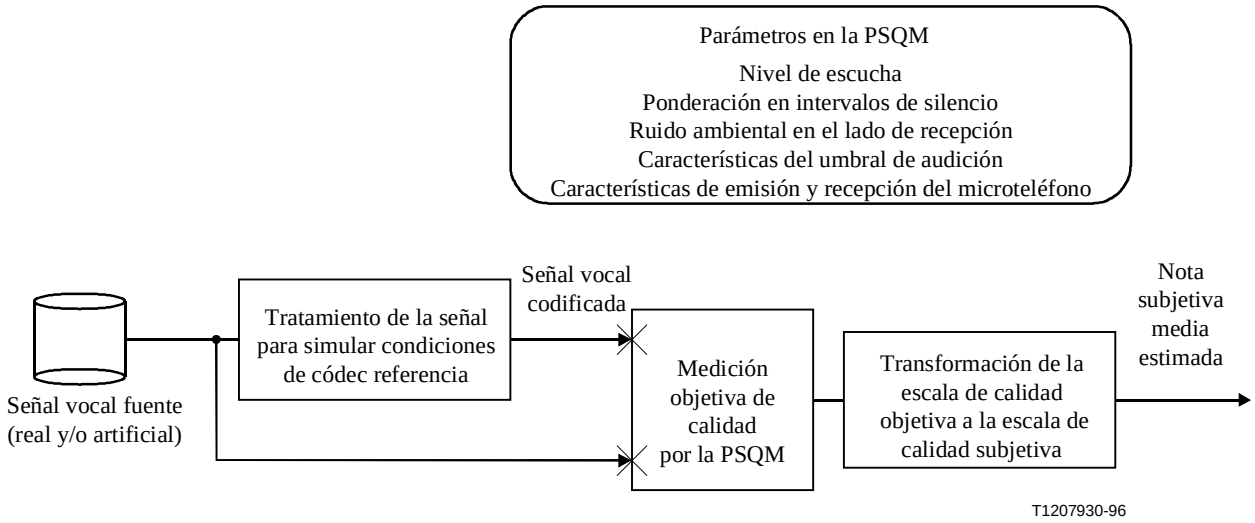

Figura 1/P.861 – Procedimiento de medición objetiva de la calidad

La medición objetiva de la calidad de los códecs vocales requiere numerosas etapas:

1) Preparación de los materiales de origen, es decir, grabación de participantes y/o generación de las voces artificiales de conformidad con la Recomendación P.50.
2) Selección de parámetros experimentales que presentarán las características dominantes del códec y podrán ser probados mediante la medición objetiva.
3) Producción de señales codificadas/de referencia.
4) Cálculo objetivo de la calidad vocal basado en la medida perceptual de la calidad vocal (PSQM), utilizando señales fuente y codificadas.
5) Transformación de la escala de calidad objetiva (calidad medida objetivamente) en la escala de calidad subjetiva (calidad evaluada subjetivamente), en caso necesario.
6) Análisis de los resultados.

A continuación se describe cada una de estas etapas.

7 Preparación del material de la señal vocal fuente

Las señales fuente para la medición objetiva pueden ser voces reales o las voces artificiales especificadas en la Recomendación P.50, en función de los objetivos del experimento.

Dado que las voces artificiales definidas en la Recomendación P.50 reproducen las principales características de la voz humana en diversos idiomas, dichas voces son útiles para evaluar objetivamente la calidad subjetiva media de un códec para dichos idiomas. Cuando se trata de la dependencia del participante que presenta un códec o del comportamiento de un códec para un idioma determinado, se recomienda utilizar de voces reales. En ninguno de esos casos se debe añadir ruido ambiental.

7.1 Voces reales

Cuando se utilizan voces reales en la medición objetiva, éstas deben ser producidas, grabadas y ecualizadas en nivel de conformidad con la cláusula 7/P.830.

Se recomienda una participación mínima de dos hombres y dos mujeres para cada condición experimental. Si se va a probar la dependencia de los participantes como factor de entidad propia, se recomienda que participe un mayor número de sujetos a saber:

– 8 hombres;

– 8 mujeres;

– 8 niños.

7.2 Voces artificiales

Cuando en una medición objetiva se utilizan voces artificiales conforme a la Recomendación P.50, se recomienda utilizar tanto voces artificiales masculinas como femeninas. Dichas señales deben atravesar un filtro con características de frecuencia adecuadas para simular las características de frecuencia de emisión de un aparato de microteléfono, y deben ser ecualizadas en nivel de la misma manera que las voces reales (véase la Recomendación P.830).

El UIT-T recomienda la utilización de la característica de frecuencia de emisión del sistema de referencia intermedio (IRS, *intermediate reference system*) modificado definido en el anexo D/P.830.

8 Selección de parámetros experimentales

Para poner de manifiesto el comportamiento de un códec, se deben examinar los efectos de diversos factores de calidad sobre el comportamiento del códec. En la Recomendación P.830 se puede hallar una orientación sobre la evaluación subjetiva de los siguientes factores de calidad:

1) niveles de entrada de la señal vocal a un códec;

2) niveles de escucha en experimentos subjetivos;

3) participantes (incluidos múltiples participantes simultáneos);

4) errores en el canal de transmisión entre un codificador y un decodificador;

5) velocidades binarias si un códec tiene más de un modo de velocidad binaria;

6) transcodificaciones;

7) desadaptación de la velocidad binaria entre un codificador y un decodificador si el códec tiene más de un modo de velocidad binaria;

8) ruido ambiental en el lado de emisión;

9) señales de información de la red como entrada a un códec;

10) música como entrada a un códec.

Dado que la medida de la calidad objetiva descrita en la presente Recomendación supone que:

1) la señal vocal fuente es "limpia" (es decir, sin ningún ruido ambiental añadido en el lado de emisión); y

2) no hay degradaciones en el canal, tales como errores en los bits de transmisión, borrados de tramas (por ejemplo, en las aplicaciones del servicio móvil radioeléctrico), o pérdida de células (por ejemplo, en las redes ATM),

los factores de calidad a los que se aplica esta Recomendación son los niveles de entrada de la señal vocal, los participantes (con exclusión de los múltiples participantes simultáneos), las velocidades binarias y las transcodificaciones.

NOTA 1 – Todavía está en estudio la medición objetiva para factores de calidad distintos de los que específicamente se señalan como aplicables en la presente Recomendación. Por lo tanto, esos factores se deben medir sólo después de haber verificado la precisión de la medida objetiva juntamente con las pruebas subjetivas conformes a la Recomendación P.830.

NOTA 2 – Aunque hay algunas indicaciones de que la medición objetiva puede predecir con precisión la calidad en condiciones de degradación de canal [10] y [11], está aún en estudio la aplicabilidad de la medición.

Además de las condiciones del códec, en la Recomendación P.830 se recomienda la utilización de condiciones de referencia en las pruebas subjetivas. Estas condiciones son necesarias para facilitar la comparación de los resultados de las pruebas subjetivas de distintos laboratorios o del mismo laboratorio en diferentes momentos. Además, cuando los resultados de la prueba objetiva se expresan en valores Q equivalentes, se deben probar las condiciones de referencia que utilizan el aparato de referencia para ruido modulado (MNRU, *modulated noise reference unit*) de banda estrecha indicado en la Recomendación P.810.

NOTA 3 – La inclusión de otros códecs normalizados como el de MIC a 64 kbit/s (Recomendación G.711), el de MICDA a 32 kbit/s (Recomendación G.726), el de LD-CELP a 16 kbit/s (Recomendación G.728) y el de CS-ACELP a 8 kbit/s (Recomendación G.729), así como el MNRU en la medición objetiva de la calidad, puede servir para mostrar el funcionamiento relativo del códec sometido a prueba y los códecs normalizados.

En la Recomendación P.830 figuran explicaciones detalladas de dichos parámetros experimentales.

9 Cálculo objetivo de la calidad

En esta cláusula se describe un método para medir la calidad de la señal vocal codificada de banda telefónica (300-3400 Hz), utilizando la medición perceptual de calidad vocal (PSQM). La finalidad de la PSQM es imitar la percepción del sonido experimentada por sujetos en situaciones de la vida real [6]. La PSQM simula experimentos en los cuales los sujetos juzgan la calidad de los códecs vocales comparando una señal codificada a una señal fuente (figura 2). Aunque este principio básico de comparación resulta especialmente adecuado para las pruebas con determinación de índices por categorías de degradación (DCR, *degradation category rating*), se pueden simular experimentos con índices por categorías absolutas (ACR, *absolute category rating*) como se muestra en las pruebas de validación [12]. En la medida en que la PSQM es una fiel representación de los procesos de percepción y juicio humanos, las diferencias inaudibles entre las señales de entrada y de salida recibirán la misma nota PSQM. En particular, si ambas señales son idénticas, la PSQM hará una predicción de calidad perfecta prescindiendo de la calidad de la señal de entrada.

En el marco de la PSQM, las señales físicas que constituyen la señal fuente y la señal codificada se corresponden con representaciones psicofísicas que se adaptan en la mayor medida posible a las representaciones internas de las señales vocales (las representaciones registradas en nuestra mente). Estas representaciones internas utilizan equivalentes psicofísicos de frecuencia (velocidades en la banda crítica) e intensidad (sonio comprimido). Se modela el enmascaramiento de una manera

sencilla: sólo se tiene en cuenta el enmascaramiento cuando coinciden dos componentes tiempo-frecuencia en ambos dominios de tiempo y frecuencia.

Dentro del método PSQM, la calidad de la señal codificada se juzga sobre la base de las diferencias en la representación interna. Esta diferencia se utiliza para calcular la perturbación debida al ruido como función del tiempo y la frecuencia. En la PSQM, la perturbación media debida al ruido está directamente relacionada con la calidad de la señal vocal codificada.

La transformación del dominio físico (externo) al dominio psicofísico (interno) tiene lugar mediante tres operaciones:

– correspondencia tiempo-frecuencia;

– transposición de la frecuencia;

– transposición de la intensidad (compresión).

Además del modelado perceptual, el método PSQM utiliza también el modelado cognitivo [7] a fin de obtener correlaciones altas entre las mediciones subjetivas y objetivas.

En la figura 3 se observa un diagrama de bloques del algoritmo PSQM.

En los cuadros 2, 3 y 4 se resumen todos los parámetros y variables de esta cláusula.

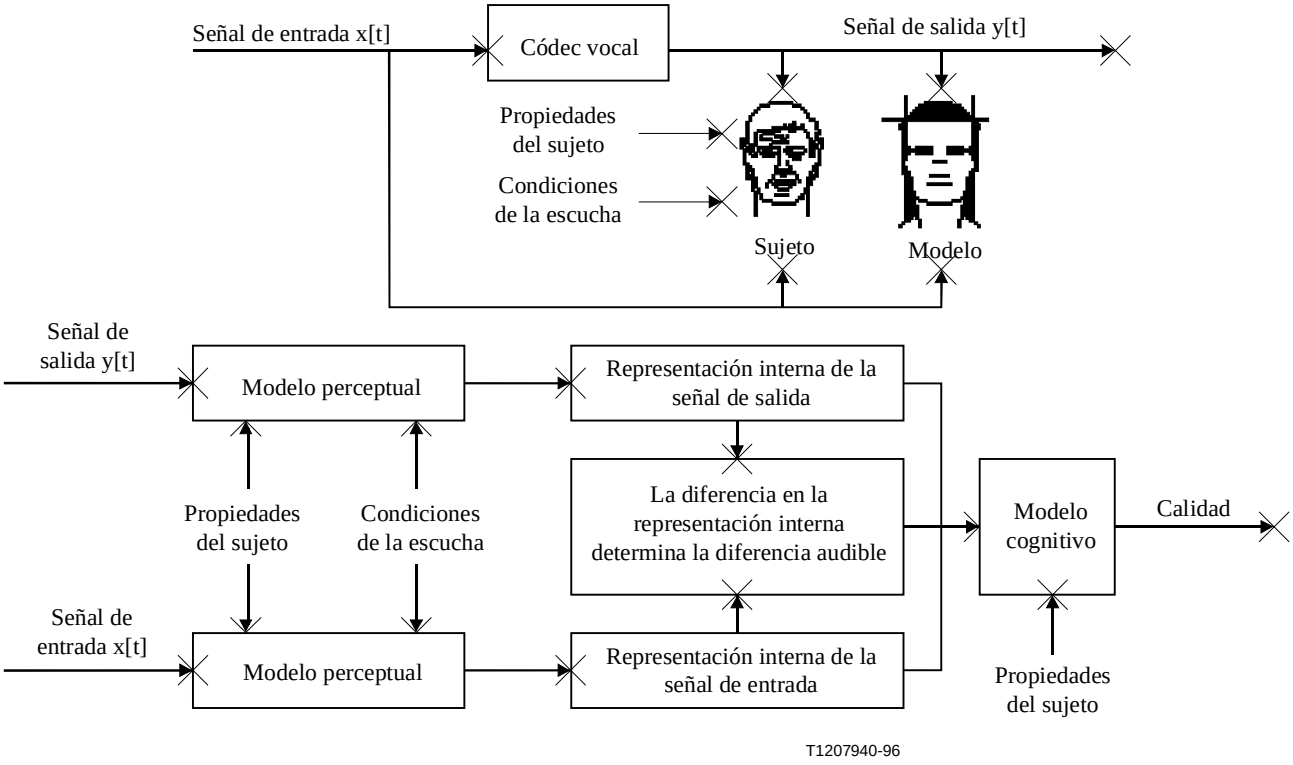

Figura 2/P.861 – Visión general de la filosofía básica aplicada al desarrollo de la PSQM – Se utiliza un modelo informatizado del sujeto, que consta de un modelo perceptual y un modelo cognitivo, para comparar la señal de salida del códec vocal con la señal de entrada

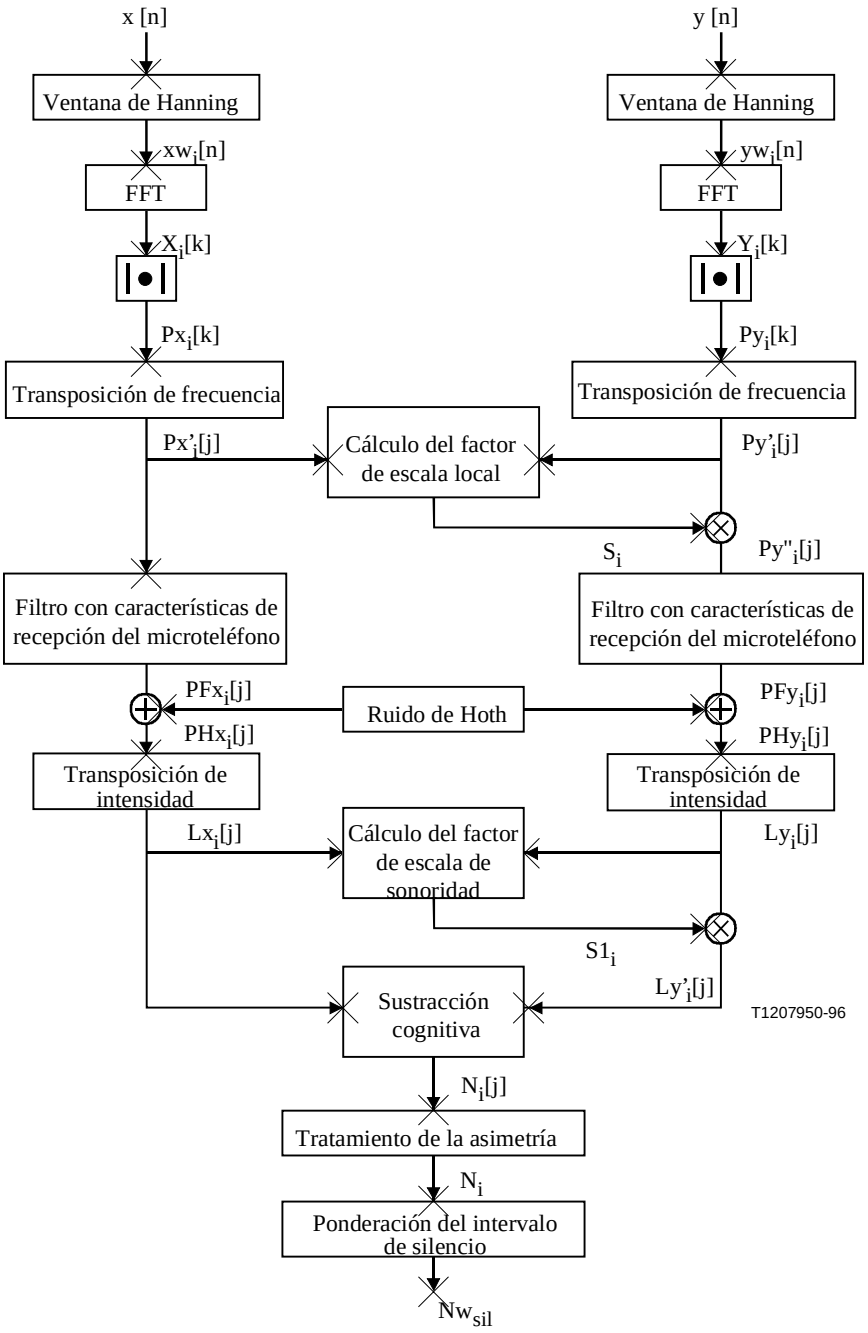

Figura 3/P.861 – Diagrama del algoritmo PSQM

Cuadro 2/P.861 – Lista de parámetros en la PSQM

Nombre	Descripción	Valor
Nb	Número de bandas en el dominio de banda crítica (Bark)	(Véase el cuadro 4)
Nf	Número de muestras en trama de tiempo	512 para frecuencia de muestreo de 16 kHz 256 para frecuencia de muestreo de 8 kHz
F[j]	Microteléfono que recibe características de frecuencia	IRS de la Recomendación P.830 (el cuadro 4 contiene la función de transferencia de potencia del IRS)
H[j]	Características de Hoth	(El cuadro 4 contiene la potencia añadida de la característica de Hoth)
$P_0[j]$	Umbral absoluto de escucha	(El cuadro 4 contiene la representación de potencia equivalente de $P_0[j]$)
$\Delta f[j]$	Anchura de banda de la banda j en Hertz	(Véase el cuadro 4)
Δz	Anchura de banda de cada subbanda en el dominio de banda crítica	0,312
γ	Exponente de la función de compresión	0,001
W_{sil}	Factor de ponderación en tramas de silencio	0,2 (provisional)
W_{sp}	Factor de ponderación en tramas vocales activas	$W_{sp} = (1 - W_{sil})/W_{sil} = 4{,}0$ (provisional)

Cuadro 3/P.861 – Variables en la PSQM

Nombre	Descripción
m	Índice en el dominio del tiempo
n	Índice en el dominio del tiempo en una trama (n: 1, 2, 3, ..., Nf)
i	Índice para tramas
j	Índice en el dominio de la transposición de frecuencias (dominio de banda crítica) (j: 1, 2, 3, ..., Nb)
k	Índice en el dominio de la frecuencia (Hz) (k: 1, 2, 3, ..., Nf/2)
x[m]	Versión con alineación de tiempo y calibración global de la señal vocal fuente muestreada
y[m]	Versión con alineación de tiempo, factor de escala global y calibración global de la señal vocal codificada muestreada
S_{global}	Factor de escala en la escala global
S_p	Factor de calibración de la potencia de tono
S_1	Factor de calibración de la sonoridad de tono
$x_i[n]$	x[m] en trama I
$y_i[n]$	y[m] en trama I
$xw_i[n]$	Versión en ventana de $x_i[n]$
$yw_i[n]$	Versión en ventana de $y_i[n]$
$X_i[k]$	FFT de $xw_i[n]$

Cuadro 3/P.861 – Variables en la PSQM *(fin)*

Nombre	Descripción
$Y_i[k]$	FFT de $yw_i[n]$
$Px_i[k]$	SPD de $xw_i[n]$
$Py_i[k]$	SPD de $yw_i[n]$
$I_f[j]$	Índice FFT del primer valor de k de $Px_i[k]$ y $Py_i[k]$ en la banda j
$I_1[j]$	Índice FFT del último valor de k de $Px_i[k]$ y $Py_i[k]$ en la banda j
$Px_i'[j]$	Densidad de potencia de tono muestreada de $xw_i[n]$
$Py_i'[j]$	Densidad de potencia del tono muestreada de $yw_i[n]$
Px_i'	Potencia de la señal vocal fuente en la trama i
Py_i'	Potencia de la señal vocal codificada en la trama i
$Py_i''[j]$	Versión a escala local de $Py_i'[j]$
$PFx_i[j]$	Versión filtrada en banda telefónica de $Px_i'[j]$
$PFy_i[j]$	Versión filtrada en banda telefónica de $Py_i''[j]$
$PHx_i[j]$	$PFx_i[j]$ más ruido de Hoth como ruido ambiental (recepción)
$PHy_i[j]$	$PFy_i[j]$ más ruido de Hoth como ruido ambiental (recepción)
S_i	Factor de escala en la escala local en la trama i
S_{av}	Promedio (media aritmética) de S_i
$Lx_i[j]$	Densidad de sonoridad comprimida muestreada de la señal vocal fuente en la trama i y banda j
$Ly_i[j]$	Densidad de sonoridad comprimida muestreada de la señal vocal codificada en trama i y banda j
Lx_i	Sonoridad comprimida momentánea de la señal vocal fuente en la trama i
Ly_i	Sonoridad comprimida momentánea de la señal vocal codificada en la trama i
Sl_i	Factor de escala en la escala de sonoridad en la trama i
$Ly_i'[j]$	Versión a escala de sonoridad de $Ly_i[j]$
$N_i[j]$	Densidad de la perturbación debida al ruido muestreada en trama i y banda j
$C_i[j]$	Factor de efecto de asimetría en trama i y banda j
N_i	Perturbación debida al ruido en la trama i
N_{wsil}	Promedio de la perturbación debida al ruido con ponderación en tramas de silencio
M_{sp}	Número de tramas vocales activas
M_{sil}	Número de tramas de silencio
N_{spav}	Promedio de N_i en tramas vocales activas
N_{silav}	Promedio de N_i en tramas de silencio

Cuadro 4/P.861 – Asignaciones de frecuencias de bandas críticas y características de filtro (basadas en una velocidad de muestreo de 16 kHz)

Número de banda, j	Frecuencia más alta [Hz]	Primer Bin FFT en banda j, I_f	Último Bin FFT en banda j, I_l	Característica de recepción, F	Umbral de escucha, P_0	Ruido de Hoth, H
0	15,6	0	0	Descartados en el tratamiento		
1	46,9	1	1	2,45E–06	3,89E+07	1,72E+04
2	78,1	2	2	9,24E–06	1,12E+06	1,72E+04
3	109,4	3	3	3,56E–05	1,26E+05	1,72E+04
4	140,6	4	4	2,59E–04	1,86E+04	1,22E+04
5	171,9	5	5	1,18E–03	6,17E+03	8,49E+03
6	203,1	6	6	7,48E–03	2,29E+03	6,31E+03
7	234,4	7	7	3,19E–02	9,33E+02	4,91E+03
8	265,6	8	8	7,31E–02	4,37E+02	3,95E+03
9	296,9	9	9	1,37E–01	2,29E+02	3,26E+03
10	328,1	10	10	2,09E–01	1,29E+02	2,74E+03
11	359,4	11	11	2,93E–01	7,76E+01	2,35E+03
12	390,6	12	12	4,25E–01	4,27E+01	2,04E+03
13	421,9	13	13	5,23E–01	3,02E+01	1,79E+03
14	453,1	14	14	5,98E–01	2,19E+01	1,59E+03
15	484,8	15	15	6,51E–01	1,66E+01	1,44E+03
16	519,2	16	16	6,94E–01	1,32E+01	1,39E+03
17	553,6	17	17	7,31E–01	1,07E+01	1,25E+03
18	590,8	18	18	7,66E–01	8,91E+00	1,22E+03
19	631,2	19	20	7,98E–01	7,59E+00	1,19E+03
20	672,9	21	21	8,37E–01	6,31E+00	1,10E+03
21	716,6	22	22	8,63E–01	5,62E+00	1,04E+03
22	760,4	23	24	8,88E–01	5,13E+00	9,45E+02
23	804,6	25	25	9,12E–01	4,68E+00	8,69E+02
24	851,4	26	27	9,35E–01	4,37E+00	8,41E+02
25	898,3	28	28	9,56E–01	4,17E+00	7,68E+02
26	947,0	29	30	9,71E–01	4,07E+00	7,33E+02
27	997,0	31	31	9,80E–01	3,98E+00	6,90E+02
28	1051,0	32	33	9,87E–01	3,98E+00	6,87E+02
29	1108,0	34	35	9,90E–01	3,98E+00	6,57E+02
30	1168,0	36	37	9,91E–01	3,98E+00	6,49E+02
31	1231,0	38	39	9,93E–01	3,98E+00	6,17E+02
32	1297,0	40	41	9,95E–01	4,07E+00	5,95E+02

Cuadro 4/P.861 – Asignaciones de frecuencias de bandas críticas y características de filtro (basadas en una velocidad de muestreo de 16 kHz) *(fin)*

Número de banda, j	Frecuencia más alta [Hz]	Primer Bin FFT en banda j, I_f	Último Bin FFT en banda j, I_l	Característica de recepción, F	Umbral de escucha, P_0	Ruido de Hoth, H
33	1366,0	42	43	1,00E+00	4,27E+00	5,68E+02
34	1437,0	44	45	1,01E+00	4,47E+00	5,37E+02
35	1509,0	46	48	1,02E+00	4,68E+00	5,04E+02
36	1582,0	49	50	1,04E+00	5,01E+00	4,80E+02
37	1658,0	51	53	1,06E+00	5,37E+00	4,51E+02
38	1736,0	54	55	1,07E+00	5,62E+00	4,37E+02
39	1817,0	56	58	1,09E+00	5,89E+00	4,20E+02
40	1902,0	59	60	1,10E+00	6,31E+00	4,05E+02
41	1991,0	61	63	1,11E+00	6,61E+00	3,97E+02
42	2084,0	64	66	1,12E+00	6,92E+00	3,86E+02
43	2184,0	67	69	1,12E+00	7,24E+00	3,82E+02
44	2289,0	70	73	1,12E+00	7,59E+00	3,74E+02
45	2401,0	74	76	1,11E+00	7,76E+00	3,67E+02
46	2520,0	77	80	1,10E+00	7,94E+00	3,63E+02
47	2647,0	81	84	1,08E+00	7,94E+00	3,56E+02
48	2781,0	85	88	1,01E+00	7,94E+00	3,46E+02
49	2922,0	89	93	8,62E–01	7,94E+00	3,37E+02
50	3069,0	94	98	6,86E–01	8,13E+00	3,25E+02
51	3225,0	99	103	5,16E–01	8,13E+00	3,16E+02
52	3392,0	104	108	3,12E–01	8,32E+00	2,92E+02
53	3572,0	109	114	1,55E–01	8,32E+00	2,69E+02
54	3765,0	115	120	3,02E–02	8,32E+00	2,47E+02
55	3971,0	121	127	2,03E–03	8,32E+00	2,25E+02
56	4193,0	128	134	1,52E–04	8,32E+00	2,06E+02

NOTA 1 – El umbral absoluto, P_0, utiliza la calibración del nivel de presión sonora (SPL, *sound pressure level*) de 0 dB = 1,0.

NOTA 2 – La primera frecuencia más alta (15,6 Hz) es equivalente a 0,156 de una banda crítica. La anchura de banda Δz es 0,312 de una banda crítica.

9.1 Inicializaciones globales

Antes de iniciar el cálculo de la perturbación debida al ruido, que es el resultado del algoritmo PSQM, se deben efectuar las siguientes inicializaciones globales para cada par de señales vocales fuente y codificadas:

– alineación de tiempo;

– aplicación de escala global para la compensación de la ganancia del sistema;

– calibración global para establecer la sonoridad de la voz.

Dado que los códecs vocales de banda telefónica suelen adoptar una frecuencia de muestreo de entrada de 8 kHz, esta Recomendación considera que tanto la señal vocal fuente como la señal codificada tienen una frecuencia de muestreo de 8 kHz o 16 kHz (es decir, frecuencia de muestreo doble).

9.1.1 Alineación de tiempo

La primera inicialización global que debe realizarse es la alineación de tiempo de la señal fuente x[m] y la señal codificada y[m]. Si las señales no están alineadas adecuadamente, no se puede aplicar la PSQM.

Si no se conoce teóricamente el retardo de tiempo en la señal codificada con respecto a la señal fuente, se puede utilizar como estimación el retardo de tiempo que proporciona la máxima correlación cruzada entre señales fuente y señales codificadas. Para señales que presentan distorsión por retardo de grupo, el retardo que conduce al mínimo valor PSQM es el correcto.

En el transcurso del procedimiento, se descartan los ceros de la izquierda y de la derecha en el registro de señales vocales y se calcula el punto inicial y el punto final mediante la detección de la actividad vocal utilizando únicamente la señal fuente. Los algoritmos para la determinación de la primera y última muestra de señal activa son los siguientes.

Cuando se determina el comienzo de la señal vocal activa en un fichero, la primera muestra que ha de declararse activa es aquella en la que la magnitud (es decir, valor absoluto) de esa muestra, más las magnitudes de las cuatro muestras precedentes totalizan 200 o más. (Cuando se trate de probar las últimas cuatro muestras para determinar el final de la actividad vocal, se considera que las muestras que preceden a la primera muestra tienen un valor cero.)

Cuando se determina el final de la señal vocal activa en un fichero, la última muestra que ha de declararse activa es aquella en la que la magnitud (es decir, valor absoluto) de esa muestra, más las magnitudes de las cuatro muestras siguientes totalizan 200 o más. (Cuando se trate de probar las últimas cuatro muestras para determinar el final de la actividad vocal, se considera que las muestras que preceden a la primera muestra tienen un valor cero.)

9.1.2 Aplicación de escala global

Después del proceso de alineación de tiempo, se aplica un factor de escala a la señal codificada y[m] para compensar la ganancia total del sistema. El factor de escala S_{global} está definido por:

$$S_{global} = \sqrt{\frac{\sum_{\text{punto inicial}}^{\text{punto final}} x^2[m]}{\sum_{\text{punto inicial}}^{\text{punto final}} y^2[m]}}$$

La señal codificada $y[m]$ se multiplica por S_{global}.

9.1.3 Calibración global

Para asegurar la precisión óptima de la medida objetiva, es necesario establecer una calibración entre el nivel de escucha y la sonoridad comprimida. Los valores señalados en el cuadro 4 se basan en la hipótesis de que una SPL de 0 dB es equivalente a un valor máximo de 1,0 en el dominio de potencia de tono, como se calcula en 9.3.1 [es decir, $\max_j(Px_i'[j]) = 1{,}0$ para una trama dada]. Se supone también que se utiliza el nivel de escucha óptimo de 78 dB de SPL junto con registros de señales vocales que tienen un nivel vocal activo de −26 dBov, como se indica en la Recomendación P.830.

Las calibraciones se realizan con una onda sinusoidal de 1 kHz a un nivel SPL de 40 dB (es decir, −64 dBov). Conviene efectuarla con una onda sinusoidal real (es decir, no entera) para evitar perturbaciones de cuantificación en la función de calibración. Un nivel SPL de 40 dB corresponde a una amplitud cero a cresta de 29,54.

La primera calibración sirve para graduar en escala el valor máximo de la representación de la potencia de tono del tono de calibración a 10 000 [es decir, si el $\max_j(Px_i'[j]) = 1{,}0$ para 0 dB SPL, el $\max_j(Px_i'[j]) = 10\,000$ para 40 dB SPL]. Este factor de calibración, S_p, se calcula por:

$$S_p = \frac{10\,000}{\max_j\left(Px_i'[j]\right)}$$

donde $Px_i'[j]$ (véase 9.3.1) se calcula para el tono de calibración. Para una implementación de la PSQM en la que se aplica un factor de escala n a la FFT, como en la rutina comercialmente disponible "four1" de *Numerical Recipes in C* [13],

$$S_p = 6{,}4661\,\mathrm{e}^{-06}$$

La segunda calibración fija la sonoridad comprimida del tono de calibración, tal como se calcula en 9.4, en 1,0 sonio comprimido. El factor de calibración viene dado por:

$$S_l = \frac{1}{Lx_i}$$

donde Lx_i se calcula para el tono de calibración. Si la primera calibración se realiza correctamente, $S_l = 240{,}05$.

NOTA 1 – El tono de calibración no debe ser filtrado a través de la característica de recepción, F, ni se debe añadir ruido de Hoth al tono de calibración antes del cálculo de Lx_i y S_l. La única finalidad de esta excepción es la calibración.

Si el nivel vocal activo en el registro digital no es −26 dBov, o si el nivel de escucha no es un SPL de 78 dB, será necesario graduar en escala los datos de la señal de entrada como corresponda.

NOTA 2 – En un registro digital de 16 bits, 0 dBov se representa por un nivel en DC de 32 767. Por lo tanto, una sinusoide con una amplitud cero a cresta de 32 767 tendrá un nivel medio cuadrático de −3,01 dBov. Con los supuestos adoptados en esta subcláusula, correspondería aproximadamente a un SPL de 101 dB.

9.2 Correspondencia tiempo-frecuencia

La correspondencia entre el dominio tiempo y el dominio tiempo-frecuencia se implementa mediante una transformada de Fourier a corto plazo con una ventana de Hanning, que da como resultado una representación tiempo-frecuencia con resolución constante en los dominios del tiempo y de la frecuencia.

9.2.1 Disposición en ventana

La señal de origen $x_i[n]$ y la señal codificada $y_i[n]$ en la trama *i* se disponen utilizando una ventana de Hanning (sen^2):

$$xw_i[n] = w[n].\ x_i[n]$$
$$yw_i[n] = w[n].y_i[n]$$

siendo $w[n]$ la función ventana.

La función ventana puede calcularse como sigue:

$$w[n] = 0{,}5\left(1-\cos\left(\frac{2\pi \mathrm{i}}{Nf}\right)\right) para\ 0 \le i \le Nf-1$$

Todos los cálculos de la cláusula 9 se definen trama por trama. Debe utilizarse una longitud de trama de 256 muestras para un muestreo a 8 kHz, y 512 muestras para un muestreo a 16 kHz, que corresponde aproximadamente a la longitud de ventana del oído, y las tramas adyacentes deben estar superpuestas en un 50%.

9.2.2 Densidad de potencia espectral (SPD, *spectral power density)* muestreada

Las densidades de potencia espectrales (SPD) de $xw_i[n]$ e $yw_i[n]$, representadas por $Px_i[k]$ y $Py_i[k]$, se calculan utilizando las transformadas rápidas de Fourier (FFT, *fast Fourier transforms*):

$$xw_i[n] \Rightarrow FFT \Rightarrow X_i[k]$$
$$yw_i[n] \Rightarrow FFT \Rightarrow Y_i[k]$$
$$Px_i[k] = (\mathrm{Re}\, X_i[k])^2 + (\mathrm{Im}\, X_i[k])^2$$
$$Py_i[k] = (\mathrm{Re}\, Y_i[k])^2 + (\mathrm{Im}\, Y_i[k])^2$$

9.3 Transposición y filtrado de frecuencia

En esta subcláusula se describe en primer lugar la transposición de la escala de Hertz a la escala de banda crítica, lo que conduce a una representación de la densidad de potencia de tono muestreada dentro de cada trama. La densidad de potencia de tono muestreada de la señal codificada se gradúa en escala dentro de cada trama, tras lo cual se filtran en banda telefónica las señales fuente y las señales codificadas y se añade ruido de Hoth para simular el entorno de escucha. Finalmente, se filtran las señales con la función de transferencia de la característica de oído externo a la del oído interno.

9.3.1 Densidad de potencia de tono muestreada

El índice de frecuencia *k* en Hz es transformado en un índice de tono *j* en el dominio de banda crítica mediante una transposición de escala de frecuencia. En primer lugar, la escala de banda crítica se divide en bandas de intervalos iguales, y para cada banda se calcula un valor de densidad de potencia de tono (una muestra) a partir de las muestras (generalmente más de una) de la densidad de potencia espectral en la banda correspondiente de la escala de Hertz. Las densidades de potencia de tono muestreada $Px_i'[j]$ y $Py_i'[j]$ para la banda *j* en la trama *i* se expresan por:

$$Px_i'[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_l[j] - I_f[j] + 1} \cdot \sum_{I_f[j]}^{I_l[j]} Px_i[k]$$

$$Py_i'[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_l[j] - I_f[j] + 1} \cdot \sum_{I_f[j]}^{I_l[j]} Py_i[k]$$

donde $I_f[j]$ es el índice de la primera e $I_l[j]$ es el índice de la última muestra en la escala de Hertz para la banda j, siendo Δf_j la anchura de banda en la banda j en Hertz, Δz la anchura de banda de cada subbanda en el dominio de banda crítica, y S_p el factor de calibración de potencia de tono indicado en 9.1.3.

9.3.2 Aplicación de escala local

Debe aplicarse un factor de escala a las señales fuente y las señales codificadas dentro de cada trama, a modo de compensación de las variaciones de ganancia lentas. Sólo se tienen en cuenta los componentes de la relación tiempo-frecuencia audibles (por encima del umbral absoluto de audibilidad para cada banda $P_0[j]$ definido en el cuadro 4). Las potencias totales de las señales fuente y las señales codificadas en una trama i, Px_i' y Py_i' se calculan utilizando la representación de frecuencia transpuesta:

$$Px_i' = \sum_{j=1}^{Nb} Px_i'[j]$$

$$Py_i' = \sum_{j=1}^{Nb} Py_i'[j]$$

En la que Nb es el número total de bandas.

Cuando tanto la potencia de la señal fuente Px_i' como la potencia de la señal codificada Py_i' son superiores al SPL de 40 dB, la potencia de la señal codificada para la banda j, $Py_i'[j]$, se multiplica por un factor de escala S_i:

$$Py_i''[j] = S_i \cdot Py_i'[j]$$

donde:

$$S_i = \frac{Px_i'}{Py_i'}$$

Si la potencia de la señal fuente Px_i' o la potencia de la señal codificada Py_i' es inferior al SPL de 40 dB, la potencia de la señal codificada para la banda j, $Py_i'[j]$ se multiplica por un factor de escala S_{av} que es el promedio de todos los factores S_i calculados anteriormente.

9.3.3 Filtrado de la banda telefónica

$Px_i'[j]$ y $Py_i''[j]$ se deben filtrar utilizando las características de recepción adecuadas para un aparato de microteléfono:

$$PFx_i[j] = F[j] \cdot Px_i'[j]$$
$$PFy_i[j] = F[j] \cdot Py_i''[j]$$

donde $F[j]$ es la respuesta de frecuencia en la banda j de las características de recepción de un microteléfono. El UIT-T recomienda la utilización de las características de recepción del IRS modificadas definidas en el anexo D/P.830 como características de frecuencia de recepción de un aparato de microteléfono. En el cuadro 4 figuran los valores de $F[j]$ para dichas características.

9.3.4 Ruido de Hoth

En la utilización normal de un aparato telefónico, la señal vocal sufre perturbaciones ocasionadas por ruidos próximos al entorno de recepción. Dentro de la PSQM, se modela este efecto añadiendo ruido de Hoth a las señales fuente y a las señales codificadas. El ruido de Hoth [8] se suma a la densidad de potencia de tono muestreada para cada valor de *j*, utilizando la función de densidad de potencia espectral indicada en la Recomendación P.800:

$$PHx_i[j] = H[j].PFx_i[j]$$
$$PHy_i[j] = H[j].PFy_i[j]$$

en la que *H[j]* es la potencia del ruido de Hoth en la banda *j* indicada en el cuadro 4.

NOTA – Todas las validaciones del método PSQM efectuadas en el UIT-T utilizaron ruido de Hoth a un nivel de 45 dBA.

9.4 Transposición de intensidad

Después de calcular las densidades de potencia de tono muestreada que tienen en cuenta el filtrado de la banda telefónica y el ruido de Hoth, se efectúa la transposición de la escala de intensidad a una escala de sonoridad que da lugar a una función de densidad de sonoridad comprimida muestreada.

A partir de las densidades de potencia de tono $PHx_i[j]$ y $PHy_i[j]$, se calculan las densidades de sonoridad comprimida de la muestra $Lx_i[j]$ y $Ly_i[j]$ utilizando una función de compresión dada por Zwicker [9]:

$$Lx_i[j] = S_l.\left(\frac{P_0[j]}{0{,}5}\right)^{\gamma}.\left[\left(0{,}5+0{,}5.\frac{PHx_i[j]}{P_0[j]}\right)^{\gamma}-1\right]$$

$$Ly_i[j] = S_l.\left(\frac{P_0[j]}{0{,}5}\right)^{\gamma}.\left[\left(0{,}5+0{,}5.\frac{PHy_i[j]}{P_0[j]}\right)^{\gamma}-1\right]$$

en la que $P_0[j]$ es el umbral interno indicado en el cuadro 4, y S_i el factor de calibración de sonoridad de tono como figura en 9.1.3. Los valores negativos de $Lx_i[j]$ y $Ly_i[j]$ se fijan en cero.

Se ha hallado un valor óptimo de γ igual a 0,001 utilizando bases de datos obtenidas de distintos experimentos de evaluación de calidad vocal.

El total de las sonoridades comprimidas momentáneas Lx_i y Ly_i (en sonio comprimido) se calculan mediante la sumatoria de las densidades de sonoridad comprimidas muestreadas $Lx_i[j]$ y $Ly_i[j]$:

$$Lx_i = \sum_{j=1}^{Nb} Lx_i[j].\Delta z$$

$$Ly_i = \sum_{j=1}^{Nb} Ly_i[j].\Delta z$$

siendo Δz la anchura de banda en el dominio de banda crítica. Las sonoridades comprimidas momentáneas Lx_i y Ly_i se utilizan en el modelado cognitivo.

9.5 Modelado cognitivo

En el marco de la PSQM, todas las operaciones que no se pueden realizar en una señal fuente sola ni en una señal codificada sola se definen como operaciones cognitivas. En esta subcláusula se examinan cuatro efectos cognitivos:

– escala de sonoridad;

– ruido cognitivo interno;

– tratamiento de la asimetría;

– tratamiento del intervalo de silencio.

9.5.1 Escala de sonoridad

En la PSQM, la densidad de sonoridad comprimida muestreada de la señal codificada se gradúa, en escala dentro de cada trama, con respecto a la sonoridad de la señal fuente:

$$Ly_i'[j] = Sl_i \,.\, Ly_i[j]$$

donde el factor de escala Sl_i se calcula a partir del total de las sonoridades comprimidas momentáneas Lx_i y Ly_i:

$$Sl_i = \frac{Lx_i}{Ly_i}$$

Cuando Lx_i o Ly_i es inferior a 0,02 sonio comprimido, Sl_i se fija en 1.

9.5.2 Densidad de la perturbación debida al ruido muestreada

La densidad de la perturbación debida al ruido muestreada $N_i[j]$ en la banda j, trama i, viene dada por el valor absoluto de la diferencia entre $Lx_i[j]$ y $Ly'_i[j]$:

$$N_i[j] = \left| Ly_i'[j] - Lx_i[j] \right| - 0{,}01$$

donde 0,01 sonio comprimido representa el ruido cognitivo interno. Si a causa de este factor, $N_i[j]$ se hiciera negativo, $N_i[j]$ se fijaría en cero.

9.5.3 Tratamiento de la asimetría

Cuando se introduce un nuevo componente de la relación tiempo-frecuencia en la señal vocal, la calidad subjetiva experimenta una degradación mayor que cuando un componente de igual sonoridad es excluido por el códec vocal. Dicha asimetría es más destacada durante los intervalos de silencio. Puede suprimirse el ruido presente en la señal fuente, con lo cual mejora la calidad. Si no hay ruido durante los intervalos de silencio en la señal fuente, cualquier diferencia entre la señal fuente y la señal codificada conduce a una disminución de calidad.

Además, cuando un componente de tiempo-frecuencia es excluido de la señal fuente (no codificado por el códec), la señal restante sigue siendo un fenómeno auditivo coherente. Si se introduce un nuevo componente tiempo-frecuencia independiente en la señal codificada (una distorsión), la nueva señal formada se puede descomponer en dos partes, la señal original y la distorsión. Esta descomposición del flujo auditivo hace más molesto el ruido.

El efecto de asimetría está cuantificado por $C_i[j]$ y ha sido tenido en cuenta en la perturbación debida al ruido en la trama i, N_i:

$$N_i = \sum_{j=1}^{Nb} N_i[j] \,.\, C_i[j] \,.\, \Delta z$$

donde:

$$C_i[j]=\left(\frac{PHy_i[j]+1}{PHx_i[j]+1}\right)^{0,2}$$

en la que $PHx_i[j]$ y $PHy_i[j]$ son las potencias de las señales fuente y las señales codificadas (tras filtrado IRS y la adición de ruido Hoth), respectivamente, dentro de la trama i y la banda j. Cuando $PHx_i[j]$ y $PHy_i[j]$ no alcanzan los 20 dB por encima del umbral absoluto de audibilidad en la banda j (es decir, $PHx_i[j]$ y $PHy_i[j]$ son inferiores a 100 * $P_0[j]$, $C_i[j]$ se fija en 1. El valor máximo de $C_i[j]$ debe limitarse a 2,0.

9.5.4 Perturbación debida al ruido que incluye el tratamiento del intervalo de silencio

En la PSQM, se tienen en cuenta los intervalos de silencio utilizando un factor de ponderación, W_{sil}, que depende del contexto de los experimentos subjetivos. Las tramas de silencio se definen como aquellas tramas para las cuales la señal fuente tiene una potencia total Px_i' (es decir, $\sum_j Px_i'[j]$) inferior al SPL de 70 dB. Si el factor de calibración global, S_p, ha sido correctamente calculado, el umbral de silencio es $Px_i' = 1{,}0*10^7$. Las tramas con Px_i' inferiores a este valor se consideran tramas de silencio.

Los promedios de sonoridades de ruido, N_{spav} y N_{silav}, pueden ahora calcularse en tramas vocales activas y en tramas de silencio, respectivamente:

$$N_{spav}=\frac{1}{M_{sp}}\sum_{i\ para\ tramas\ vocales\ activas} N_i$$

$$N_{silav}=\frac{1}{M_{sil}}\sum_{i\ para\ tramas\ de\ silencio} N_i$$

donde M_{sp} es el número de tramas vocales activas y M_{sil} es el número de tramas de silencio.

La influencia de los intervalos de silencio depende directamente de la longitud de dichos intervalos. Si la señal vocal fuente no contiene ningún intervalo de silencio, la influencia será nula. Si la señal vocal fuente contiene un cierto porcentaje de tramas de silencio, la influencia será proporcional a dicho porcentaje. Utilizando un conjunto de condiciones límite evidentes, se puede demostrar que la ponderación correcta es:

$$N_{w_{sil}}=\frac{W_{sp}\cdot p_{sp}}{W_{sp}\cdot p_{sp}+p_{sil}}.N_{spav}+\frac{p_{sil}}{W_{sp}\cdot p_{sp}+p_{sil}}.N_{silav}$$

siendo p_{sil} la fracción de tramas de silencio, p_{sp} la fracción de tramas vocales activas ($p_{sil} + p_{sp} = 1{,}0$), W_{sil} el factor de ponderación en los intervalos de silencio, $_{sp}=\frac{1-W_{sil}}{W_{sil}}$, y $N_{w_{sil}}$ la perturbación debida al ruido corregida con un factor de ponderación W_{sil} para el intervalo de silencio.

Esta perturbación debida al ruido $N_{w_{sil}}$, denominada valor PSQM en las subcláusulas siguientes, puede utilizarse para predecir la calidad vocal percibida subjetivamente obtenida por el método de determinación de índices por categorías absolutas (ACR) que utiliza la escala de calidad de escucha.

NOTA 1 – El valor de $N_{w_{sil}}$ debe tener un límite superior de 6,5.

NOTA 2 – Para el material vocal que comprende largos periodos de intervalos de silencio, la ponderación es distinta de la utilizada para un material vocal que únicamente tiene cortos intervalos de silencio. Además, el ruido en la grabación de los materiales fuente influye también en la ponderación del intervalo de silencio. Para el material vocal carente de intervalos de silencio, la ponderación no es aplicable y $N_{w_{sil}}$ pasa a ser igual

a N_{spav}. Se examinaron un cierto número de bases de datos vocales para determinar la ponderación óptima en los intervalos de silencio. Dichas bases de datos constaban de material vocal con alrededor del 50% de intervalos de silencio. La ponderación óptima hallada variaba entre 0,0 y 0,5 [10] [11] y [12]. Sigue aún en estudio la determinación del valor de W_{sil} para señales vocales con intervalos de silencio. Con carácter provisional, se recomienda una ponderación de 0,2 para materiales vocales con alrededor del 50% de intervalos de silencio.

10 Transformación de la escala de calidad objetiva a la escala de calidad subjetiva

El resultado del algoritmo descrito en la cláusula 9, denominado valor PSQM, indica el grado de degradación subjetiva de la calidad debida a la codificación vocal. Por lo tanto, cuando no es necesaria la evaluación subjetiva de la calidad en una escala específica, por ejemplo, para optimizar los parámetros de un códec o simplemente para comparar el comportamiento de códecs, el propio valor PSQM es bastante útil. No obstante, para la evaluación subjetiva de la calidad en escalas de calidad como las notas medias de opinión (MOS, *mean opinion scores*) y los valores Q equivalentes, el valor PSQM se transforma como se indica a continuación.

10.1 Notas medias de opinión

En la evaluación subjetiva de la calidad de funcionamiento de los códecs, se utiliza generalmente el método ACR que utiliza la escala de calidad de escucha especificada en la Recomendación P.800, la cual proporciona la calidad objetiva expresada de MOS. Dado que la relación entre los valores MOS y PSQM no es necesariamente la misma para los diferentes idiomas, ni siquiera para distintas pruebas subjetivas con un mismo idioma, es difícil determinar una función única que transforme el valor PSQM al valor MOS estimado. En la práctica, por consiguiente, es necesario obtener con anticipación dichas funciones de transformación para cada idioma y para cada prueba subjetiva.

NOTA – El valor absoluto de la MOS depende del contexto del experimento subjetivo. La MOS estimada que se obtiene por una función de transformación predeterminada predice la calidad subjetiva en el experimento subjetivo cuyo contexto equivale a los utilizados al deducir la función de transformación.

Cuando se presenten los resultados en el dominio de la MOS estimada, se deberá comunicar la función de transformación del valor PSQM al valor MOS.

10.2 Valores Q equivalentes

Es difícil comparar las MOS obtenidas en diferentes experimentos subjetivos dado que el juicio subjetivo está influido por los entornos experimentales, por ejemplo, la gama de calidad vocal en el experimento. Por consiguiente, a veces se utiliza el valor Q equivalente como escala de calidad subjetiva. El valor Q equivalente está determinado como el valor Q del MNRU definido en la Recomendación P.810 para el cual la MOS es equivalente a la de la señal vocal codificada.

En la medición objetiva, el valor Q equivalente se puede calcular directamente a partir de los valores PSQM para señales vocales codificadas y de las condiciones del MNRU, sin transformar el valor PSQM a la MOS (véase la figura 4). Cuando se presentan los resultados en el dominio Q equivalente estimado, se deben comunicar las características del valor Q en relación con las del valor PSQM ilustradas en la figura 4.

NOTA – El valor Q equivalente es relativamente no fiable en las regiones en las que dicho valor Q es alto y bajo, dado que la curva de Q en función de PSQM se hace casi plana en dichas regiones. Por consiguiente, es necesario tener cuidado cuando se trabaja en el dominio Q con señales vocales de muy alta y muy baja calidad.

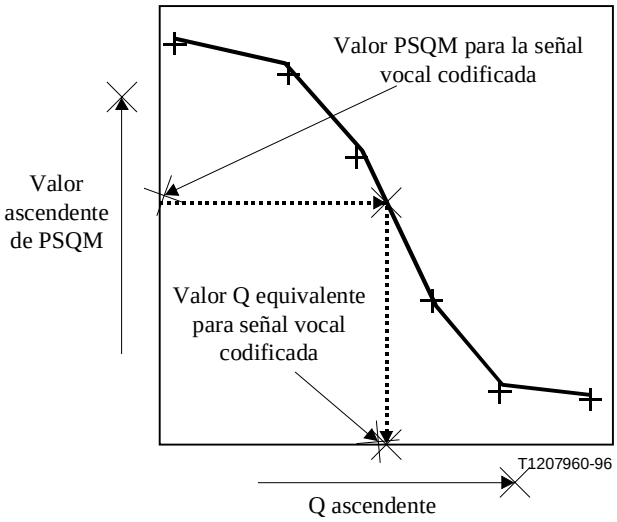

Figura 4/P.861 – Determinación del valor Q equivalente de la señal vocal codificada

11 Análisis de resultados

El análisis de los resultados de la medición objetiva debe realizarse con base en el valor PSQM en la MOS estimada o en el valor Q equivalente estimado.

Para cada condición de prueba, deben calcularse separadamente y comunicarse las notas medias de los participantes masculinos, femeninos y su promedio.

No se recomienda el cálculo de desviaciones típicas separadas para cada condición de prueba. Los límites de confidencia se deben evaluar teniendo en cuenta la variación de la calidad objetiva de unos a otros participantes y frases tipo y las pruebas de significación realizadas por técnicas convencionales de análisis de varianza.

NOTA – El análisis estadístico aquí descrito es diferente del de la evaluación subjetiva en el cual la media de la calidad subjetiva se evalúa estadísticamente teniendo en cuenta las variaciones de unos a otros sujetos así como de participantes y frases tipo. Dado que la PSQM no puede calcular las distribuciones de apreciaciones subjetivas sino únicamente la media de las mismas, es imposible realizar el análisis sobre los sujetos. El cálculo de la distribución de las apreciaciones subjetivas sigue siendo aún objeto de estudio. Por lo tanto, cuando sea necesario efectuar el análisis de los sujetos, se deberán realizar experimentos subjetivos conformes a la Recomendación P.830.

Bibliografía

[1] NTT: Transmission performance objective evaluation model for fundamental factors, *CCITT Contribution COM XII-174*, noviembre de 1983.

[2] LALOU (J.): The information index: an objective measure of speech transmission performance, *Annales des Telecommunications*, volume 45, N.º 1-2, págs. 47-65 CNET/Francia, 1990.

[3] BNR: Evaluation of non-linear distortion via the coherence function, *CCITT Contribution COM XII-60*, abril de 1982.

[4] KUBICHEK (R.F.), QUINCY (E.A.), KISER (K.L.): Speech Quality Assessment Using Expert Pattern Recognition Techniques, *Proceedings of the IEEE Pacific Rim Conference on Computers, Communication, and Signal Processing*, junio de 1989.

[5] Royal PTT, Netherlands: Measuring the quality of audio devices, *CCITT Contribution COM XII-114*, Ginebra, diciembre de 1991.

[6] BEERENDS (J.G.), STEMERDINK (J.A.): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation, *J. Audio Eng. Soc.*, vol. 42, N.º 3, págs. 115-123, marzo de 1994.

[7] BEERENDS (J.G.): Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, *Speech Quality Assessment*, Workshop papers, Bochum, págs. 1-9, noviembre de 1994.

[8] Suplemento N.º 13 del CCITT a las Recomendaciones de la serie P, cláusula 2, *Espectros de ruido*, *Libro Azul*, Fascículo 5, UIT, Ginebra, 1988.

[9] ZWICKER (Feldtkeller): *Das Ohr als Nachrichtenempfänger*, *S. Hirzel Verlag*, Stuttgart, 1967.

[10] Royal PTT, The Netherlands: Correlation of a perceptual quality speech measure with the subjective quality of the CCITT LD-CELP (G.728) speech codec, *ITU-T Contribution COM 12-10*, Ginebra, marzo de 1993.

[11] Royal PTT, The Netherlands: Correlation between the PSQM and the subjective results of ITU-T 8 kbit/s 1993 speech codec test, *ITU-T Contribution COM 12-31*, Ginebra, septiembre de 1994.

[12] NTT: Review of validation tests for objective speech quality measures, *ITU-T Contribution COM 12-74*, Ginebra, mayo de 1996.

[13] PRESS (W.H.) *et al.* Numerical Recipes in C, The Art of Scientific Computing – *Cambridge University Press*, Cambridge, Massachusetts, 1988.

APÉNDICE I

Contenido del disquete que acompaña a la Recomendación P.861

I.1 Introducción

A fin de ayudar a los lectores de esta Recomendación a desarrollar su propia implementación de la PSQM, se ha incluido en ella un disquete que contiene vectores de prueba y el detalle de los valores calculados. El presente apéndice describe los ficheros del disquete.

I.2 Directorio\test

Este directorio contiene ficheros destinados a probar una implementación de la precisión de la PSQM. Los ficheros contenidos en este directorio son:

`longs.cod`	`longs.src`	`outlong.txt`
`outshort.txt`	`shorts.cod`	`shorts.src`

El contenido de los ficheros se describe a continuación.

`longs.cod` Fichero de señales vocales codificadas que se utiliza al calibrar la PSQM. Se almacena primero el LSB, que es retardado por 22 muestras procedentes del fichero de señales vocales de origen.

`longs.src` El fichero de señales vocales de origen para crear `longs.cod`. Se almacena primero el LSB.

`outlong.txt` Trama a trama para el vector de prueba largo. Además de las variables de calibración (S_p y S_1), el fichero contiene la siguiente información incluida en el fichero:

factor de escala global calculado (S_{global});

puntos inicial y final de los ficheros de señales vocales de origen y codificadas;

perturbación de ruido trama a trama (N_i);

indicador trama a trama de silencio de trama (1 = silencio, 0 = no silencio).

`outshort.txt` Valores de variable paso a paso para el vector de prueba corto. Requiere los siguientes valores en el programa:

retardo = 0;

punto inicial = 0;

punto final = 511;

S_{global} = 1,0.

En el fichero se proporcionan los valores intermedios siguientes:

secuencia de entrada ($x_i[n]$, $y_i[n]$);

versión en ventana de secuencia de entrada ($xw_i[n]$, $yw_i[n]$);

densidad de potencia de espectro muestreada ($Px_i[k]$, $Py_i[k]$);

factor de escala local (S_i);

densidad de potencia de tono muestreada ($Px'_i[j]$, $Py'_i[j]$);

resultados del filtrado en banda telefónica ($PFx_i[j]$, $PFy_i[j]$);

resultados de la adición de ruido Hoth ($PHx_i[j]$, $PHy_i[j]$);

densidad de sonoridad comprimida muestreada ($Lx_i[j]$, $Ly_i[j]$);

factor de escala de sonoridad local (Sl_i);

densidad de perturbación de ruido muestreado ($N_i[j]$);

factor de efecto de asimetría ($C_i[j]$);

perturbación de ruido (N_i).

`shorts.cod` Fichero de señales vocales codificadas que se utiliza al calibrar la PSQM. Se almacena primero el LSB, que es retardado por 0 muestras procedentes del fichero de señales vocales de origen.

`shorts.src` El fichero de señales vocales de origen para crear `shorts.cod`. Se almacena primero el LSB.

APÉNDICE II

Medición objetiva de la calidad de los códecs vocales de banda telefónica (300-3400 Hz), mediante bloques de normalización de la medición (MNB, *measuring normalizing slocks*)

II.1 Introducción

El UIT-T ha estudiado métodos objetivos para medir la calidad de la señal vocal codificada. En el periodo de estudios 1993-1996, se comparó la eficacia de varias medidas de la calidad objetiva de códecs vocales de banda telefónica (300-3400 Hz). Basándose en los resultados de la prueba de validación, se concluyó que el método denominado medición perceptual de la calidad vocal (PSQM, *perceptual speech quality measure)* era el que mejor estimaba la calidad subjetiva, con independencia de los idiomas utilizados, los hablantes y los códecs. Asimismo, se vio que, en promedio, cuando se utiliza la PSQM el error de estimación es comparable con la fiabilidad estadística de los datos subjetivos. Esto llevó a preparar la Recomendación P.861, basándose en el algoritmo PSQM.

La medida objetiva de la calidad descrita en el cuerpo principal de esta Recomendación es un instrumento muy eficaz para estimar la calidad subjetiva de los códecs vocales. Ahora bien, como se describe en la cláusula alcance de esta Recomendación (véase el cuadro 1), los factores de calidad que pueden evaluarse utilizando dicha Recomendación son limitados.

Para ampliar el alcance de la medición objetiva de la calidad de los códecs vocales, el UIT-T se encuentra preparando una medida objetiva de la calidad que puede aplicarse a la evaluación de señales vocales degradadas por errores en el canal de transmisión, tales como errores en los bits en redes móviles, pérdida de células en las redes ATM, pérdida de paquetes en la telefonía por Internet, etc.

En este apéndice se proporciona una medida objetiva a la calidad cuyo algoritmo se encuentra actualmente disponible en detalle en el UIT-T y se espera que pueda aplicarse a la evaluación de los errores en los bits y las borraduras en las tramas. Se prevé también que el algoritmo consignado en el presente apéndice amplíe el alcance de la medición objetiva de la calidad en lo que concierne a los diferentes tipos de códecs que habrán de utilizarse (cuadro II.1). Hay que señalar que los factores de prueba incluidos en el alcance de esta Recomendación deberían evaluarse basándose en la propia Recomendación P.861.

Aunque en [1], [2] y [3] del presente apéndice puede encontrarse información sobre la eficacia de la medida objetiva de la calidad, se requiere mayor información. En particular, habría que investigar la aplicabilidad de esta medida a otros tipos de degradación de canal, por ejemplo, la pérdida de células/paquetes, en comparación con otras medidas objetivas de la calidad que ha propuesto el UIT-T ([4], [5] y [6]). Por tales motivos, el UIT-T no recomienda adoptar por el momento ningún nuevo algoritmo que amplíe el alcance de esta Recomendación, pero proporciona un ejemplo de las medidas objetivas de la calidad que pueden tener mayor alcance que dicha Recomendación.

El UIT-T seguirá comparando/validando la eficacia de las medidas objetivas de la calidad, incluida la descrita en este apéndice. Esto llevará a preparar una nueva Recomendación aplicable a la evaluación de los errores en el canal de transmisión, para ampliar el ámbito de esta Recomendación vigente.

II.2 Cálculo de la medida objetiva

En esta subcláusula se describe el cálculo de una medida objetiva, basada en bloques de normalización de la medición (MNB, *measuring normalizing blocks*). Los MNB se prepararon en respuesta a la observación de que los oyentes se adaptan y reaccionan de manera distinta a las desviaciones espectrales con diferentes escalas de tiempo y de frecuencia. Así, pues, en lo que atañe

a la aplicación de la estimación de la calidad vocal, la máxima coherencia perceptual en una amplia gama de tipos de distorsión requiere realizar un conjunto de análisis a múltiples escalas de frecuencia y de tiempo. Al medir las desviaciones espectrales, deben suprimirse las desviaciones en una escala, para no volverlas a computar como parte de las desviaciones en otras escalas. Se observó también que, al pasar de escalas grandes a otras más reducidas, lo más probable es que se emulen los patrones de adaptación y reacción de los oyentes a las desviaciones espectrales. Esta observación ha llevado a definir una estructura jerárquica para los MNB.

Cuadro II.1/P.861 – Relación entre las tecnologías de codificación, los factores experimentales y las aplicaciones en esta Recomendación

Factores de prueba	Nota
Niveles de entrada de señal vocal a un códec	1
Niveles de escucha en experimentos subjetivos	3
Factores de dependencia del hablante	1
Múltiples hablantes simultáneos	3
Errores de bits y borraduras de tramas	2
Pérdida de células y paquetes	3
Velocidades binarias, si el códec tiene más de un modo de velocidad binaria	1
Transcodificaciones	1
Desadaptación de velocidades binarias entre codificador y decodificador si el códec tiene más de un modo de velocidad binaria	3
Ruido ambiental en el lado emisión	3
Señales de información de la red como entrada a un códec	3
Música como entrada a un códec	3
Retardo	4
Transposición temporal a corto plazo de la señal audio	3
Transposición temporal a largo plazo de la señal audio	5
Recorte en tiempo de la señal vocal	3
Recorte en amplitud de la señal vocal	3
Tecnologías de codificación	
Forma de onda	1
Celp e híbridos ≥ 4 kbit/s	1
Celp e híbridos < 4 kbit/s	2
VOCODER (codificadores de señales vocales)	2
Otros codificadores	2
Aplicaciones	
Optimización del codificador	1
Evaluación del codificador	1
Selección del codificador	3
Planificación de la red	6
Pruebas de red reales	7
Dispositivos de medición que no ocasionan interrupción del servicio	4

Notas relativas al Cuadro II.1/P.861

NOTA 1 – Si las condiciones de su prueba consisten en uno de los elementos calificados con un 1, debe utilizar el algoritmo descrito en esta Recomendación.

NOTA 2 – Si las condiciones de su prueba consisten en sólo uno de los elementos calificados con un 2, o en una combinación de elementos calificada con 1 ó 2, el algoritmo descrito en el presente apéndice debe proporcionar resultados satisfactorios.

NOTA 3 – Se dispone de información insuficiente sobre la precisión de la medida objetiva con respecto a esta variable.

NOTA 4 – Es sabido que las medidas objetivas proporcionan predicciones imprecisas cuando se utilizan junto con esta variable o, en otros casos, cuando no está previsto utilizarlas con dicha variable.

NOTA 5 – Es sabido que la medida objetiva de esta Recomendación proporciona predicciones imprecisas cuando hay una fluctuación lenta de fase apreciable (más del 10% de la longitud de trama). La aplicabilidad de la medida descrita en este apéndice queda en estudio. Queda en estudio la aplicabilidad de ambas medidas objetivas cuando la fluctuación lenta de fase es pequeña.

NOTA 6 – Con las precauciones necesarias, las medidas objetivas podrían utilizarse a ciertos efectos de planificación de la red. Cabe señalar al lector que existen importantes factores en la planificación de la red a los que no es aplicable la presente Recomendación (véase la sección "Factores de prueba" de este cuadro).

NOTA 7 – Con las precauciones necesarias, las medidas objetivas podrían utilizarse para realizar ciertas pruebas de red reales. Cabe señalar al lector que puede haber factores o tecnologías en una conexión de red real a los cuales no es aplicable la presente Recomendación o este apéndice (véanse las secciones "Factores de prueba" y "Tecnologías de codificación" de este cuadro).

Se consideran aquí dos tipos de bloques de normalización de la medición. E1 primero es el bloque de normalización de la medición en tiempo (TMNB, *time measuring normalizing block*), mientras que el segundo es el bloque de normalización de la medición en frecuencia (FMNB, *frequency measuring normalizing block*). Cada uno de estos bloques toma señales de referencia [R(t,f)] y de prueba [T(t,f)] transformadas perceptualmente como entradas y las devuelve como salidas, junto con un conjunto de mediciones. Estos dos bloques constituyentes se definen, respectivamente, en las figuras II.1 y II.2. El TMNB integra en una determinada escala de frecuencias, a continuación mide las diferencias y normaliza la señal de prueba en múltiples tiempos. Por último, las partes positiva y negativa de las mediciones se integran en el tiempo. En un FMNB ocurre lo contrario. Un FMNB integra en una cierta escala de tiempo, a continuación mide las diferencias y normaliza la señal de prueba en múltiples frecuencias. Por último las partes positiva y negativa de las mediciones se integran en frecuencia. El diseño de ambos tipos de MNB hace que sean equipotentes. Esta importante propiedad se ilustra en la figura II.3 y significa simplemente que un segundo pase a través de un MNB dado no alterará en mayor medida la señal de prueba y que dicho pase dará lugar a un vector de medición compuesto por ceros. La equipotencia de los MNB permite conectarlos en cascada, y no obstante seguir midiendo la desviación en una escala de tiempo o de frecuencia dadas una sola vez.

Para medir las desviaciones espectrales en múltiples escalas de tiempo y frecuencia, se ha formado una estructura jerárquica de los TMNB y FMNB que funciona en escalas decrecientes (figura II.4). Cuando se utiliza para medir la distancia junto con una transformación perceptual (como se describe más abajo), esta estructura parece ser muy eficaz para emular patrones de adaptación y reacción de los oyentes a desviaciones espectrales. Esta estructura arroja resultados en 12 mediciones. Debido a la naturaleza jerárquica de este tipo de estructuras, las mediciones a partir de capas distintas de la superior no son muy significativas por separado, pero se ha visto que una combinación lineal de las mediciones es un buen indicador de la distancia perceptual entre las dos señales. El valor resultante de esta combinación lineal se denomina distancia al auditorio (AD, *auditory distance*):

$$AD = \sum_{i=1}^{N} peso_i \ . m_i \ .$$

La distancia al auditorio es una magnitud positiva. Cuando las señales de referencia y de prueba son similares, AD es pequeña. A medida que las señales de referencia y prueba se separan perceptualmente, AD aumenta. Puede emplearse una función logística o cualquier otro tipo de "función limitadora" para hacer corresponder AD a un intervalo finito. Esto permite correlacionar más adecuadamente AD con juicios de calidad subjetiva o de degradación, que normalmente tienen una gama finita. Los pesos, $peso_i$, se han elegido para maximizar esta correlación.

NOTA – Este apéndice contiene suficiente información para implementar algoritmos en un lenguaje de programación. Pueden validarse nuevas implementaciones utilizando la información disponible en el fichero ftp://ftp.its.bldrdoc.gov/dist/voice/verify.zip.

II.2.1 Especificaciones de entrada-salida

La entrada al algoritmo está constituida por un par de ficheros vocales denominados de referencia y prueba. El fichero de referencia contiene una representación digital de la señal de referencia que suele ser la entrada al dispositivo objeto de prueba (DUT, *device under test*), y se designa por *x* en las ecuaciones. El fichero llamado de prueba contiene una representación digital de la señal de prueba, que suele ser la salida del DUT y se designa por *y* en las ecuaciones. La velocidad de muestreo es de 8000 muestras por segundo y la precisión recomendada es al menos de 16 bits por muestra. Puede adoptarse una precisión menor, si el usuario está dispuesto a aceptar la correspondiente pérdida de sensibilidad. Además, pueden adoptarse una precisión mayor y velocidades de muestreo superiores (por ejemplo, 16 000 muestras por segundo), si se pone cuidado en asegurar que la transformación al dominio de frecuencia produzca los resultados especificados en II.2.4. Además, si se dispone de ficheros con velocidad de muestreo elevada, podrían muestrearse con una velocidad menor utilizando los programas disponibles en la biblioteca de herramientas de soporte lógico del UIT-T que se han publicado en la Recomendación G.191 (1996). Los ficheros de entrada deben contener al menos un segundo de señal vocal de anchura de banda telefónica (los ficheros que contienen sólo pausas en una conversación natural no resultan útiles). La duración de los ficheros empleados para preparar estos algoritmos va de 3 a 9 segundos.

El algoritmo genera un único valor de salida no negativo llamado distancia al auditorio (AD). AD es una estimación de la distancia perceptual entre las señales de referencia y de prueba. Por esa razón, cuando el DUT y la configuración de la prueba son transparentes, las señales de referencia y de prueba serán idénticas y AD nula. A medida que el DUT genere mayor distorsión, las señales de referencia y prueba variarán desde el punto de vista perceptual y AD aumentará.

II.2.2 Retardo de tiempo

Se supone que los dos ficheros tienen la misma longitud y están sincronizados, esto es, que se ha suprimido todo retardo en el DUT y la configuración de la prueba. Si estos retardos se conocen a priori, pueden eliminarse mediante una temporización adecuada durante la adquisición de los datos. Si estos retardos no se conocen a priori, podrán estimarse utilizando la técnica descrita en la sección 7 de ANSI T1 Standard on Multimedia Communications Delay, Synchronization, and Frame Rate Measurement (ANSI Standard T1.801.04-1997), y suprimirse acto seguido editando uno o ambos ficheros.

II.2.3 Preparación de la señal

El contenido de referencia se representa en el vector *x*, y el contenido de la prueba en el vector *y*. Se suprime a continuación el valor medio en cada una de las N1 entradas de ambos vectores:

$$x(i) = x(i) - \frac{1}{N1} \cdot \sum_{j=1}^{N1} x(j), \;\; y(i) = y(i) - \tfrac{1}{N1} \cdot \sum_{j=1}^{N1} y(j), \;\; 1 \le i \le N1.$$

NOTA – En este apéndice se utiliza la notación *Formación (i) = Formación (i) +/–/· /÷ factor de normalización.* Aunque esta notación no es matemáticamente coherente, es indicativa de la forma en que puede implementarse este algoritmo en forma codificada. La formación *(array)* original se modifica (normaliza) en la forma indicada y el valor modificado se utiliza en cálculos posteriores.

Esto permite eliminar cualquier componente de corriente continua (CC) que pueda encontrarse en las señales de prueba y de referencia (la componente CC de una señal es inaudible).

A continuación, se normaliza cada uno de los vectores a un nivel medio cuadrático común:

$$x(i) = x(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} x(j)^2 \right]^{-1/2}, \; y(i) = y(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} y(j)^2 \right]^{-1/2}, 1 \le i \le N1$$

Esta operación elimina aproximadamente cualquier ganancia fija en el DUT o en la configuración de la prueba. Así pues, los valores de AD que arroja este algoritmo no quedarán afectados por una ganancia fija.

II.2.4 Transformación al dominio de la frecuencia

A continuación las señales se transforman al dominio de la frecuencia utilizando la transformada rápida de Fourier (FFT). El tamaño de la trama es de 16 ms (128 muestras por señal vocal muestreada a 8000 Hz) y la superposición de tramas es del 50%. Las muestras más allá de la trama final completa se descartan. Cada trama de muestras se multiplica (muestra por muestra) por la ventana de Hamming de longitud 128:

$$w(i) = 0{,}54 - 0{,}46 \cos\left(\frac{2\pi(i-1)}{127} \right), 1 \le i \le 128.$$

Después de la multiplicación por la ventana de Hamming, cada trama se transforma en un vector del dominio de la frecuencia de 128 puntos utilizando la FFT. Para cada trama, se retienen los valores cuadráticos de las muestras de frecuencia 1 a 65 (de componente CC a Nyquist). Los resultados se almacenan en las matrices X e Y. Estas matrices contienen 65 filas y N2 columnas, donde N2 es el número de tramas extraídas de las muestras originales N1 en *x* e *y*.

Hay que señalar que en estas matrices las filas representan el eje de frecuencias (con el subíndice *i* en la descripción algorítmica) y las columnas el eje de tiempos (con el subíndice *j*).

Debido a que el establecimiento de escalas FFT no se encuentra normalizado, habrá que procurar asegurar que se dé a la FFT utilizada en este algoritmo la escala necesaria para satisfacer la siguiente condición: cuando una trama de 128 muestras de valores reales, cada una de las cuales con valor 1 es una entrada a la FFT sin disposición en ventana, el valor complejo en el "bin" de CC de la salida FFT debe ser $128 + 0 \cdot j$.

II.2.5 Selección de tramas

En el cálculo de AD sólo se utilizan tramas que satisfacen o superan los umbrales de energía en X e Y. Para X, el umbral de energía se fija 15 dB por debajo de la energía de la trama de cresta en X:

$$energía\,en\,x(j) = \sum_{i=1}^{65} X(i,j),\ umbral\,en\,x = 10^{\frac{-15}{10}} \cdot \max_{j}(energía\,en\,x(j)).$$

Para Y, el umbral de energía se fija 35 dB por debajo de la energía de la trama de cresta en Y:

$$energía\,en\,y(j) = \sum_{i=1}^{65} Y(i,j),\ umbral\,en\,y = 10^{\frac{-35}{10}} \cdot \max_{j}(energía\,en\,y(j)).$$

Se toman las tramas que satisfacen o superan dichos umbrales de energía:

$$\{energía\,en\,x(j) \geq umbral\,en\,x\}\ \text{Y}\ \{energía\,en\,y(j) \geq umbral\,en\,y\} \Rightarrow \text{se retiene la trama } j.$$

Si una trama contiene una o más muestras iguales a cero, se elimina esa trama de X e Y. Estas matrices contienen ahora 65 filas y N3 columnas, donde N3 es el número de tramas que se han retenido. Si N3 = 0, los ficheros de entrada no contienen señales adecuadas, y termina el algoritmo.

Hay que señalar que los niveles umbral son diferentes para X e Y, ya que hay que prever la inclusión de tramas en Y que tengan potencia reducida debido a una perturbación perceptualmente significativa (por ejemplo, recorte temporal), pero que conserven potencia suficiente para ser comparadas con éxito con el material vocal original.

II.2.6 Aproximación de la sonoridad percibida

Se procede a transformar logarítmicamente cada una de las muestras del dominio de frecuencias en X e Y a una aproximación de la sonoridad percibida:

$$X(i,j) = 10 \cdot \log_{10}(X(i,j)),\ \ Y(i,j) = 10 \cdot \log_{10}(Y(i,j)),\ \ 1 \leq i \leq 65, 1 \leq j \leq N3.$$

II.2.7 Bloque de normalización de la medición de frecuencias (FMNB)

Un FMNB se aplica a X e Y en la escala temporal más larga disponible, definida por la longitud de los ficheros de entradas. Se extraen cuatro mediciones, que se almacenan en el vector de medición m. Estas mediciones se realizan en los bordes inferior y superior de la banda telefónica vocal (en las cuales resulta más fácil detectar ruido y cambios en la respuesta de frecuencia). Se utilizan los vectores temporales f1, f2 y f3:

$$f1(i) = \frac{1}{N3}\sum_{j=1}^{N3} Y(i,j) - \frac{1}{N3}\sum_{j=1}^{N3} X(i,j), 1 \leq i \leq 65, \text{ Medir}$$

$$f2(i) = f1(i) - f1(17), 1 \leq i \leq 65, \text{ Normalizar la medición a 1 kHz}$$

$$Y(i,j) = Y(i,j) - f2(i), 1 \leq i \leq 65, 1 \leq j \leq N3, \text{ Normalizar } Y$$

$$f3(i) = \frac{1}{4}\sum_{j=1}^{4} f2(1 + 4 \cdot (i-1) + j), 1 \leq i \leq 16, \text{ Suavizar la medición}$$

$$m(1)=f3(1),\ m(2)=f3(2),\ m(3) = f3(13),\ m(4), \text{ Salvar 4 mediciones}$$

II.2.8 Cálculo de bloques de normalización de la medición del tiempo

En la estructura de MNB, la parte central de la banda sufre dos niveles adicionales de división de banda binaria, lo que produce banda de una anchura aproximada de 2-3 Bark. Los extremos superior e inferior de la banda son tratados una vez por un TMNB separado. Por último, se realiza una

medición residual. Hay en total 9 TMNB en esta estructura, que se reproduce gráficamente en la figura II.4. La estructura de MNB genera 8 mediciones, aparte de las generadas por el FMNB inicial (descrito en II.2.7). Se utilizan las variables temporales *t*0, *t*1, ... *t*9.

TMNB-0 (parte inferior de la banda, anchura de 1,9 Bark):

$$t0(j) = \frac{1}{5}\sum_{i=2}^{6} Y(i,j) - \frac{1}{5}\sum_{i=2}^{6} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t0(j), 2 \le i \le 6, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(5) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t0(j),0), \text{ Salvar parte positiva de la medición}$$

TMNB-1 (parte central de la banda, capa superior, 10 Bark de anchura):

$$t1(j) = \frac{1}{36}\sum_{i=7}^{42} Y(i,j) - \frac{1}{36}\sum_{i=7}^{42} Y(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t1(j), 7 \le i \le 42, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(6) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t1(j),0), \text{ Salvar parte positiva de la medición}$$

$$m(7) = -\frac{1}{N3}\sum_{j=1}^{N3} \text{mín}(t1(j),0), \text{ Salvar parte negativa de la medición}$$

TMNB-2 (parte superior de la banda, 3 Bark de anchura):

$$t2(j) = \frac{1}{23}\sum_{i=43}^{65} Y(i,j) - \frac{1}{23}\sum_{i=43}^{65} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t2(j), 43 \le i \le 65, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(8) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t2(j),0), \text{ Salvar parte positiva de la medición}$$

TMNB-3 (parte central de la banda, capa central, 5 Bark de anchura):

$$t3(j) = \frac{1}{12}\sum_{i=7}^{18} Y(i,j) - \frac{1}{12}\sum_{i=7}^{18} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t3(j), 7 \le i \le 18, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(9) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t3(j),0), \text{ Salvar parte positiva de la medición}$$

TMNB-4 (parte central de la banda, capa central, 5 Bark de anchura):

$$t4(j) = \frac{1}{24}\sum_{i=19}^{42} Y(i,j) - \frac{1}{24}\sum_{i=19}^{42} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t4(j), 19 \le i \le 42, 1 \le j \le N3, \text{ Normalizar } Y$$

TMNB-5 (parte central de la banda, capa inferior, 2,5 Bark de anchura):

$$t5(j) = \frac{1}{5}\sum_{i=7}^{11} Y(i,j) - \frac{1}{5}\sum_{i=7}^{11} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t5(j), 7 \le i \le 11, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(10) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t5(j),0), \text{ Salvar parte positiva de la medición}$$

TMNB-6 (parte central de la banda, capa inferior, 2,5 Bark de anchura):

$$t6(j) = \frac{1}{7}\sum_{i=12}^{18} Y(i,j) - \frac{1}{7}\sum_{i=12}^{18} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t6(j), 12 \le i \le 18, 1 \le j \le N3, \text{ Normalizar } Y$$

TMNB-7 (parte central de la banda, capa inferior, 2,5 Bark de anchura):

$$t7(j) = \frac{1}{10}\sum_{i=19}^{28} Y(i,j) - \frac{1}{10}\sum_{i=19}^{28} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t7(j), 19 \le i \le 28, 1 \le j \le N3, \text{ Normalizar } Y$$

$$m(11) = \frac{1}{N3}\sum_{j=1}^{N3} \text{máx}(t7(j),0), \text{ Salvar parte positiva de la medición}$$

TMNB-8 (parte central de la banda, capa inferior, 2,5 Bark de anchura):

$$t8(j) = \frac{1}{14}\sum_{i=29}^{42} Y(i,j) - \frac{1}{14}\sum_{i=29}^{42} X(i,j), 1 \le j \le N3, \text{ Medir}$$

$$Y(i,j) = Y(i,j) - t8(j), 29 \le i \le 42, 1 \le j \le N3, \text{ Normalizar } Y$$

Medición residual:

$$t9(i,j) = Y(i,j) - X(i,j), 1 \le i \le 65, 1 \le j \le N3, \text{ Medición residual}$$

$$m(12) = \frac{1}{N364}\sum_{i=2}^{65}\sum_{j=1}^{N3} \text{máx}(t9(i,j),0), \text{ Salvar parte positiva de la medición residual}$$

Hay que señalar que si se retienen dos mediciones (parte positiva y parte negativa) para cada uno de los 9 TMNB en la estructura, se obtendrán en total 18 mediciones. La naturaleza jerárquica de la estructura de MNB, junto con la propiedad de equipotencia de los MNB produce la dependencia lineal entre estas 18 mediciones. Sólo puede disponerse de 7 mediciones de MNB linealmente independientes. Estas mediciones se combinan con la única medición residual y las 4 mediciones de FMNB, lo que arroja un total de 12 mediciones.

II.2.9 Combinación lineal de mediciones para la estructura de MNB

Se precede a combinar linealmente las 12 mediciones para generar un valor de AD. Los pesos utilizados en esta combinación lineal figuran en el cuadro II.2:

$$\text{AD} = \sum_{i=1}^{12} m(i) \cdot peso_i .$$

Cuando las 12 mediciones arrojan valores nulos, AD es cero.

Cuadro II.2/P.861 – Pesos para la combinación lineal de mediciones en la estructura de MNB

i	$peso_i$
1	0,0000
2	–0,0023
3	–0,0684
4	0,0744
5	0,0142
6	0,0100
7	0,0008
8	0,2654
9	0,1873
10	2,2357
11	0,0329
12	0,0000

NOTA – Puede parecer ineficaz computar los valores de la medición y multiplicarlos por un factor de ponderación nulo (esto es, $peso_1$ y $peso_{l2}$). Sin embargo, hay dos motivos que justifican el cómputo de estas mediciones. En primer lugar, el conjunto de mediciones, m_l a m_{l2}, representa el conjunto completo de valores de medición linealmente independientes disponibles como parte de la estructura de MNB. En segundo lugar, este conjunto de ponderaciones ($peso_i$) son las mejores ponderaciones para las aplicaciones y condiciones establecidas en el cuadro II.1. Para otras aplicaciones y condiciones, puede requerirse una función de ponderación diferente en que estas mediciones no arroje valores nulos en estas mediciones.

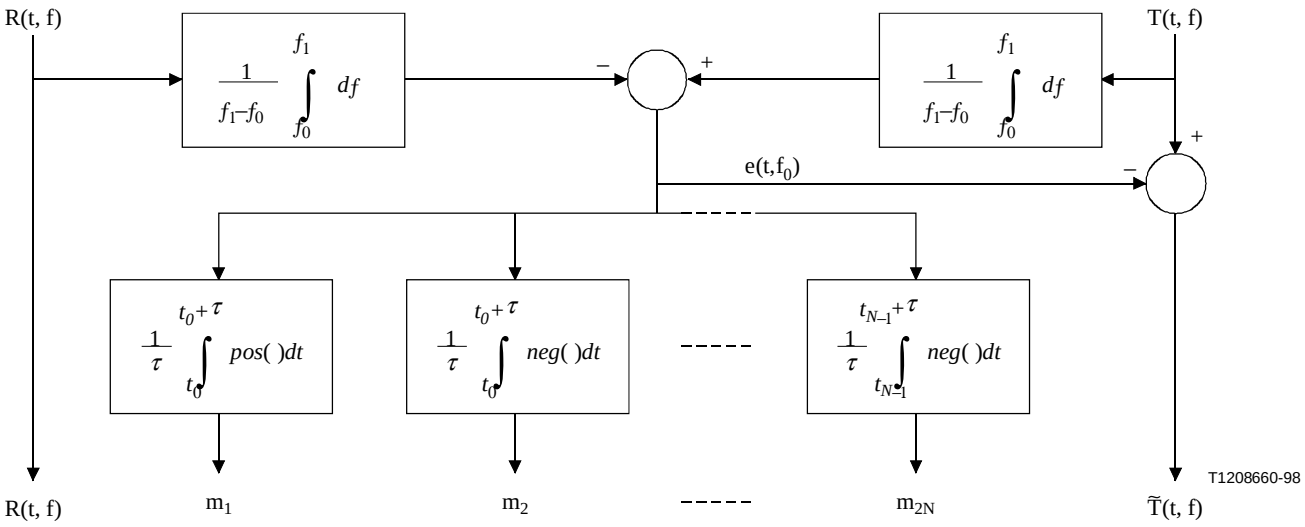

Figura II.1/P.861 – Bloque de normalización de la medición en tiempo (TMNB)

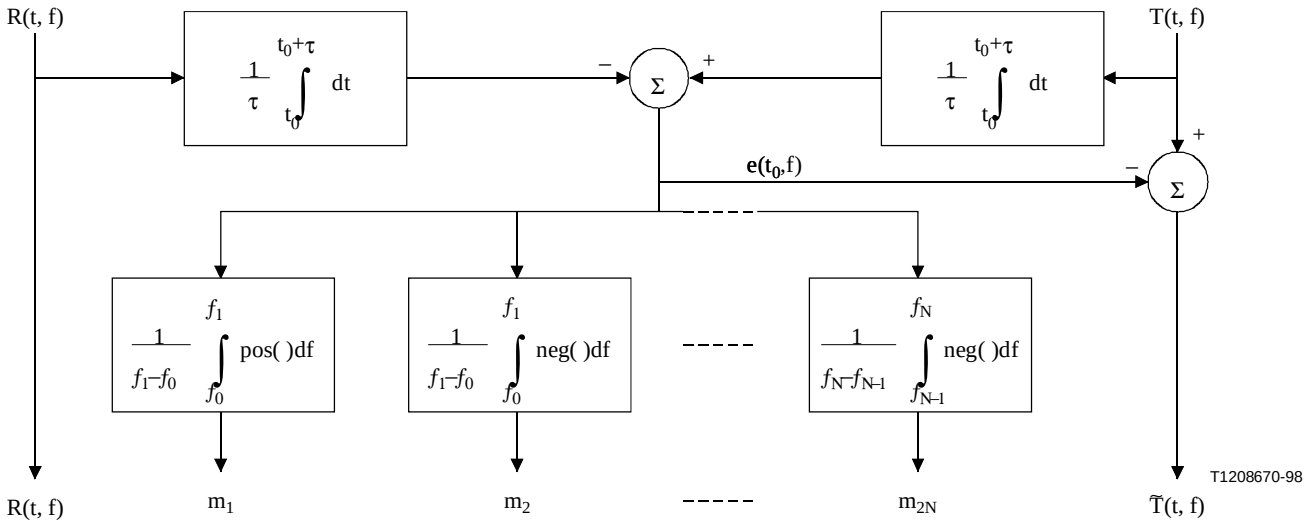

Figura II.2/P.861 – Bloque de normalización de la medición en frecuencia (FMNB)

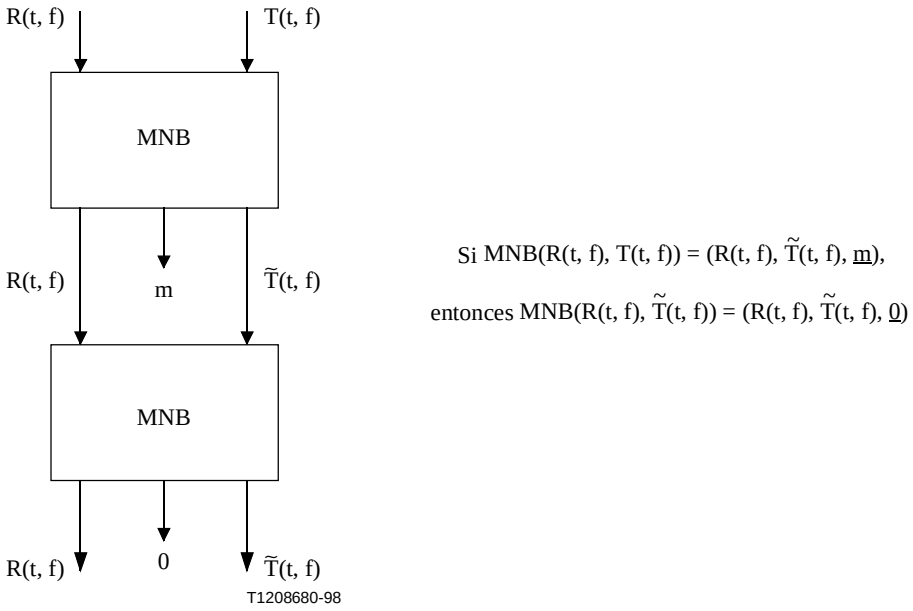

Figura II.3/P.861 – Los MNB son equipotentes

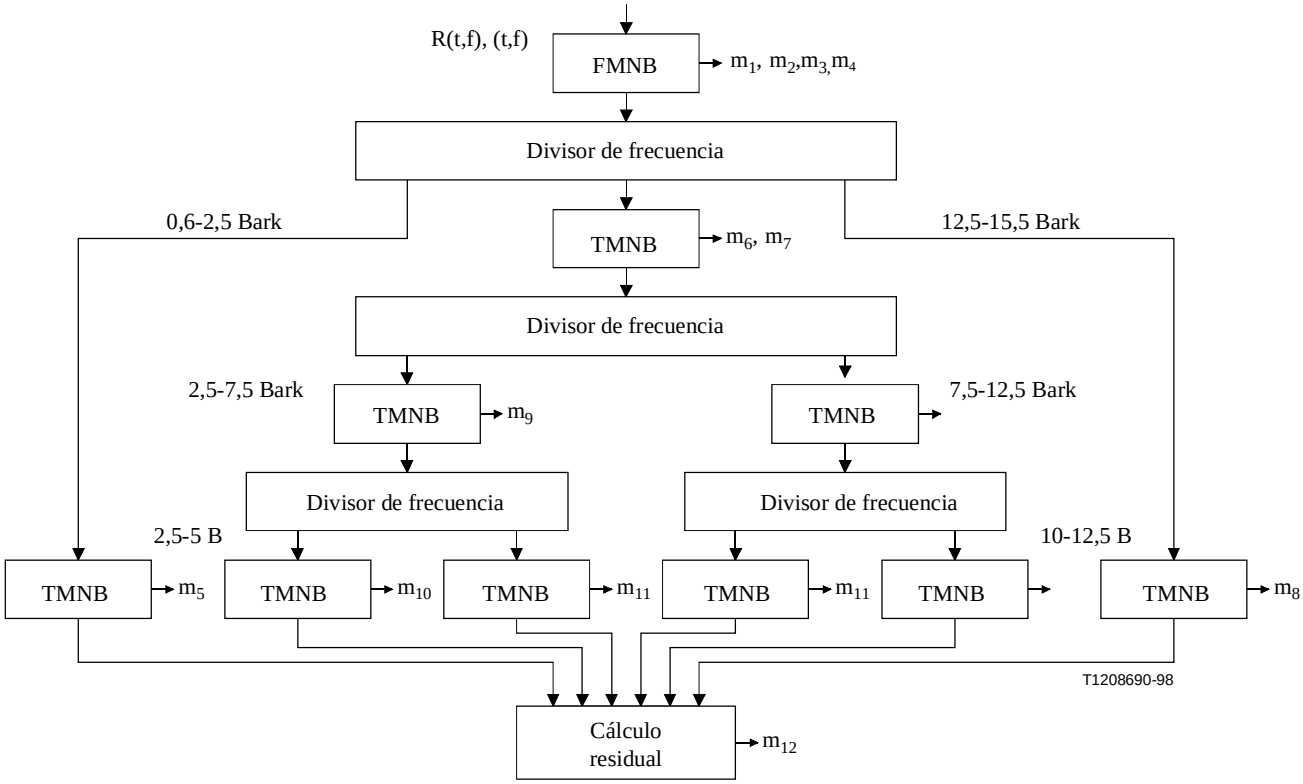

Figura II.4/P.861 – Estructura jerárquica de MNB

Bibliografía

[1] UIT-T, Contribución COM 12-25: Summary of available performance results for the MNB voice quality measurement algorithm, Estados Unidos, febrero de 1998.

[2] UIT-T, Contribución COM 12-26: Comparison of two methods for objective assessment of the quality of coded speech, *AT&T*, febrero de 1998.

[3] UIT-T, Contribución COM 12-27: Results of objective voice quality experiments, Estados Unidos, febrero de 1998.

[4] UIT-T, Contribución COM 12-20: Improvement of the P.861 Perceptual Speech Quality Measure, KPN, febrero de 1998.

[5] UIT-T, Contribución COM 12-21: An experimental investigation of the accumulation of perceived error in time-varying speech distortions, BT, febrero de 1998.

[6] UIT-T, Contribución COM 12-34: TOSQA – Telecommunication objective speech quality assessment, Alemania, febrero de 1998.

SERIES DE RECOMENDACIONES DEL UIT-T

Serie A	Organización del trabajo del UIT-T
Serie B	Medios de expresión: definiciones, símbolos, clasificación
Serie C	Estadísticas generales de telecomunicaciones
Serie D	Principios generales de tarificación
Serie E	Explotación general de la red, servicio telefónico, explotación del servicio y factores humanos
Serie F	Servicios de telecomunicación no telefónicos
Serie G	Sistemas y medios de transmisión, sistemas y redes digitales
Serie H	Sistemas audiovisuales y multimedios
Serie I	Red digital de servicios integrados
Serie J	Transmisiones de señales radiofónicas, de televisión y de otras señales multimedios
Serie K	Protección contra las interferencias
Serie L	Construcción, instalación y protección de los cables y otros elementos de planta exterior
Serie M	RGT y mantenimiento de redes: sistemas de transmisión, circuitos telefónicos, telegrafía, facsímil y circuitos arrendados internacionales
Serie N	Mantenimiento: circuitos internacionales para transmisiones radiofónicas y de televisión
Serie O	Especificaciones de los aparatos de medida
Serie P	**Calidad de transmisión telefónica, instalaciones telefónicas y redes locales**
Serie Q	Conmutación y señalización
Serie R	Transmisión telegráfica
Serie S	Equipos terminales para servicios de telegrafía
Serie T	Terminales para servicios de telemática
Serie U	Conmutación telegráfica
Serie V	Comunicación de datos por la red telefónica
Serie X	Redes de datos y comunicación entre sistemas abiertos
Serie Y	Infraestructura mundial de la información
Serie Z	Lenguajes de programación