(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 403 509**

(51) Int. CI.:

G10L 19/00 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **11.09.2007 E 11008485 (2)**

(97) Fecha y número de publicación de la concesión europea: **13.02.2013 EP 2410516**

(54) Título: **Método y sistema para la evaluación integral y diagnóstica de la calidad de la voz de escucha**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
20.05.2013

(73) Titular/es:

DEUTSCHE TELEKOM AG (50.0%)
Friedrich-Ebert-Allee 140
53113 Bonn, DE y
FRANCE TÉLÉCOM (50.0%)

(72) Inventor/es:

BARRIAC, VINCENT DIPL.-ING.;
CÔTÉ, NICOLAS DIPL.-ING;
GAUTIER-TURBIN, VALÉRIE DR.;
MÖLLER, SEBASTIAN PROF.DR.-ING;
RAAKE, ALEXANDER DR.-ING.;
WÄLTERMANN, MARCEL DIPL-ING.;
HEUTE, ULRICH y
SCHOLZ, KIRSTIN

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Método y sistema para la evaluación integral y diagnóstica de la calidad de la voz de escucha.

Campo de la invención

La invención está relacionada con sistemas de comunicación en general, y específicamente con un método y un sistema para determinar la calidad de la transmisión de un sistema de comunicación, en particular de un sistema de comunicación adaptado para la transmisión de voz

Antecedentes de la invención

Para la planificación, el diseño, la instalación, la optimización y la monitorización de redes de telecomunicación que proporcionan capacidades de transmisión de voz, se debe tener en cuenta la calidad que experimenta el usuario del servicio asociado. Normalmente, la calidad se cuantifica realizando experimentos de percepción con individuos en un entorno de laboratorio. Para evaluar la calidad de la voz transmitida, los individuos de prueba se someten a una situación de solo escucha o de conversación o bien reciben muestras de voz bajo estas condiciones, y evalúan la calidad de lo que han escuchado sobre varias escalas de evaluación. El Sector de Estandarización de Telecomunicaciones de la Unión Internacional de Telecomunicaciones proporciona guías para dichos experimentos y propone utilizar varias escalas de evaluación como se describen, por ejemplo, en la Rec. P.800 de la ITU-T, 1996, en la Rec. P.830 de la ITU-T, 1996, o en el Manual sobre Telefonometría de la ITU-T, 1992. La escala que se utiliza con más frecuencia es una escala de evaluación de categorías absolutas de 5 puntos sobre “calidad global”. La evaluación promedio de las evaluaciones objetivas obtenidas sobre esta escala se denomina Puntuación Media de Opinión, MOS. Las puntuaciones MOS se pueden cualificar en función de si se han obtenido en una situación de solo escucha o en una conversación y en función del contexto de los canales de transmisión de banda estrecha (ancho de banda de audio 300-3400 Hz), banda ancha (50-7000 Hz) o mixta (banda estrecha y banda ancha), tal y como se describe, por ejemplo, en la Rec. P.800.1 de la ITU-T (2006).

Debido a los esfuerzos y costes necesarios para llevar a cabo test subjetivos, se han desarrollado algoritmos que estiman la puntuación subjetiva esperada en un experimento de percepción en función de las señales de voz, o de los parámetros que caracterizan la red de telecomunicaciones. Las señales de voz se pueden generar de forma artificial, por ejemplo utilizando simulaciones, o se pueden grabar en redes en funcionamiento. En función de si existe disponibilidad o no de señales de voz en la entrada del canal de transmisión que se consideran, se pueden distinguir distintos tipos de modelos basados en señales:

- un modelo de referencia completo, el cual estima las puntuaciones subjetivas de calidad de escucha calculando una distancia o similitud entre representaciones adecuadas de la señal de entrada y la de salida, o deduciendo una medida de la distorsión a partir de la comparación de las señales de entrada y salida, y transformando el resultado sobre una escala asociada a la calidad subjetiva,

- un modelo sin referencia, el cual estima las puntuaciones subjetivas de calidad de escucha únicamente en función de la señal de salida; esto se puede realizar, por ejemplo, mediante la generación de una referencia artificial por parte del propio algoritmo, y llevando a cabo un análisis posterior de comparación de señales, tal y como se ha enunciado más arriba, y

- un modelo da calidad de conversación, el cual estima la puntuación de calidad para una situación de solo escucha, solo habla y/o una conversación.

En el documento “Estimation of the Quality Dimension ‘Directness/Frecuency content’ for the Instrumental Assessment of Speech Quality (Estimación de la Dimensión de Calidad ‘Contenido Direccionalidad/Frecuencia’ para la Evaluación mediante Equipos de Calidad de la Voz)” de K. Scholz y otros, Interspeech 2006 – ICSLP, vol. 3, 2006, páginas 1523-1526, se identifican tres dimensiones como relevantes para la evaluación de la calidad de una señal de voz que se ha transmitido sobre una red de telecomunicaciones moderna, estas son “contenido Direccionalidad/Frecuencia”, “Continuidad” y “Nivel de ruido”, en donde la calidad de voz total percibida de una señal de voz en términos de puntuación media de opinión (MOS) se puede expresar como una combinación lineal ponderada de las tres dimensiones de calidad que da como resultado una calidad global que cubre aproximadamente el 90% de la varianza total de las evaluaciones de la calidad de voz. Con respecto a una evaluación mediante equipos de la calidad de voz, se describe un método para la estimación de la dimensión de calidad “contenido Direccionalidad/Frecuencia”.

Existen varios tipos de modelos de referencia completos para los canales de transmisión de voz y audio. Estos consisten, en general, en un paso de proceso previo de las señales de entrada y salida, una transformación en una representación interna, un paso de comparación que devuelve un índice, seguido por unos pasos de integración y transformación que devuelven una puntuación de calidad estimada.

Para la transmisión de voz de banda estrecha, los modelos de referencia completos incluyen el modelo PESQ

descrito en la Recomendación P.862 de la ITU-T (2001), su precursor PSQM descrito en la Recomendación P.861 de la ITU-T (1998), el modelo TOSQA descrito en la Contribución Com 12-19 de la ITU-T (2001), así como el PAMS descrito en el documento "The Perceptual Analysis Measurement System for Robust End-to-end Speech Quality Assessment (El Sistema para Medida del Análisis de Percepción para la Evaluación Robusta de la Calidad de Voz Extremo a Extremo)" de A.W. Rix y M.P. Hollier, Proc. ICASSP del IEEE, 2000, vol. 3, pág. 1515-1518. Se describen otros modelos en el documento "Objective Modelling of Speech Quality with Psychoacoustically Validated Auditory Model (Modelado Objetivo de la Calidad de la Voz con un Modelo Auditivo Validado Psicoacústicamente)" de M. Hansen y B. Kollmeier, 2000, J. Audio Eng. Soc., vol. 48, pág. 395-409, en el documento "Objective Estimation of Perceived Speech Quality – Part I: Development of the Measuring Normalizing Block Technique (Estimación Objetiva de la Calidad de Voz Percibida – Parte I: Desarrollo de la Técnica de Bloques de Normalización de Mediciones)" de S. Voran, Trans. Speech Audio Process. del IEEE, 1999, vol. 7, núm. 4, pág. 371-382, en el documento "Instrumentelle Verfahren zur Sprachqualitätsschätzung – Modelle Auditiver Tests (Métodos Instrumentales para Estimar la Calidad de la Voz – Modelos de Pruebas Auditivas)", de J. Berger, 1998, tesis doctoral, Universidad de Kiel, Shaker Verlag, Aachen, en el documento "Psychoakustisch motivierte Maße zur instrumentellen Sprachgütebeurteilung (Dimensiones psicoacústicas motivadas para determinar la evaluación de la calidad de la voz)" de M. Hauenstein, 1997, tesis doctoral, Universidad de Kiel, Shaker Verlag, Aachen, y en el documento "An Objective Measure for Predicting Subjective Quality of Speech Coders (Una medida objetiva para Predecir la Calidad Subjetiva de Codificadores de Voz)" de S. Wang, A. Sekey y A. Gersho, 1992, J. Sel. Areas Comun. del IEEE, vol. 10, núm. 5, pág. 819-829.

El modelo de Wang, Sekey y Gersho utiliza una Distorsión Espectral de Corteza (BSD) la cual no incluye el efecto de enmascaramiento. El modelo PSQM (Medida de la calidad de la voz Percibida) se deriva del modelo PAQM (Medida de la Calidad de Audio Percibida) y se especializaba únicamente en la evaluación de la calidad de la voz. El PSQM incluye como nuevos efectos cognitivos la medida de la perturbación de nivel de ruido en un intervalo de silencio y una asimetría de la distorsión de percepción entre componentes que se encuentra o se introduce en el canal de transmisión. El modelo de Voran, denominado Bloque de Normalización de Medidas, utiliza una distancia auditiva entre las dos señales transformadas perceptivamente. El modelo de Hansen y Kollmeier utiliza un coeficiente de correlación entre las dos señales de voz transformadas a un estadio neural superior de percepción. El modelo PAMS (Sistema de Medición de Análisis de Percepción) es una extensión de la medida BSD que incluye nuevos elementos para anular los efectos debidos al retardo variable en los sistemas de Voz sobre IP y los filtros lineales en interfaces analógicas. El modelo TOSQA (Evaluación Objetiva de la Calidad de Voz en Telecomunicaciones; Berger, 1998) evalúa un canal de transmisión extremo a extremo incluyendo los terminales utilizando una medida de la similitud entre ambas señales transformadas perceptivamente. El modelo PESQ (Evaluación de la Percepción de la Calidad de la Voz) es una combinación de dos modelos precursores, PSQM y PAMS incluyendo la ecualización parcial de la respuesta en frecuencia.

Únicamente se han realizado unas pocas propuestas para canales de transmisión de voz de banda ancha (50-7000 Hz) o mezcla de banda estrecha y banda ancha. La ITU-T recomienda actualmente una extensión de su modelo PESQ de la Rec. P.862.2 (2005), denominado PESQ de banda ancha, WB-PESQ, que consiste, principalmente, en sustituir las características del filtro de entrada del PESQ por un filtro paso alto, y aplicarlo tanto a las señales de voz de banda estrecha como de banda ancha. Además, la versión de 2001 del TOSQA (Contr. COM 12-19 de la ITU-T, 2001) ha demostrado ser capaz de estimar el MOS también en un contexto de banda ancha, como el WB-PAMS (Del. Contr. D.001 de la ITU-T, 2001).

En la literatura se han descrito varios estudios para evaluar la consistencia de las estimaciones del WB-PESQ con evaluaciones subjetivas como, por ejemplo con el Del. Contr. D.070 de la ITU-T (2005), el documento "Objective Quality Assessment of Wideband Speech by an Extension of the ITU-T Recommendation P.862 (Evaluación Objetiva de la Calidad de la Voz en Banda ancha con una Extensión de la Recomendación P.862 de la ITU-T)" de A. Takahashi y otros, 2005, en los Proc. de la 9ª Conf. Int. de Comunicación y Tecnologías de Voz (Interspeech Lisboa 2005), Lisboa, pág. 3153-3156, el documento "Objective Quality Assessment of Wideband Speech Coding (Evaluación Objetiva de la Calidad de la Codificación de Voz en Banda ancha)" de N. Kitawaki y otros, 2005, en Trans. de Comun. del IEICE, vol. E88-B(3), pág. 1111-1118, o el documento "Analysis of a Quality Prediction Model for Wideband Speech Quality, the WB-PESQ (Análisis del Modelo de Predicción de Calidad para Calidad de Voz en Banda ancha, el WB-PESQ)" de N. Côté y otros, 2006, en Proc. del 2º Tutorial y Taller de Investigación sobre Calidad de Percepción de Sistemas de la ISCA, Berlín, pág. 115-122.

El procedimiento de evaluación consiste, en general, en analizar la relación entre las evaluaciones de audición obtenidas en un test de solo escucha, MOS_LQS (MOS Subjetiva de Calidad de Escucha), y sus puntuaciones correspondientes estimadas mediante un aparato MOS_LQO (MOS Objetiva de Calidad de Escucha). Por ejemplo, en Takahashi y otros (2005), se evaluaron tres codificadores de voz en banda ancha con WB-PESQ, y se encontró una preferencia por el códec G.722.1, con el que la MOS_LQO es significativamente menor que la MOS_LQS. En Kitawaki y otros (2005) se observó el mismo efecto para el códec G.722.2, aunque el coeficiente de correlación promedio es de aproximadamente 0,90. El WB-PESQ ha demostrado ser capaz de predecir la clasificación del códec en las evaluaciones de los oyentes, pero no ha sido capaz de cuantificar la diferencia perceptiva entre los codecs.

La siguiente tabla muestra los coeficientes de correlación de Pearson de la base de datos AQUAVIT (AQUAVIT – Evaluación de la Calidad de Señales Audiovisuales sobre Internet y UMTS, Proyecto P.905 de Eurescom, Marzo de 2001) para tres modelos de banda ancha:

Test:	Ancho de banda:	WB-PESQ	TOSQA-2001	WB-PAMS
1	Banda Mixta	0,952	0,966	0,946
2a	Banda Estrecha	0,981	0,954	0,981
2b	Banda Ancha	0,977	0,982	0,992

Como se puede observar a partir de estos datos, los modelos conocidos ya proporcionan puntuaciones estimadas de la calidad con una correlación importante. Sin embargo, los modelos típicamente no tienen la misma precisión para la voz transmitida en banda estrecha y en banda ancha. Además, si se detecta una mala calidad de una ruta de transmisión, a partir de la puntuación estimada de calidad no se puede deducir ninguna información sobre el origen de la pérdida de calidad.

Por lo tanto, es un objeto de la presente invención mostrar una aproximación nueva y mejorada para determinar una medida de la calidad de la voz asociada a una ruta de la señal de un sistema de transmisión de datos utilizado para la transmisión de voz. Otro objeto de la invención es proporcionar una medida de la calidad de la voz con una alta precisión para la voz transmitida en banda estrecha y en banda ancha. Otro objeto adicional de la invención es proporcionar una medida de la calidad de la voz a partir de la cual se pueda deducir un origen de la pérdida de calidad en la ruta de la señal.

Resumen de la invención

Los objetos de la presente invención se consiguen mediante cada uno de los contenidos de las respectivas reivindicaciones independientes adjuntas. Los contenidos de las respectivas reivindicaciones dependientes adjuntas incluyen refinamientos o modos de realización ventajosos y/o preferidos.

Los inventores han descubierto que, además de una estimación de la calidad de voz total, tal y como se expresa, por ejemplo, sobre una escala de calidad total de acuerdo con la Rec. P.800 de la ITU-T (1996), las dimensiones de percepción son importantes para la formación de la calidad. Además, las dimensiones de percepción proporcionan una imagen más detallada y analítica de la calidad de la voz transmitida, por ejemplo, por comparación entre los canales de transmisión, o para analizar las fuentes de componentes particulares del canal de transmisión sobre la calidad percibida. Las dimensiones se pueden definir sobre la base de características de la señal, tal y como se propone, por ejemplo, en el Contr. COM 12-4 de la ITU-T (2004) o el Contr. COM 12-16 de la ITU-T (2006), o sobre la base de una descomposición perceptiva de los eventos de sonido, tal y como se describe en el documento “Underlying Quality Dimensions of Modern Telephone Connections (Dimensiones de la Calidad Subyacente de las Conexiones de Telefonía Modernas)” de M. Wältermann y otros, 2006, en Proc. de la 9ª Conf. Int. de Procesamiento del Lenguaje hablado (Interspeech 2006 – ICSLP), Pittsburgh, PA, pág. 2170-2173. La invención propone, con gran ventaja, métodos para determinar dichas dimensiones individuales y para integrarlas en un modelo completo de referencia basado en señales para la estimación de la calidad de la voz. El término “dimensión de percepción” de una señal de voz se utiliza en la presente solicitud para describir un rasgo característico de una señal de voz que percibe por separado un oyente de la señal de voz.

De este modo, la invención propone, preferiblemente, una forma específica de modelo completo de referencia, el cual estima diferentes puntuaciones asociadas a la calidad de la voz, en particular para una situación de solo escucha.

En consecuencia, en un primer modo de realización, un método inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través de una ruta de la señal de un sistema de transmisión de datos produciendo dicha señal de salida, comprende los pasos de procesamiento previo de dichas señales de entrada y/o salida, determinar una tasa de interrupción de la señal de salida procesada previamente y/o determinar una medida de la intensidad de los tonos musicales presentes en la señal de salida procesada previamente, y determinar dicha medida de la calidad de la voz a partir de dicha tasa de interrupción y/o dicha medida de la intensidad de los tonos musicales. Este método está adaptado para determinar la dimensión de percepción asociada a la continuidad de la señal de salida.

Típicamente, tanto la señal de entrada como la de salida se procesan previamente, por ejemplo con el objetivo de ajustar los niveles. Debido a que en este primer modo de realización, sin embargo, típicamente sólo se sigue procesando la señal de salida procesada previamente, también puede ser una ventaja procesar previamente únicamente la señal de salida.

Con el fin de detectar interrupciones y/o tonos musicales en la señal, se determina lo más preferiblemente un

espectro de frecuencia discreto de la señal de salida procesada previamente dentro de al menos un intervalo de tiempo predefinido, en donde el espectro de frecuencia discreto es, preferiblemente, un espectro de corta duración generado mediante una transformación de Fourier discreta (DFT). El espectro de frecuencia discreto resultante de conformidad con la ventaja comprende valores de amplitud espectral para las parejas de frecuencia/tiempo basados en una tasa de muestreo predefinido y una serie de bandas de frecuencia predefinidas.

Preferiblemente, las bandas de frecuencia predefinidas se encuentran dentro de una rango de frecuencias predefinido con un límite inferior entre 0 Hz y 500 Hz y un límite superior entre 3 kHz y 20 kHz. El rango de frecuencias predefinido se selecciona en función de la aplicación, en particular en función de si las señales de voz son señales de banda estrecha, de banda ancha o de banda completa. Típicamente, los canales de transmisión de voz de banda estrecha se asocian con un rango de frecuencias entre 300 Hz y 3,4 kHz, mientras que los canales de transmisión de voz de banda ancha se asocian con un rango de frecuencias entre 50 Hz y 7 kHz. La banda completa se asocia, típicamente, con una frecuencia límite superior por encima de 7 kHz, el cual, en función de la aplicación, puede ser, por ejemplo, 10 kHz, 15 kHz, 20 kHz o incluso más alto. Por lo tanto, en función de la aplicación, las bandas de frecuencias predefinidas se encuentran, preferiblemente, dentro de uno de los rangos de frecuencias de más arriba.

En consecuencia, para las aplicaciones en las que las señales de voz sean señales de banda estrecha, las bandas de frecuencia predefinidas se encuentran, preferiblemente, dentro del rango de frecuencias típico de la banda de telefonía, esto es, en un rango esencialmente entre 300 Hz y 3,4 kHz. Ventajosamente, para aplicaciones de voz de banda ancha o de mezcla de banda estrecha y banda ancha el límite inferior es 50 Hz y el límite superior se encuentra entre 7 kHz y 8 kHz. Además, para aplicaciones de banda completa, el límite superior se encuentra, preferiblemente, por encima de 7 kHz, en particular por encima de 10 kHz, en particular por encima de 15 kHz, en particular por encima de 20 kHz.

Además, las bandas de frecuencia predefinidas son, preferiblemente, esencialmente equidistantes, en particular para la detección de tonos musicales.

El término espectro de frecuencia de corta duración se refiere a un espectro de densidad de amplitud, el cual se genera, en general, mediante una FFT (transformada rápida de Fourier) para un intervalo predefinido. En un espectro de frecuencia de corta duración, el intervalo de análisis es únicamente de corta duración lo cual proporciona una buena instantánea de la composición de la frecuencia, sin embargo, a costa de la resolución en frecuencia. Por lo tanto, la tasa de muestreo utilizada para generar el espectro de frecuencias discreto de la señal de salida procesada previamente se encuentra, preferiblemente, entre 0,1 ms y 200 ms, en particular entre 1 ms y 20 ms, en particular entre 2 ms y 10 ms.

Las interrupciones en la señal de salida procesada previamente se detectan, de forma ventajosa, mediante la determinación de un gradiente del espectro de frecuencia discreto, en donde el inicio de una interrupción se identifica mediante un gradiente que se encuentra por debajo de un primer umbral y el final de una interrupción se identifica mediante un gradiente que se encuentra por encima de un segundo umbral.

Para la detección de tonos musicales se establece, preferiblemente, un valor de amplitud esperado para cada par de frecuencia/tiempo del espectro de frecuencia discreto, en donde dichos tonos musicales se detectan mediante la determinación de pares de frecuencia/tiempo para las que el valor de la amplitud espectral es mayor que el valor de la amplitud esperada y la diferencia entre el valor de la amplitud espectral y el valor de la amplitud esperada excede un umbral predefinido.

En este primer modo de realización de un método inventivo la medida de la calidad de la voz se determina, preferiblemente, mediante el cálculo de una combinación lineal de la tasa de interrupción y la medida de la intensidad de tonos musicales detectados. Sin embargo, dentro del alcance de la invención también se encuentran combinaciones no lineales.

En un segundo modo de realización de un método inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través de una ruta de la señal de un sistema de transmisión de datos produciendo dicha señal de salida, comprende los pasos de: procesar previamente de dichas señales de entrada y/o salida, determinar, a partir de dichas señales de entrada y salida al menos un parámetro de calidad que sea una medida del ruido de fondo introducido en la señal de salida con respecto a la señal de entrada, y/o el centro de gravedad del espectro de dicho ruido de fondo, y/o la amplitud de dicho ruido de fondo, y/o el nivel de ruido de alta frecuencia introducido en la señal de salida con respecto a la señal de entrada, y/o el ruido correlacionado con la señal introducido en la señal de salida con respecto a la señal de entrada, en donde dicha medida de la calidad de la voz se determina a partir de dicho al menos un parámetro de calidad. Este método está adaptado para determinar la dimensión de percepción asociada al nivel de ruido de la señal de salida con respecto a la señal de entrada.

En las señales de entrada y salida procesadas previamente se detectan, de forma ventajosa, intervalos de actividad de la voz e intervalos de pausa en la voz. El parámetro de calidad que se mide para el ruido de fondo de forma más

ventajosa se determina comparando los espectros en frecuencia discretos de las señales de entrada y de salida procesados previamente dentro de dichas pausas de la voz. Preferiblemente, los espectros de frecuencia discretos se establecen como espectros de frecuencia de corta duración tal como se ha descrito más arriba. Preferiblemente, los espectros de frecuencia discretos se comparan calculando una diferencia sofométrica ponderada entre los espectros en un rango de frecuencias predefinido con un límite inferior entre 0 Hz y 0,5 Hz y un límite superior entre 3,5 kHz y 8,0 kHz.

Los inventores han descubierto que los valores apropiados de los límites con respecto al ruido de fondo para aplicaciones de banda estrecha son esencialmente 0 Hz para el límite inferior y esencialmente 4 kHz para el límite superior. Para aplicaciones de banda ancha preferiblemente el límite inferior es esencialmente 0 Hz y el límite superior se encuentra entre 7 kHz y 8 kHz. Por supuesto, en función de la aplicación o propósito, también se pueden elegir otros rangos de frecuencia.

Además, el método comprende, preferiblemente, el paso de calcular la diferencia entre el centro de gravedad del espectro de dicho ruido de fondo y un valor predefinido que representa un centro de gravedad ideal, en donde dicho valor predefinido es, en particular, igual a 2 kHz, debido a que el centro de gravedad en un rango de frecuencia entre 0 y 4 kHz para el “ruido blanco” tendría este valor.

El parámetro de calidad, que es una medida para el nivel de ruido de alta frecuencia, se determina, preferiblemente, como una proporción señal a ruido en un rango de frecuencias predefinido con un límite inferior entre 3,5 kHz y 8,0 kHz y un límite superior entre 5 kHz y 30 kHz.

Para aplicaciones de banda estrecha se ha encontrado que son preferibles un límite inferior de esencialmente 4 kHz y un límite superior de esencialmente 6 kHz. Para aplicaciones de banda ancha y/o de banda completa el límite inferior se encuentra preferiblemente entre 7 kHz y 8 kHz y el límite superior se encuentra preferiblemente por encima de 7 kHz, en particular por encima de 10 kHz, en particular por encima de 15 kHz, en particular por encima de 20 kHz.

Para determinar el parámetro de calidad, que es una medida para el ruido correlacionado con la señal, preferiblemente en un rango de frecuencias predefinido, se sustrae de un espectro de corta duración de la magnitud promedio de la señal de salida procesada previamente un espectro de corta duración de la magnitud promedio de la señal de entrada procesada previamente y un espectro de corta duración de la magnitud media del ruido de fondo estimado. Esta diferencia se normaliza con un espectro de corta duración de la magnitud promedio de la señal de entrada procesada previamente para describir el ruido correlacionado con la señal en la señal de salida procesada previamente. El espectro resultante se evalúa para determinar el parámetro de dimensión “ruido correlacionado con la señal”, en donde dicho rango de frecuencias predefinido tiene un límite inferior entre 0 Hz y 8 kHz y un límite superior entre 3,5 kHz y 20 kHz.

Un rango de frecuencias, que ha resultado ser el más preferible con respecto al ruido correlacionado con la señal, en particular para aplicaciones de banda estrecha, tiene un límite inferior de esencialmente 3 kHz y un límite superior de esencialmente 4 kHz.

La medida de la calidad de la voz asociada al nivel de ruido se determina, preferiblemente, calculando una combinación lineal o no lineal de algunos parámetros de calidad seleccionados de los de más arriba.

En un tercer modo de realización un método inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través de una ruta de señal de un sistema de transmisión de datos que produce dicha señal de salida, comprende los pasos de: procesar previamente dichas señales de entrada y/o salida, transformar el espectro de frecuencias de la señal de salida procesada previamente, en donde la escala de frecuencias se transforma en una escala de tonos, en particular la escala de Bark, y la escala de niveles se transforma en una escala de intensidad, detectar la parte de la señal de salida transformada que comprende voz, y determinar dicha medida de la calidad de la voz como un valor del tono promedio de la parte de señal detectada. Este método está adaptado para determinar la dimensión de percepción asociada a la intensidad de la señal de salida respecto a la señal de entrada.

Si las señales de entrada y salida son archivos digitales de voz, la medida de la calidad de la voz se determina, preferiblemente, en función del nivel digital y/o el modo de reproducción de dichos archivos digitales de voz y/o un nivel de presión de sonido predefinido.

En este tercer modo de realización, típicamente se procesa previamente tanto la señal de entrada como la de salida, por ejemplo, para el propósito de ajuste de niveles. Sin embargo, como también en este tercer modo de realización típicamente solo se sigue procesando la señal de salida procesada previamente, también puede ser beneficioso procesar previamente únicamente la señal de salida.

En un cuarto modo de realización, un método inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través

de una ruta de la señal de un sistema de transmisión de datos produciendo dicha señal de salida, comprende los pasos de: procesar previamente dichas señales de entrada y salida, determinar una respuesta en frecuencia y/o una función de ganancia correspondiente de la ruta de la señal a partir de las señales de entrada y salida procesadas previamente, determinar al menos un valor característico que represente una característica predefinida de la respuesta en frecuencia y/o la función de ganancia y determinar dicha medida de la calidad de la voz a partir de dicho al menos un valor característico.

Este método está adaptado para determinar la dimensión de percepción asociada al contenido de direccionalidad y/o frecuencia de la señal de salida respecto a la señal de entrada, en donde dicha al menos una característica predefinida comprende, preferiblemente, un ancho de banda de la función de ganancia, y/o un centro de gravedad de la función de ganancia, y/o una pendiente de la función de ganancia, y/o una profundidad de los picos y/o valles de la función de ganancia, y/o una anchura de los picos y/o valles de la función de ganancia. Sin embargo, también se puede utilizar cualquier otra característica asociada a la dimensión de percepción “contenido direccionalidad/frecuencia” de las señales de voz a analizar. Un ancho de banda se determina, lo más preferiblemente, como un ancho de banda rectangular equivalente (ERB) de la respuesta en frecuencia, debido a que ésta es una medida que proporciona una aproximación a los anchos de banda de los filtros del oído humano.

La función de ganancia se transforma ventajosamente en la escala de Bark, la cual es una escala psicoacústica propuesta por E. Zwicker que se corresponde con las bandas de frecuencia críticas del oído.

Además, las características predefinidas se determinan, preferiblemente, basándose en un intervalo de la respuesta en frecuencia y/o la función de ganancia seleccionados. En la práctica, la función de ganancia se descompone, preferiblemente, en una suma de una primera y una segunda funciones, en donde dicha primera función representa una función de ganancia suavizada y dicha segunda función representa un curso estimado de los picos y valles de la función ganancia.

Las características predefinidas determinadas se combinan para proporcionar la medida de la calidad de la voz, la cual es una estimación de la dimensión de percepción “contenido direccionalidad/frecuencia”, en donde, por ejemplo, se calcula una combinación lineal de los valores de las características. Sin embargo, lo más preferible es determinar la medida de la calidad de la voz mediante el cálculo de una combinación no lineal de los valores de las características, que se adapta para ajustarse a la banda de audio respectiva del canal de transmisión de voz que se está considerando.

El paso de procesamiento previo en cualquiera de los métodos descritos más arriba comprende, preferiblemente, los pasos de seleccionar una ventana en el dominio del tiempo para las señales de entrada y/o salida a procesar, y/o filtrar la señal de entrada y/o salida, y/o ajustar en el tiempo las señales de entrada y salida, y/o ajustar el nivel de las señales de entrada y salida, y/o corregir las distorsiones en frecuencia de la señal de entrada y/o salida, y/o seleccionar para su procesamiento únicamente la señal de salida. El ajuste de nivel de las señales de entrada y salida comprende, preferiblemente, normalizar a un nivel de señal predefinido tanto la señal de entrada como la de salida, en donde dicho nivel de señal predefinido es, esencialmente, de forma ventajosa 79 dB SPL, 73 dB SPL o 65 dB SPL.

Debido a que los métodos descritos más arriba se utilizan, lo más preferiblemente, para determinar las dimensiones de percepción individuales de las señales de voz en un modelo de referencia completo, en un quinto modo de realización un método inventivo para determinar una medida de la calidad de la voz de una señal de salida con respecto a una señal de entrada, en donde dicha señal de entrada pasa a través de una ruta de la señal de un sistema de transmisión de datos que produce dicha señal de salida, comprende los pasos de: procesar dichas señales de entrada y salida para determinar una primera medida de la calidad de la voz, determinar al menos una segunda medida de la calidad de la voz mediante la realización de un método de acuerdo con cualquiera de los modos de realización primero, segundo, tercero o cuarto descritos más arriba, y calcular una tercera medida de la calidad de la voz a partir de la primera medida de la calidad de la voz y la al menos una segunda medida de la calidad de la voz. El cálculo de la tercera medida de la calidad de la voz puede comprender calcular una combinación lineal o no lineal de las medidas de la calidad de la voz primera y segunda.

La primera medida de la calidad de la voz se determina, preferiblemente, mediante un método basado en un modelo de referencia completo conocido como, por ejemplo, el modelo PESQ o el TOSQA.

Preferiblemente, se determinan al menos dos segundas medidas de la calidad de la voz aplicando diferentes métodos. Lo más preferiblemente, se determinan cuatro segundas medidas de la calidad de la voz llevando a cabo, respectivamente los métodos de los modos de realización primero, segundo, tercero y cuarto descritos más arriba.

Las medidas de la calidad de la voz primera, segunda y/o tercera proporcionan, ventajosamente, una estimación de la evaluación de calidad subjetiva de la ruta de la señal esperada por un usuario promedio, en particular como un valor en la escala MOS, de aquí en adelante denominado también como puntuación MOS.

Un dispositivo inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con

respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través de una ruta de la señal de un sistema de transmisión de datos que produce dicha señal de salida, se adapta para ejecutar un método de acuerdo con uno cualquiera de los modos de realización primero, segundo, tercero o cuarto de los descritos más arriba.

Preferiblemente, el dispositivo comprende una unidad de procesamiento previo con entradas para recibir dichas señales de voz de entrada y salida, y una unidad de procesamiento conectada a la salida de la unidad de procesamiento previo, en donde dicha unidad de procesamiento comprende, preferiblemente, un microprocesador y una unidad de memoria.

Un sistema inventivo para determinar una medida de la calidad de la voz de una señal de voz de salida con respecto a una señal de voz de entrada, en donde dicha señal de entrada pasa a través de una ruta de la señal de un sistema de transmisión de datos que produce dicha señal de salida, comprende una primera unidad de procesamiento para determinar una primera medida de la calidad de la voz a partir de dichas señales de voz de entrada y salida, al menos un dispositivo tal como el que se ha descrito más arriba para determinar una segunda medida de la calidad de la voz a partir de dichas señales voz de entrada y salida y una unidad de consolidación conectada a las salidas de la primera unidad de procesamiento y de cada uno de dichos al menos un dispositivo, en donde dicha unidad de consolidación dispone de una salida para proporcionar dicha medida de la calidad de la voz y está adaptada para calcular un valor de salida a partir de las salidas de la primera unidad de procesamiento y cada uno de dicho al menos un dispositivo en función de un algoritmo predefinido.

Los dispositivos para determinar una segunda medida de la calidad de la voz tienen, preferiblemente, salidas respectivas para proporcionar dicha segunda medida de la calidad de la voz, que es una estimación de la calidad asociada con una dimensión de percepción individual respectiva.

Se proporcionan, preferiblemente, al menos dos dispositivos para determinar una segunda medida de la calidad de la voz, y más preferiblemente se proporciona un dispositivo para cada una de las dimensiones de percepción “contenido direccionalidad/frecuencia”, “continuidad”, “nivel de ruido” e “intensidad” descritas más arriba.

En un modo de realización preferido, el sistema comprende, además, una unidad de clasificación conectada a la salida de la unidad de consolidación para correlacionar la medida de la calidad de la voz con una escala predefinida, en particular con la escala MOS.

Breve descripción de las figuras

Se muestra en

Fig. 1 una vista esquemática de un modelo de referencia completo de la técnica anterior, y

Fig. 2 una vista esquemática de un modo de realización preferido de un sistema inventivo.

Descripción detallada de la invención

A continuación, se describen con más detalle ejemplos de modos de realización de la invención preferidos haciendo referencia a las figuras.

En la Fig. 1 se muestra esquemáticamente una configuración típica de un modelo de referencia completo conocido a partir la técnica anterior. Se proporciona a una unidad 210 de procesamiento previo una señal x(k) de entrada y una señal y(k) de salida, resultante de la transmisión de la señal x(k) de entrada a través de un canal 100 de transmisión. La unidad 210 se adapta, por ejemplo, para la determinación de la ventana del dominio del tiempo, filtrado previo, ajuste de tiempo, ajuste de niveles y/o corrección de la distorsión en frecuencia de las señales de entrada y salida dando como resultado las señales x’(k) e y’(k) procesadas previamente. Estas señales procesadas previamente se transforman en una representación interna mediante las respectivas unidades 221 y 222 de transformación, dando como resultado, por ejemplo, una representación provocada mediante percepción de ambas señales. La unidad 230 de comparación lleva a cabo una comparación de las dos representaciones internas en un índice unidimensional. Este índice, típicamente, está relacionado con la similitud y/o distancia de las tramas de las señales de entrada y salida, o se proporciona como un índice de distorsión estimado para la trama de la señal de salida comparada con la trama de la señal de entrada. Una unidad 240 de integración en el dominio del tiempo integra los índices para las tramas individuales en el tiempo de un índice para una muestra de voz completa. La puntuación de calidad estimada resultante, por ejemplo proporcionada como una puntuación MOS, es generada por la unidad 250 de transformación.

En la Fig. 2 se representa esquemáticamente un modo de realización preferido de un sistema inventivo 10 para determinar la medida de la calidad de la voz.

El sistema 10 que se muestra está adaptado para un nuevo modelo de referencia completo basado en señales para estimar la calidad de la voz transmitida tanto por banda estrecha como por banda ancha. Las características de este enfoque comprenden una estimación de cuatro puntuaciones dimensionales motivadas perceptivamente con la

ayuda de los estimadores 300, 400, 500 y 600 dedicados, la integración de una puntuación de calidad de escucha básica obtenida con la ayuda de un modelo de referencia completo y las puntuaciones dimensionales en una estimación de calidad global, y separar la salida de la puntuación de calidad global y las puntuaciones dimensionales con el fin de planificar, diseñar, optimizar, implementar, analizar y monitorizar la calidad de la voz.

El sistema que se muestra en la Fig. 2 comprende un estimador 300 para la dimensión de percepción “contenido direccionalidad/frecuencia”, un estimador 400 para la dimensión de percepción “continuidad”, un estimador 500 para la dimensión de percepción “nivel de ruido” y un estimador 600 para la dimensión de percepción “intensidad”. En el modo de realización que se muestra cada uno de los estimadores 300, 400, 500 y 600 comprende, respectivamente, una unidad 310, 410, 510 y 610 de procesamiento previo y, respectivamente, una unidad 320, 420, 520 y 620 de procesamiento. Sin embargo, también se puede proporcionar una unidad de procesamiento común para varios de los estimadores o para todos ellos.

Se proporciona una unidad 710 de consolidación de interferencias que combina una estimación de calidad básica obtenida mediante un estimador 200 básico basado en un modelo de referencia completo conocido con las estimaciones de calidad proporcionadas por los estimadores dimensionales 300, 400, 500 y 600. La estimación de calidad combinada se clasifica a continuación en la escala MOS mediante la unidad 720 de clasificación.

De forma especialmente ventajosa se proporciona un perfil de calidad de diagnóstico como una salida del sistema 10, que comprende una puntuación de calidad global estimada (MOS) y varias estimaciones dimensionales de percepción.

Como entrada a cada una de las unidades 200, 300, 400, 500 y 600 se proporcionan la señal x(k) de voz de referencia limpia, la señal y(k) de voz distorsionada, y en el caso de entrada digital, la frecuencia de muestreo. En el caso de que las interfaces acústicas formen parte de los canales de transmisión, las señales de voz son equivalentes a las señales eléctricas equivalente, que se aplican o se han obtenido en estas interfaces.

El estimador 200 básico se puede basar en cualquier modelo de referencia completo conocido como, por ejemplo, PESQ o TOSQA. Los componentes del estimador 200 básico se corresponden con los que se muestran en la Fig. 1.

Las unidades 310, 410, 510 y 610 de procesamiento previo están adaptadas, preferiblemente, para llevar a cabo un ajuste de tiempo entre las señales x(k) e y(k). El ajuste de tiempo puede ser el mismo que el que se utiliza en el estimador 200 básico o se puede adaptar de forma particular para cada estimador dimensional individual respectivo.

El estimador 300 de “contenido direccionalidad/frecuencia” está basado en parámetros medidos de la respuesta en frecuencia del canal 100 de transmisión. Estos parámetros comprenden, preferiblemente, el ancho de banda rectangular equivalente (ERB) y el centro de gravedad (Θ_G) de la respuesta en frecuencia. Ambos parámetros se miden sobre la escala de Bark. Otros parámetros apropiados comprenden la pendiente de la respuesta en frecuencia así como la profundidad y la anchura de los picos y valles de la respuesta en frecuencia.

La medida de la calidad de la voz proporcionada por el estimador 300 se determina, preferiblemente, calculando una combinación lineal de los parámetros de más arriba, esto es, mediante la siguiente ecuación

$$\hat{DF} = C_1 + C_2 \cdot ERB + C_3 \cdot \Theta_G + C_4 \cdot S + C_5 \cdot D + C_6 \cdot W$$

en donde

C_1-C_6	Constantes,
ERB:	Ancho de banda rectangular equivalente,
Θ_G:	Centro de gravedad,
S:	Pendiente,
D, W:	Profundidad y anchura de picos y valles.

Preferiblemente, las constantes C_1-C_6 se ajustan a un conjunto de muestras de voz apropiadas para el propósito respectivo. Esto se puede conseguir, por ejemplo, utilizando métodos de entrenamiento basados en redes neuronales artificiales.

A continuación se muestra un ejemplo de la ecuación de más arriba determinada por los inventores basada en un conjunto de ejemplo de muestras de voz y utilizando únicamente el ERB y el Θ_G.

$$\hat{DF} = -20.5865 + 0.2466\frac{ERB}{Bark} + 1.8730\frac{\Theta_G}{Bark}$$

Sin embargo, el cálculo de la medida de la calidad de la voz asociada al "contenido direccionalidad/frecuencia" no se encuentra limitado a una combinación lineal de los parámetros anteriores, sino que también comprende de forma especialmente ventajosa el cálculo de términos no lineales.

Por lo tanto, en un modo de realización más preferido la medida de la calidad de la voz proporcionada por el estimador 300 se determina calculando la siguiente ecuación:

$$\hat{DF} = \sum_{n=0}^{N}\sum_{m=0}^{M}\sum_{j=1}^{5}\sum_{i=1}^{5} C_{i,j,n,m}\cdot V_i^n \cdot V_j^m$$

en donde

$$V_1 = ERB\,;V_2 = \Theta_G\,;V_3 = S\,;V_4 = D\,;V_5 = W$$

$$N, M \in \{0, 1, 2, 3, ...\}$$

$C_{i,j,n,m}$: constantes con al menos un $C_{i,j,n,m} \neq 0$ con n > 0 y m > 0.

A continuación se ofrece un ejemplo preferido de la ecuación no lineal anterior:

$$\hat{DF} = -2.059\cdot C_A \cdot C_B + 4.485\cdot C_A^2 + 24.334\cdot C_A + 5.677\cdot C_B + 54.096$$

con

$$C_A = 3.79 - 0.38\cdot\frac{ERB}{Bark}$$

$$C_B = 2.12 - 0.23\cdot\frac{\Theta_G}{Bark}$$

En el modo de realización que se muestra, el estimador 400 para estimar la dimensión "continuidad" de calidad de la voz, denominado también más adelante como C-Meter, está basado en la estimación de dos parámetros de la señal: una tasa de interrupción de la señal de voz así como tonos musicales presentes dentro de una señal de voz.

A continuación se describe la funcionalidad de un ejemplo del modo de realización preferido del estimador 400.

La detección de una tasa de interrupción de la señal está basada en un algoritmo que detecta interrupciones de una señal de voz basándose en un análisis de la progresión temporal del gradiente de energía de la señal de la voz.

El algoritmo para la detección de interrupciones calcula primero el espectro de corta duración

$$X(\mu, i) = DFT\{x(k, i)\}$$

de la señal de voz distorsionada *x(k)*. En esta fórmula, el parámetro μ representa el índice de frecuencia de los valores de la DFT. El parámetro *i* indica el número de la trama actual de longitud *M* = 40 muestras (≙5 ms). Durante el cálculo del espectro $X(\mu,i)$ de corta duración, cada trama *x(k,i)* se pondera utilizando una ventana de Hamming. Las siguientes tramas no se solapan en este cálculo.

Para cada índice μ de frecuencia, el gradiente temporal $G_\mu(\mu,i,i+1)$ de la energía de la señal se calcula:

$$G_\mu(\mu, i, i+1) = |X(\mu, i+1)|^2 - |X(\mu, i)|^2 .$$

El sumatorio sobre todos los gradientes temporales $G_\mu(\mu,i,i+1)$ dentro de la región de frecuencia de la banda telefónica (μ_u ≙ 300 Hz - μ_o ≙ 3,4 kHz) proporciona el gradiente $G(i,i+1)$:

$$G(i,i+1) = \sum_{\mu=\mu_u}^{\mu o} G_\mu(\mu,i,i+1).$$

La normalización del gradiente $G(i,i+1)$ a la energía de la i-ésima trama proporciona el gradiente normalizado $G^n(i,i+1)$:

$$G^n(i,i+1) = \min\left(\frac{G(i,i+1)}{\sum_{\mu=\mu_u}^{\mu o} |X(\mu,i)|^2}, 1\right).$$

El resultado del gradiente de la energía se encuentra entre -1 y +1. Un gradiente de la energía con un valor de aproximadamente -1 indica una disminución extrema de energía tal y como sucede al inicio de una interrupción. Al final de una interrupción se observa un aumento extremo de energía que produce un gradiente de energía de aproximadamente +1.

El algoritmo detecta el inicio de una interrupción en el caso de que se produzca un gradiente de energía de $G^n(i,i+1) < -0{,}99$. El final de una interrupción se indica mediante el primer gradiente de energía siguiente de $G^n(i,i+1)=1$. Utilizando el conocimiento sobre la longitud total de una señal *x(k)* de voz y los indicadores de las interrupciones de inicio y fin, se puede calcular una tasa *Ir* de interrupción.

Para la utilización de este algoritmo para la estimación de la tasa de interrupción dentro del estimador 400 instrumental para la "continuidad", preferiblemente se adaptan algunas constantes dentro de este algoritmo con respecto a unos datos de prueba predefinidos para proporcionar estimaciones óptimas para la tasa de interrupción con un fin determinado.

La detección de tonos musicales está basada en la idea de la "Aproximación Relativa" descrita en el documento "Objective Evaluation of Acoustic Quality Based on a Relative Approach (Evaluación Objetiva de la Calidad Acústica Basada en una Aproximación Relativa)" de K. Genuit, 1996 en: Proc. Internoise'96, Liverpool, Reino Unido.

Como se describe en el documento "Application of the Relative Approach to Optimize Packet Loss Concealment Implementations (Aplicación de la Aproximación Relativa para Optimizar Implementaciones de Ocultación de Pérdida de Paquetes)" de F. Kettler y otros, 2003 en: Fortschritte der Akustik – DAGA 2003, Aachen, 18-20 de marzo de 2003, Deutsche Gesellschaft für Akustik, DEGA e.V., la idea que subyace a la "Aproximación Relativa" es comparar el valor de la señal real actual con una estimación del valor de la señal actual a partir de la historia de la señal para detectar cambios en el tiempo dentro de las señales acústicas que resultan inesperados y desagradables para el oído humano. Tal y como se describe en Genuit (1996) y Kettler (2003), la "Aproximación Relativa" incluye un modelo de la audición en el método de análisis.

En el C-Meter, esto es, en el estimador 400, se aplica directamente la idea de la "Aproximación Relativa" al espectro de corta duración de una señal de voz. Para detectar tonos musicales, se analiza un espectro de corta duración de una señal de voz dentro de bandas de frecuencia equidistantes. Los tonos musicales se detectan para aquellos pares tiempo-frecuencia *t, f,* en los que la amplitud espectral *X(t,f)* cumple dos condiciones: (1) la amplitud espectral actual real *X(t,f)* es mayor que la amplitud espectral actual esperada $\hat{X}(t,f)$, la cual es la media de los valores de amplitud espectral precedentes:

$$\hat{X}(t,f) = \frac{1}{N}\sum_{i=10}^{1} X(t-i,f);$$

y (2) la diferencia entre la amplitud espectral actual real y la estimación de la amplitud espectral real excede cierto umbral.

De este modo, de forma especialmente ventajosa, no se utiliza ningún modelo de audición en el C-Meter 300, al revés que en la conocida "Aproximación Relativa". En el C-Meter 300 únicamente se aplica la idea básica de la "Aproximación Relativa" de comparación del valor de la señal actual real con una estimación de la señal actual.

A partir de los resultados de la detección de los tonos musicales dentro de un archivo de voz se derivan dos parámetros que describen las características de los tonos musicales: un parámetro que indica la amplitud media de los tonos musicales MT_a, y un parámetro que indica la frecuencia de la ocurrencia de los tonos musicales, MT_f.

La estimación de la continuidad de la señal de voz se obtiene como una combinación lineal de los parámetros dimensionales "tasa de interrupción" e "intensidad del tono musical":

$$\hat{C} = 0.9274 - 0.7297 \cdot Ir - 0.0029 \cdot MT_a \cdot MT_f .$$

La ecuación anterior representa únicamente un ejemplo de modelo en el que se puede basar el estimador 300. Por supuesto, también se incluye dentro del alcance de la invención un modelo modificado o alterado. En particular, también se puede tener en cuenta que en la percepción humana de la dimensión “continuidad” tienen influencia más parámetros además de la “tasa de interrupción y la “intensidad del tono musical”. Ejemplos de dichos parámetros adicionales comprenden “tasa de recorte anterior/posterior” y “tasa de pérdida de paquetes”, ya que se prevé que también afecten a la percepción humana de la dimensión “continuidad”.

En el modo de realización que se muestra, el estimador 500 para la dimensión de percepción “nivel de ruido”, denominado también de aquí en adelante como N-Meter, está basado en la evaluación mediante equipos de cuatro parámetros que los inventores han encontrado que están asociados a la percepción humana de un nivel de ruido de la señal: un ruido de fondo de la señal BG_N, un parámetro que tiene en cuenta la distribución espectral del ruido de fondo de la señal FS_N, el nivel de ruido de alta frecuencia HF_N, y un ruido correlacionado con la señal SC_N. Se obtiene una estimación del “nivel de ruido” de un archivo de voz, $\hat{N}$, mediante una combinación lineal de estos cuatro parámetros

$$\hat{N} = \beta_0 + \beta_1 \cdot BG_N + \beta_2 \cdot FS_N + \beta_3 \cdot HF_N + \beta_4 \cdot SC_N .$$

El parámetro dimensional “ruido de fondo”, BG_N, se basa en un análisis del nivel de ruido durante las pausas de la voz:

$$BG_N = 10 \cdot \log_{10}\left[\frac{1}{96}\sum_{\mu=1}^{96} B_\mu \cdot \left(\frac{1}{K}\cdot\sum_{k=1}^{K}\left(\hat{\Phi}_{nn}(\Omega_{\mu,k}) - \Phi_{xx}(\Omega_{\mu,k})\right)\Big|_{k=pause}\right)\right].$$

Aquí, $\hat{\Phi}_{nn}(\Omega_\mu, k)\big|_{k=pause}$ describe el espectro de la densidad de potencia del archivo de voz procesado durante las pausas de la voz y, por lo tanto, se supone que describe el ruido de fondo contenido en un archivo de voz. $\Phi_{xx}(\Omega_\mu, k)\big|_{k=pause}$ describe el espectro del archivo de voz original durante las pausas de la voz. Se asume que la diferencia entre ambos espectros describe la cantidad de nivel de ruido añadido a la señal de la voz debido al procesamiento. Se promedian las diferencias entre ambos espectros correspondientes a todos los segmentos de tiempo $k = 1 \ldots K$. La diferencia media entre ambos espectros se pondera sofométricamente y se promedia para todos los valores de frecuencia desde 0 a 4 kHz, lo que se corresponde con el promedio de todos los índices de frecuencia $\mu = 1 \ldots 96$.

El parámetro dimensional “dispersión en frecuencia” FS_N, tiene en cuenta la forma espectral del ruido de fondo. Se supone que el contenido en frecuencia del nivel de ruido influye en la precepción humana del nivel de ruido. El ruido blanco parece menos molesto que el nivel de ruido coloreado. Además, el nivel de ruido más intenso parece más molesto que el nivel de ruido menos intenso. Estas suposiciones se encuentran verificadas por el test de audición de la dimensión “nivel de ruido” descrita en “Untersuchungen zur messtechnischen Erfassung und systematischen Beeinflussung der Sprachqualitäts- dimensión ‘Rauschhaftiegkeit’ (Los estudios de medición y grabación sobre la influencia sistemática sobre la calidad de voz de la dimensión 'Nivel de ruido')” de Ch. Kühnel, 2007, Tesis de Diploma, Instituto de Teoría de Circuitos y Sistemas, Universidad de Christian-Albrechts, Kiel. En la evaluación mediante equipos del “nivel de ruido” estas suposiciones se modelan mediante el parámetro dimensional FS_N:

$$FS_N = \left|f_{TP} - f_{opt}\right| \cdot A_{TP} .$$

$|f_{TP} - f_{opt}|$ describe la desviación del centro de gravedad del espectro del nivel de ruido con respecto del centro de gravedad ideal. En el caso de “ruido blanco” en el rango de frecuencias desde 0 Hz a 4 kHz, el espectro correspondiente es plano dentro del rango de frecuencias desde 0 Hz a 4 kHz y, por lo tanto, el centro de gravedad del espectro del nivel de ruido se encuentra en $f_{opt} = 2\ kHz$. En el caso de nivel de ruido coloreado, el centro de gravedad se desvía con respecto a este centro de gravedad ideal. El parámetro A_{TP} describe la energía del espectro de nivel de ruido. Por lo tanto, este parámetro modela el efecto de que el nivel de ruido más intenso es más molesto que el nivel de ruido menos intenso. Este efecto se modela en combinación con una desviación del centro de gravedad con respecto a su punto ideal.

Esto significa que se supone que siempre se produce una desviación del centro de gravedad con respecto a su punto ideal.

El parámetro dimensional “nivel de ruido de alta frecuencia” HF_N, se determina como una proporción señal ruido en el rango de frecuencias entre 4 kHz y 6 Hz.

$$NSR(\Omega_\mu, k) = 10 \cdot \log_{10} \frac{B_\mu \cdot \hat{\Phi}_{nn}(\Omega_\mu, k)\big|_{k=pause}}{A_\mu \cdot \Phi_{xx}(\Omega_\mu, k)\big|_{k=speech}}$$

En la presente solicitud, $\hat{\Phi}_{nn}(\Omega_\mu, k)\big|_{k=pause}$ describe el espectro de densidad de potencia del archivo de voz procesado durante las pausas de voz y $\Phi_{xx}(\Omega_\mu, k)\big|_{k=speech}$ describe el espectro del archivo de voz original durante la voz. Mientras que el sonido se pondera sofométricamente, el espectro de la voz se pondera utilizando el A-norm que modela la sensibilidad del oído humano. La proporción señal ruido *NSR(Ω_μ, k)* por índice de frecuencia Ω_μ e índice de tiempo *k* se integra sobre todos los índices de frecuencia y tiempo para proporcionar una estimación del nivel de ruido de alta frecuencia HF_N. Se utiliza una función de promedio sofisticada que utiliza diferentes L_p-norms.

Por ejemplo, para determinar el parámetro dimensional “ruido correlacionado con la señal” SC_N, primero se determina una diferencia entre un minuendo y un sustraendo. El minuendo se obtiene a partir de la relación entre el espectro de la magnitud media $|\bar{Y}(\mu)|$ de la señal de salida procesada previamente menos el espectro de la magnitud media $|\bar{X}(\mu)|$ de la señal original procesada previamente y el espectro de la magnitud media $|\bar{X}(\mu)|$ de la señal original procesada previamente. Los espectros medios $|\bar{X}(\mu)|$ e $|\bar{Y}(\mu)|$ se calculan como la media de los espectros de la magnitud de corta duración |*X(μ,n)*| e |*Y(μ,n)*| en los segmentos de señal con actividad de voz. Aquí, el parámetro *n* indica el número de segmento de señal considerado. El sustraendo se obtiene a partir de la relación entre el espectro de la magnitud media $|\bar{N}(\mu)|$ del ruido de fondo estimado y el espectro $|\bar{X}(\mu)|$ de la magnitud media de la señal original procesada previamente. El espectro de la magnitud media $|\bar{N}(\mu)|$ se calcula como el espectro promedio de la magnitud de corta duración |*Y(μ,n)*| en las pausas de la voz.

A continuación se representa la fórmula respectiva para calcular el espectro de ruido correlacionado con la señal:

$$NC(\mu) = \frac{|\bar{Y}(\mu)| - |\bar{X}(\mu)|}{|\bar{X}(\mu)|} - \frac{|\bar{N}(\mu)|}{|\bar{X}(\mu)|}.$$

siendo

$|\bar{Y}(\mu)|$: Espectro de la magnitud media de la señal de salida procesada previamente calculado dentro de segmentos de señal con actividad de voz,

$|\bar{X}(\mu)|$: Espectro de la magnitud media de la señal original procesada previamente, esto es, la señal de entrada, calculado dentro de segmentos de señal con actividad de voz,

$|\bar{N}(\mu)|$: Espectro de la magnitud media del ruido de fondo estimado,

μ: Índice de frecuencia,

en donde,

$$N(\mu) = \left(\frac{1}{K} \cdot \sum_{k=1}^{K} \left(\hat{\Phi}_{nn}(\Omega_{\mu,k}) \right) \big|_{k=pause} \right)$$

El parámetro dimensional “ruido correlacionado con la señal”, SC_N, se determina como una función del espectro de más arriba del ruido correlacionado con la señal esencialmente entre 3 kHz y 4 kHz:

$SC_N = f(NC(\mu))$

con

μ: Índices de frecuencia que se corresponden con frecuencias entre 3 kHz y 4 kHz.

El estimador 600 para la dimensión “intensidad” de la calidad de la voz, denominado también de aquí en adelante como L-Meter, está basado en el modelo de audición descrito en el documento “Procedure for Calculating the Loudness of Temporally Variable Sounds (Procedimiento para Calcular la Intensidad de Sonidos Temporalmente

Variables)" de E. Zwicker, 1977, J. Acoust. Soc. Ame., vol. 62, Nº 3, pág. 675-682. La señal de voz degradada se transforma en el dominio perceptivo. En particular, la escala de frecuencias se transforma en una escala de tonos y la escala de niveles se transforma en una escala de intensidades.

Sin embargo, el modelo de audición también se puede actualizar de forma ventajosa a uno más reciente como el modelo descrito en el documento "A Model of Loudness Applicable to Time-Varying Sounds (Un Modelo de Intensidad Aplicable a Sonidos que Varían en el Tiempo)" de B.R. Glasberg y B.C.J. Moore, 2002, J. Audio Eng. Soc., vol. 50, pág. 331-341, que está más relacionado con señales de voz.

Además, se utiliza una Detección de Actividad de Voz (VAD) para encontrar las partes de voz en la señal. El medidor de intensidad no tiene en cuenta las partes de señal que son sólo de ruido.

La medida de la calidad de la voz proporcionada por el medidor de intensidad 600 se corresponde con un promedio entre la parte de voz y la escala de tonos de la señal de voz degradada.

En particular, la intensidad se estima como una media sobre la escala de Bark (24 puntos) de una trama de 16 ms de la señal de salida de acuerdo con la siguiente ecuación:

$$\overline{Loudness}[n] = \frac{1}{24}\sum_{i=1}^{24} Loudness[i,n]$$

A continuación, se calcula una media sobre la parte de voz de acuerdo con la siguiente ecuación:

$$\overline{\overline{Loudness}} = \frac{1}{N}\sum_{i=1}^{N} \overline{Loudness}[N]$$

Estas *N* tramas de las partes de voz se localizan mediante un algoritmo de Detección de Actividad de Voz.

Para determinar la intensidad perceptiva real, se utilizan dos parámetros de entrada, el nivel de salida utilizado en el test auditivo (en dB SPL) que se corresponde con el nivel digital (en dB ovl) del archivo de voz, y el modo de reproducción, esto es, reproducción monoaural o biaural.

Los niveles digitales que se utilizan típicamente comprenden -26 dB ovl y -30 dB ovl, los valores de salida típicos comprenden 79 dB SPL (monoaural), 73 dB SPL (biaural) y 65 dB SPL (terminal de manos libres).

A continuación se describe la funcionalidad de la unidad 710 de consolidación.

Se utiliza la salida proporcionada por el estimador 200 básico para proporcionar una puntuación R_0 de referencia sobre la escala R extendida del modelo E definido en el rango de valores [0:130]. La escala *R* extendida es una versión ampliada de la escala *R* utilizada en el E-model. El E-model es un modelo de la calidad de la voz parametrizado, esto es, un modelo que utiliza parámetros en lugar de señales de voz, descrito en la recomendación G.107 de la ITU-T (2005). La escala *R* extendida se describe, por ejemplo, en el documento "Impairment Factor Framework for Wide-Band Speech Codecs (Esquema del Factor de Pérdida de Calidad para Códec de Voz de Banda Ancha)" de S. Möller y otros, 2006, Trans. sobre Procesamiento de Audio, Voz y Lenguaje del IEEE, vol. 14, núm. 6.

Este resultado tiene en cuenta únicamente la degradación no lineal debida a la parte de procesamiento como el códec de voz, algoritmos de ocultación de ruido, etc.

La salida del L-Meter 600 se transforma en un factor *Ie_loud* de pérdida de calidad mediante una función predefinida:

$$Ie\_loud = f(\overline{Loudness})$$

Este factor de pérdida de calidad también se define en el rango de valores [0:130]. Esta función puede ser no monótona debido a que se pueden considerar como degradaciones valores de la voz demasiado altos y demasiado bajos.

Las salidas del resto de estimadores 300, 400 y 500 también se transforman en factores de pérdida de calidad. Debido a que la degradación es una función de la intensidad, la salida del L-meter 600 también es un parámetro, lo que da como resultado las siguientes ecuaciones para los respectivos factores de pérdida de calidad:

$$Ie\_cont = g(\hat{C}, \overline{Loudness})$$

$$Ie\_direct = h(\hat{DF}, \overline{Loudness})$$

$$Ie\_noisiness = l(\hat{N}, \overline{Loudness})$$

Se proporciona una puntuación MOS_i para cada dimensión utilizando una función de asociación entre la puntuación R_i para esta dimensión y la MOS_i de acuerdo con las siguientes ecuaciones:

$R_i = R_0 - Ie_i$

$MOS_i = f(R_i)$

La puntuación R total, R_{ov}, se calcula a partir de la referencia R_0 y los diferentes factores Ie_i de pérdida de calidad utilizando la siguiente ecuación:

$R_{ov} = R_0 – Ie\_loud – Ie\_cont – Ie\_direct – Ie\_noisiness$

A partir de esto se determina la puntuación MOS como una función de la puntuación R total:

$MOS_{ov} = f(R_{ov})$

La invención se puede aplicar, por ejemplo, a cualquiera de los siguientes tipos de sistemas de telecomunicación que se corresponden con los canales de transmisión 100 de las Fig. 1 y 2:

- Redes públicas conmutadas, por ejemplo, red fija PSTN de cable, GSM, WCDMA, CDMA, etc.,

- Móvil-Pulsar para hablar, Voz sobre IP e interconexiones PSTN a VoIP, Tetra y

- componentes de procesamiento de voz utilizados comúnmente como, por ejemplo, codificadores, sistemas de reducción de nivel de ruido, control de ganancia adaptativo, nivel de ruido de confort, y sus combinaciones,

- canales de transmisión de banda estrecha, banda mixta, banda ancha y banda completa,

- 3G y redes de próxima generación incluyendo tecnologías de procesamiento de voz avanzadas, interfaces acústicas y aplicaciones de manos libres.

Los escenarios de aplicación de la técnica inventiva comprenden

- planificación de redes de telecomunicaciones, incluyendo equipamiento de terminales,

- optimización de componentes de red,

- comparación de redes y componentes de red,

- monitorización de redes y componentes,

- diagnósticos de funcionamientos defectuosos de redes y otros problemas, y

- cálculo y optimización de la carga de una red.

En consecuencia, dentro del concepto técnico de la presente divulgación se incluye también la utilización de cualquiera de los métodos para determinar una medida de la calidad de la voz descrita en la presente solicitud para cualquiera de los sistemas de telecomunicaciones citados más arriba y para cualquiera de los escenarios de aplicación citados más arriba.

Los métodos, dispositivos y sistemas propuestos permiten que la invención se pueda utilizar de forma especialmente ventajosa en aplicaciones de banda estrecha, banda ancha, banda completa y también para banda mixta, esto es, para determinar una medida de la calidad de la voz con respecto a un canal de transmisión adaptado para la transmisión de voz dentro del rango de frecuencias de la banda o bandas respectivas.

El alcance de la presente invención se define en las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un método para determinar una medida de la calidad de la voz de una señal de voz de salida (y) con respecto a una señal de voz de entrada (x), en donde dicha señal de entrada (x) pasa a través de una ruta de la señal (100) de un sistema de transmisión de datos que produce dicha señal de salida (y), que comprende los pasos de

- procesar previamente dichas señales de entrada y/o salida,

- detectar pausas de la voz en las señales de entrada y salida procesadas previamente,

- determinar, a partir de las señales de entrada (x_3) y de salida (y_3) procesadas previamente, al menos un parámetro de la calidad que sea una medida de

  - el ruido de fondo introducido en la señal de salida con respecto a la señal de entrada, y/o

  - el centro de gravedad del espectro de dicho ruido de fondo, y/o

  - la amplitud de dicho ruido de fondo, y/o

  - el ruido de alta frecuencia introducido en la señal de salida con respecto a la señal de entrada, y/o

  - el ruido correlacionado con la señal introducido en la señal de salida con respecto a la señal de entrada, en donde

el al menos un parámetro de calidad que es una medida del ruido de fondo se determina comparando los espectros en frecuencia discretos de las señales de entrada y salida procesadas previamente dentro de dichas pausas de la voz, y

- determinar dicha medida de la calidad de la voz a partir de dicho al menos un parámetro de calidad.

2. El método de la reivindicación 1, en el que la comparación de dichos espectros en frecuencia discretos comprende el cálculo de una diferencia ponderada sofométricamente entre los espectros en un rango de frecuencias predefinido con un límite inferior entre 0 Hz y 0,5 kHz y un límite superior entre 3,5 kHz y 8,0 kHz.

3. El método de una cualquiera de las reivindicaciones 1 ó 2, que comprende un paso para el cálculo de la diferencia entre el centro de gravedad del espectro de dicho ruido de fondo y un valor predefinido que representa un centro de gravedad ideal, en el que, en particular, dicho valor predefinido es igual a 2 kHz.

4. El método de una cualquiera de las reivindicaciones 1 a 3, en el que el parámetro de calidad que es una medida del ruido de alta frecuencia se determina como una proporción ruido señal en un rango de frecuencias predefinido con un límite inferior entre 3,5 kHz y 8,0 kHz y un límite superior entre 5 kHz y 30 kHz.

5. El método de una cualquiera de las reivindicaciones 1 a 4, que comprende los pasos de

- determinar un espectro de corta duración de la magnitud media de la señal de salida procesada previamente, de la señal de entrada procesada previamente y de un ruido de fondo estimado,

- sustraer de dicho espectro de corta duración de la magnitud media de la señal de salida procesada previamente el espectro de corta duración de la magnitud media de la señal de entrada procesada previamente y el espectro de corta duración de la magnitud media del ruido de fondo estimado,

- normalizar el resultado de la sustracción a un espectro de corta duración de la magnitud media de la señal de entrada procesada previamente, y

- determinar el parámetro de la calidad que es una medida del ruido correlacionado con la señal a partir del resultado normalizado dentro de un rango de frecuencias predefinido con un límite inferior entre 0 Hz y 8 kHz y un límite superior entre 3,5 kHz y 20 kHz.

6. El método de una cualquiera de las reivindicaciones 1 a 5, en el que el paso de procesamiento previo comprende los pasos de

- seleccionar una ventana en el dominio del tiempo para la señal de entrada y/o de salida a procesar, y/o

- filtrar la señal de entrada y/o de salida, y/o

- alinear el tiempo de las señales de entrada y salida, y/o

- ajustar el nivel de las señales de entrada y salida, y/o

- corregir las distorsiones en frecuencia de la señal de entrada y/o de salida, y/o

- seleccionar únicamente la señal de salida para su procesamiento.

7. El método de la reivindicación 6, en el que dicho ajuste de nivel de las señales de entrada y salida comprende normalizar ambas señales de entrada y salida respecto a un nivel de señal predefinido.

8. El método de la reivindicación 7, en el que dicho nivel de señal predefinido es, esencialmente, 79 dB SPL, 73 dB SPL ó 65 dB SPL.

9. Un dispositivo (300, 400, 500, 600) para determinar una medida de la calidad de la voz de una señal de voz de salida (y) con respecto a una señal de voz de entrada (x), en el que dicha señal de entrada (x) pasa a través de una ruta (100) de la señal de un sistema de transmisión de datos que produce dicha señal de salida (y), adaptado para llevar a cabo un método de acuerdo con una cualquiera de las reivindicaciones 1 a 8.

10. El dispositivo de la reivindicación 9, que comprende

- una unidad de procesamiento previo (310, 410, 510, 610) con entradas para recibir dichas señales de voz de entrada (x) y salida (y), y

- una unidad de procesamiento (320, 420, 520, 620) conectada a la salida de la unidad de procesamiento previo (310, 410, 510, 610).

11. Un método para determinar una medida de la calidad de la voz de una señal de salida (y) con respecto a una señal de entrada (x), en el que dicha señal de entrada (x) pasa a través de una ruta (100) de la señal de un sistema de transmisión de datos que produce dicha señal de salida (y), que comprende los pasos de

- procesar dichas señales de entrada y salida para determinar una primera medida de la calidad de la voz,

- determinar al menos una segunda medida de la calidad de la voz mediante la ejecución de un método de acuerdo con una cualquiera de las reivindicaciones 1 a 8, y

- calcular una tercera medida de la calidad de la voz a partir de la primera medida de la calidad de la voz y la al menos una segunda medida de la calidad de la voz.

12. El método de la reivindicación 11, en donde dicha primera medida de la calidad de la voz se determina mediante un método basado en el modelo de referencia completa PESQ o TOSQA.

13. Un sistema (10) para determinar una medida de la calidad de la voz de una señal de voz de salida (y) con respecto a una señal de voz de entrada (x), en donde dicha señal de entrada (x) pasa a través de una ruta (100) de la señal de un sistema de transmisión de datos que produce dicha señal de salida (y), que comprende

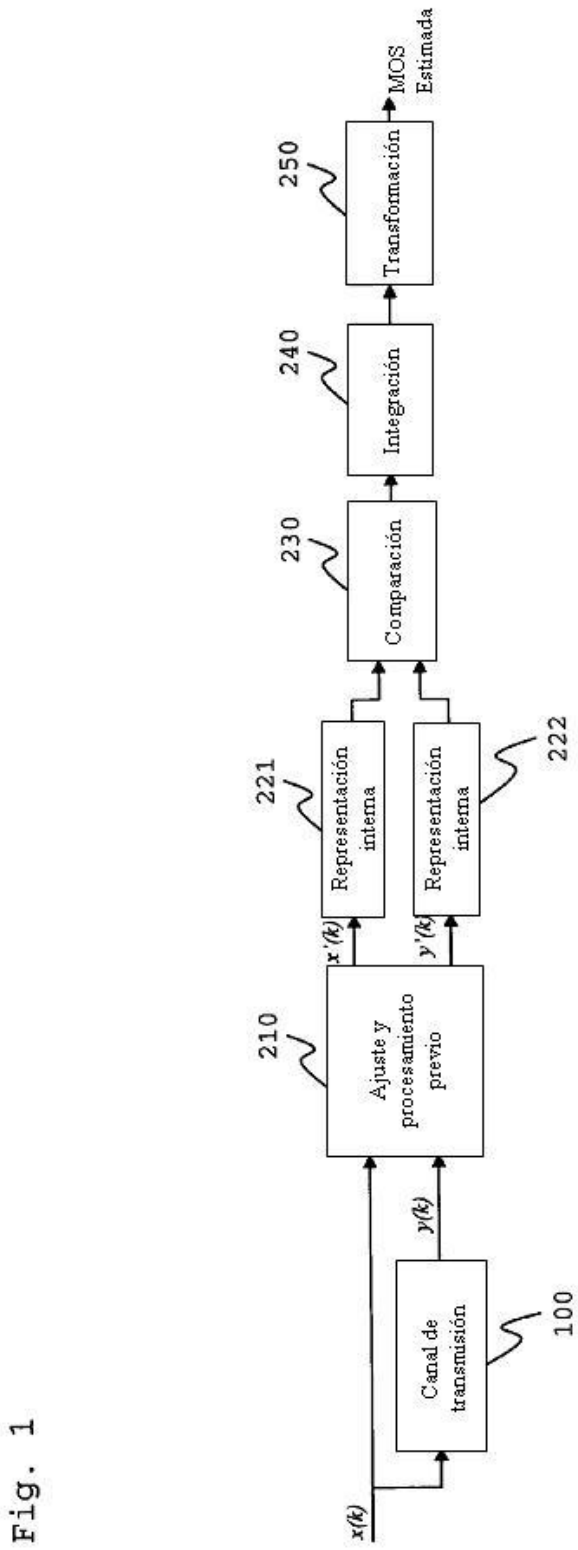
- una primera unidad de procesamiento (200) para determinar una primera medida de la calidad de la voz a partir de dichas señales de voz de entrada y salida,

- al menos un dispositivo (300, 400, 500, 600) de acuerdo con la reivindicación 9 ó 10 para determinar una segunda medida de la calidad de la voz a partir de dichas señales de voz de entrada y salida, y

- una unidad de consolidación (710) conectada a las salidas de la primera unidad de procesamiento (200) y cada uno de dicho al menos un dispositivo (300, 400, 500, 600), en donde dicha unidad de consolidación (710) tiene una salida para proporcionar dicha medida de la calidad de la voz y está adaptada para calcular un valor de salida a partir de las salidas de la primera unidad de procesamiento (200) y cada uno de dicho al menos un dispositivo (300, 400, 500, 600) en función de un algoritmo predefinido.

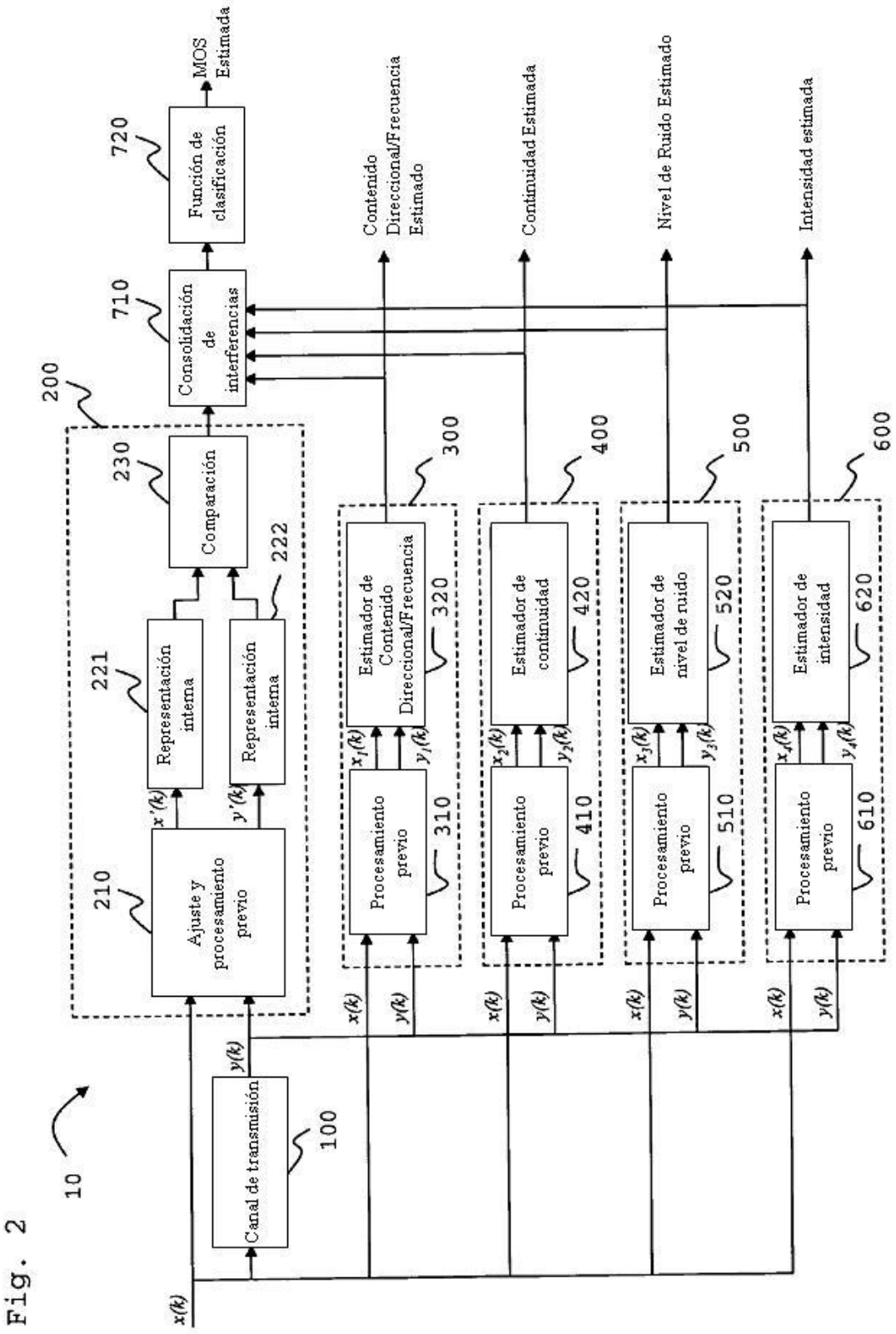
14. El sistema de acuerdo con la reivindicación 13 que comprende, además, una unidad de clasificación (720) conectada a la salida de la unidad de consolidación (710) para correlacionar la medida de la calidad de la voz dentro de una escala predefinida, en particular la escala MOS.

Fig. 1



Técnica Anterior

Fig. 2


10
Canal de transmisión
100
x(k)
y(k)
200
Ajuste y procesamiento previo
210
Representación interna
221
Representación interna
222
x'(k)
y'(k)
Comparación
230
Consolidación de interferencias
710
Función de clasificación
720
MOS Estimada
Procesamiento previo
310
Estimador de Contenido Direccional/Frecuencia
320
300
Contenido Direccional/Frecuencia Estimado
Procesamiento previo
410
Estimador de continuidad
420
400
Continuidad Estimada
Procesamiento previo
510
Estimador de nivel de ruido
520
500
Nivel de Ruido Estimado
Procesamiento previo
610
Estimador de intensidad
620
600
Intensidad estimada