(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 769 061**

(51) Int. CI.:

G10L 19/06 (2013.01)
G10L 21/0208 (2013.01)
G10L 19/125 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **23.09.2016 PCT/EP2016/072701**

(87) Fecha y número de publicación internacional: **30.03.2017 WO17050972**

(96) Fecha de presentación y número de la solicitud europea: **23.09.2016 E 16770500 (3)**

(97) Fecha y número de publicación de la concesión europea: **11.12.2019 EP 3353783**

(54) Título: **Codificador y método para codificar una señal de audio con ruido de fondo reducido que utiliza codificación predictiva lineal**

(30) Prioridad:

25.09.2015 EP 15186901
21.06.2016 EP 16175469

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.06.2020

(73) Titular/es:

FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (100.0%)
Hansastraße 27c
80686 München, DE

(72) Inventor/es:

FISCHER, JOHANNES;
BÄCKSTRÖM, TOM y
JOKINEN, EMMA

(74) Agente/Representante:

ARIZTI ACHA, Monica

DESCRIPCIÓN

Codificador y método para codificar una señal de audio con ruido de fondo reducido que utiliza codificación predictiva lineal

La presente invención se refiere a un codificador para codificar una señal de audio con ruido de fondo reducido que utiliza codificación predictiva lineal, un método correspondiente y un sistema que comprende codificador y un decodificador. En otras palabras, la presente invención se refiere a un enfoque conjunto de codificación y/o mejora de voz, tal como, por ejemplo, codificación y mejora conjuntas de voz por incorporación en un codificador-decodificador (códec) CELP (codificación lineal predictiva excitada por código).

Ya que se han extendido dispositivos de voz y comunicación y es probable que se usen en condiciones adversas, ha incremento la demanda de métodos de mejora de voz que puedan hacer frente a entornos adversos. En consecuencia, por ejemplo, en teléfonos móviles ahora es común utilizar métodos de atenuación de ruido como un paso/bloque de procesamiento previo para todo el procesamiento posterior de voz tal como codificación de voz. Existen diferentes enfoques que incorporan mejora de voz en codificadores de voz [1, 2, 3, 4]. En tanto que estos diseños mejoran la calidad de voz transmitida, el procesamiento en cascada no permite una reducción al mínimo/optimización perceptual, conjuntas de calidad, o ha sido al menos difícil una reducción al mínimo conjunta de ruido de cuantificación e interferencia.

El objetivo de los códecs de voz es permitir transmisión de voz de alta calidad con una cantidad mínima de datos transmitidos. Para alcanzar ese objetivo se necesitan representaciones eficientes de la señal, tal como modelado de la envolvente espectral de la señal de voz por predicción lineal, la frecuencia fundamental por un predictor de largo plazo y el resto con un libro de códigos de ruido. Esta representación es la base de códecs de voz que utilizan el paradigma de codificación lineal predictiva excita por código (CELP), que se utiliza en la mayoría de las normas de codificación de voz tal como multitasa adaptativa (AMR), AMR de banda ancha (AMR-WB), codificación unificada de voz y audio (USAC) y servicio de voz mejorado (EVS) [5, 6, 7, 8, 9, 10, 11].

Para comunicación de voz natural, los hablantes a menudo utilizan dispositivos en modos de manos libres. En estos escenarios el micrófono por lo general está lejos de la boca, por lo cual la señal de voz puede distorsionarse fácilmente por interferencias tal como reverberación o ruido de fondo.

La degradación no solo afecta la calidad de voz percibida, sino también la inteligibilidad de la señal de voz y por lo tanto puede impedir de forma severa la naturalidad de la conversación. Para mejorar la experiencia de comunicación, entonces es benéfico aplicar métodos de mejora de voz para atenuar ruidos y reducir los efectos de reverberación. El campo de mejora de voz es maduro y están fácilmente disponibles muchos métodos [12]. Sin embargo, la mayoría de algoritmos existentes se basan en métodos de superposición y suma, tal como transformadas como la transformada corta de Fourier en el tiempo (STFT), que aplica esquemas de partición en ventanas basados en superposición y suma, mientras que, en contraste, los códecs CELP modelan la señal con un filtro predictivo lineal/predictor lineal y aplican partición en ventanas solo en el residuo. Estas diferencias fundamentales hacen difícil unir métodos de mejora y codificación. Aún es claro que la optimización conjunta de mejora y codificación puede mejorar potencialmente la calidad, reducir el retardo y complejidad computacional.

Por lo tanto, existe una necesidad de un enfoque mejorado.

El documento EP1 944 761 A1 divulga un método para transmitir una señal digital y(n), y(n) que comprende una señal útil s(n) y una señal de perturbación p(n). El método comprende los pasos de: - recibir los Coeficientes de Predicción Lineal (LPC) A y la señal y e (n), y e (n) siendo una señal de LPC codificado de y(n). -estimar la matriz de autocorrelación s de la señal útil s(n), de la matriz de autocorrelación p de la señal de perturbación p(n) y LPC A p de la señal de perturbación p(n); - calcular un LPC A s modificado usando A y un sp, Ap estimado; - generar un flujo de datos modificado y e'(n) que incluya el LPC A s modificado,

El documento "Codebook driven short-term predictor parameter estimation for speech enhancement" (Sriram Srinivasan, Jones Samuelsson, y W. Bastiaan Kleijn) divulga una nueva técnica para la estimación de parámetro predictivos lineales a corto plazo de discurso y ruido desde datos ruidosos y su posterior uso en esquemas de mejora de forma de onda.

El documento US 6.263.307 B1 divulga un filtro de supresión acústica que incluye filtrado de atenuación con una estimación libre de ruido basada en un libro de código de frecuencias espectrales.

Es un objeto de la presente invención proporcionar un concepto mejorado para procesar una señal de audio utilizando codificación predictiva lineal. Este objeto se resuelve por la materia de las reivindicaciones independientes.

Las realizaciones de la presente invención muestran un codificador para codificar una señal de audio con ruido de

fondo reducido utilizando codificación predictiva lineal. El codificador comprende un estimador de ruido de fondo configurado para estimar ruido de fondo de la señal de audio, un reductor de ruido de fondo configurado para generar señal de audio de ruido de fondo reducido al restar el ruido de fondo estimado de la señal de audio de la señal de audio, y un predictor configurado para someter la señal de audio a análisis de predicción lineal para obtener un primer conjunto de coeficientes de filtro de predicción lineal (LPC) y para someter a la señal de audio de ruido de fondo reducido a análisis de predicción lineal para obtener un segundo conjunto de coeficientes de filtro de predicción lineal (LPC). Además, el codificador comprende un filtro de análisis compuesto de una cascada de filtros en el dominio del tiempo controlados por el primer conjunto obtenido de coeficientes CELP y el segundo conjunto obtenido de coeficientes LPC.

La presente invención se basa en el hallazgo que un filtro de análisis mejorado en un entorno de codificación predictiva, lineal incrementa las propiedades de procesamiento de señal de codificador. De forma más específica, utilizando una cascada o una serie de filtros en el dominio del tiempo conectados en serie mejora la velocidad de procesamiento o el tiempo de procesamiento de la señal de audio de entrada si los filtros se aplican a un filtro de análisis del entorno de codificación predictiva, lineal. Esto es ventajoso ya que se omiten la conversión de tiempo-frecuencia convencionalmente utilizada y la conversión de frecuencia-tiempo inversa de la señal de audio en el dominio del tiempo, entrante para reducir ruido de fondo al filtrar bandas de frecuencia que están dominadas por ruido. En otras palabras, al llevar a cabo la reducción o cancelación de ruido de fondo como una parte del filtro de análisis, se puede llevar a cabo reducción de ruido de fondo en el dominio del tiempo. Por lo tanto, el procedimiento de superposición y suma de, por ejemplo, se omite una MDCT/IDMCT (transformada discreta de coseno, modificada [inversa]), que se puede utilizar para conversión de tiempo/frecuencia/tiempo. Este método de superposición y suma limita la característica de procesamiento en tiempo real del codificador, ya que la reducción de ruido de fondo no se puede llevar a cabo en un solo cuadro, sino solo en cuadros consecutivos.

En otras palabras, el codificador descrito es capaz de llevar a cabo la reducción de ruido de fondo y por lo tanto el procesamiento completo del filtro de análisis en un solo cuadro de audio, y por lo tanto permite procesamiento en tiempo real de una señal de audio. El procesamiento en tiempo real puede referirse a un procesamiento de la señal de audio sin un retardo perceptible para usuarios participantes. Un retardo perceptible puede presentarse, por ejemplo, en una teleconferencia si un usuario tiene que esperar una respuesta del otro usuario debido a un retardo de procesamiento de la señal de audio. Este retardo máximo, permitido puede ser menos de 1 segundo, de manera preferente por debajo de 0,75 segundos, o incluso de manera más preferente por debajo de 0,25 segundos. Se tiene que señalar que estos tiempos de procesamiento se refiere al procesamiento completo de la señal de audio del emisor al receptor y por lo tanto incluyen, además del procesamiento de señal del codificador también el tiempo de transmisión de la señal de audio y el procesamiento de señal en el decodificador correspondiente.

De acuerdo con realizaciones, la cascada de filtros en el dominio del tiempo, y por lo tanto el filtro de análisis, comprenden dos veces un filtro de predicción lineal que utiliza el primer conjunto obtenido de coeficientes LPC y una vez un inverso de un filtro de predicción lineal adicional que utiliza el segundo conjunto obtenido de coeficientes LPC. Este procesamiento de señal puede ser llamado como filtración Wiener. Por lo tanto, en otras palabras, la cascada de filtros en el dominio del tiempo puede comprender un filtro Wiener.

De acuerdo con realizaciones adicionales, el estimador de ruido de fondo puede estimar una autocorrelación de ruido de fondo como una representación de ruido de fondo de la señal de audio. Además, el reductor de ruido de fondo puede generar la representación de audio de ruido de fondo reducido al restar la autocorrelación del ruido de fondo de una autocorrelación estimada de la señal de audio, en donde la correlación de audio estimada de la señal de audio es la representación de la señal de audio y en donde la representación de la señal de audio de ruido de fondo reducido es una autocorrelación de la señal de audio de ruido de fondo reducido. Utilizando la estimación de funciones de autocorrelación en lugar de utilizar la señal de audio en el dominio del tiempo para calcular los coeficientes LPC y llevar a cabo la reducción de ruido de fondo permite un procesamiento de señal completamente en el dominio del tiempo. Por lo tanto, la autocorrelación de la señal de audio y la autocorrelación del ruido de fondo se pueden calcular por convolución o al utilizar una integral de convolución de un cuadro de audio o una subparte del cuadro de audio. Por lo tanto, la autocorrelación del ruido de fondo se puede llevar a cabo en un cuadro o incluso solo en un subcuadro, que se puede definir como el cuadro o la parte del cuadro donde no está presente ninguna (casi) señal de audio de primer plano tal como voz. Además, la autocorrelación de la señal de audio de ruido de fondo reducido se puede calcular al restar la autocorrelación de ruido de fondo y la autocorrelación de la señal de audio (que comprende ruido de fondo). Utilizando la autocorrelación de la señal de audio de ruido de fondo reducido y la señal de audio (convencionalmente que tiene ruido de fondo) se permite calcular los coeficientes LPC para la señal de audio de ruido de fondo reducido y la señal de audio, respectivamente. Los coeficientes LPC de ruido de fondo reducido pueden ser llamados como el segundo conjunto de coeficientes LPC, donde los coeficientes LPC de la señal de audio pueden ser llamados como el primer conjunto de coeficientes LPC. Por lo tanto, la señal de audio se puede procesar completamente en el dominio del tiempo, ya que la aplicación de la cascada de filtros en el dominio del tiempo también lleva a cabo su filtración en la señal de audio en el dominio del tiempo.

Antes de que se describan realizaciones en detalle utilizando las figuras anexas, se va a señalar que se les da a los

mismos o elementos funcionalmente iguales los mismos números de referencia en las figuras y que se omite una descripción repetida para elementos provistos con los mismos números de referencia. Por lo tanto, las descripciones provistas para elementos que tienen los mismos números de referencia son mutuamente intercambiables.

Las realizaciones de la presente invención se analizarán posteriormente con referencia a las figuras anexas, en donde:

La figura 1 muestra un diagrama de bloques esquemático de un sistema que comprende el codificador para codificar una señal se audio y un decodificador;

La figura 2 muestra un diagrama de bloques esquemático de a) un esquema de codificación de mejora en cascada, b) un esquema de codificación de voz CELP, y c) el esquema conjunto de codificación y mejora;

La figura 3 muestra un diagrama de bloques esquemático de la realización de la figura 2 con una notación diferente;

La figura 4 muestra una gráfica de líneas esquemáticas de la SNR de magnitud perceptual (relación de señal a ruido), como se define en la ecuación 23 para el enfoque conjunto propuesto (J) y el método en cascada (C), en donde la señal de entrada se degradó por ruido de carro no estacionario, y los resultados se presentan para dos velocidades de bits diferentes (7,2 kbit/s indicada por subíndice 7 y 13,2 kbit/s indicada por subíndice 13);

La figura 5 muestra una gráfica de líneas esquemáticas de la SNR de magnitud perceptual, como se define en la ecuación 23 para el enfoque conjunto propuesto (J) y el método en cascada (C), en donde la señal de entrada se degradó por un ruido blanco estacionario, y los resultados se presentan para dos velocidades de bits diferentes (7,2 kbit/s indicada por subíndice 7 y 13,2 kbit/s indicada por subíndice 13);

La figura 6 muestra un gráfico esquemático que muestra una ilustración de las puntuaciones MUSHRA para los diferentes angloparlantes (femenino (F) y masculino (M)), para dos interferencias diferentes (ruido blanco (W) y ruido de carro (C)), para dos SNR de entrada diferentes (10 dB (1) y 20 dB (2)), en donde todos los elementos se codificaron a dos velocidades de bits (7,2 kbit/s (7) y 13,2 kbit/s (13)), para el enfoque conjunto propuesto (JE) y el mejora en cascada (CE), en donde REF fue la referencia oculta, LP el anclaje de paso bajo de 3,5 kHz, y Mix la mezcla distorsionada;

La figura 7 muestra un gráfico de diferentes puntuaciones MUSHRA, simuladas a través de dos velocidades de bits diferentes, comparando el nuevo mejora conjunto (JE) con un enfoque en cascada (CE); y

La figura 8 muestra un diagrama de flujo esquemático de un método para codificar una señal de audio con ruido de fondo reducido utilizando codificación predictiva lineal.

En lo siguiente, se describirán realizaciones de la invención en detalle adicional. Los elementos mostrados en las figuras respectivas que tienen la misma o una funcionalidad similar tienen asociados con los mismos, los mismos signos de referencia.

A continuación, se describirá un método para codificación y mejora conjuntos, con base en filtración Wiener [12] y codificación CELP. Las ventajas de esta función son que 1) la inclusión de filtración Wiener en la cadena de procesamiento no incrementa el retardo algorítmico del códec CELP, y que 2) la optimización conjunta reduce al mínimo simultáneamente distorsión debido a cuantificación y ruido de fondo. Además, la complejidad computacional del esquema conjunto es menor que aquella del enfoque en cascada. La implementación se basa en trabajo reciente a cerca de partición en ventanas residuales en códecs tipo CELP [13, 14, 15], que permite incorporar la filtración Wiener en los filtros del códec CELP de una nueva forma. Con este enfoque se puede demostrar que se mejora tanto la calidad objetiva como subjetiva en comparación con un sistema en cascada.

El método propuesto para codificación y mejora conjuntos de voz, evita de esta forma acumulación de errores debido a procesamiento en cascada y mejora además la calidad de salida perceptual. En otras palabras, el método propuesto evita acumulación de errores debido a procesamiento en cascada, ya que se lleva a cabo una reducción al mínimo conjunta de interferencia y cuantificación por una filtración Wiener óptima en un dominio perceptual.

La figura 1 muestra un diagrama de bloques esquemático de un sistema 2 que comprende un codificador 4 y un decodificador 6. El codificador 4 se configura para codificar una señal de audio 8’ con ruido de fondo reducido utilizando codificación predictiva lineal. Por lo tanto, el codificador 4 puede comprender un estimador de ruido de fondo 10 configurado para estimar una representación de ruido de fondo 12 de la señal de audio 8’. El codificador

puede comprender además un reductor de ruido de fondo 14 configurado para generar una representación de una señal de audio de ruido de fondo reducido 16 al restar la representación del ruido de fondo estimado 12 de la señal de audio 8' de una representación de la señal de audio 8. Por lo tanto, el reductor de ruido de fondo 14 puede recibir la representación del ruido de fondo 12 del estimador de ruido de fondo 10. Una entrada adicional del reductor de ruido de fondo puede ser la señal de audio 8' o la representación de la señal de audio 8. Opcionalmente, el reductor de ruido de fondo puede comprender un generador configurado para generar de forma interna la representación de la señal de audio 8, tal como, por ejemplo una autocorrelación 8 de la señal de audio 8'.

Además, el codificador 4 puede comprender un predictor 18 configurado para someter la representación de la señal de audio 8 a análisis de predicción lineal para obtener un primer conjunto de coeficientes de filtro de predicción lineal (LPC) 20a y para someter la representación de la señal de audio de ruido de fondo reducido 16 a análisis de predicción lineal para obtener un segundo conjunto de coeficientes de filtro de predicción lineal 20b. Similar al reductor de ruido de fondo 14, el predictor 18 puede comprender un generador para generar de forma interna la representación de la señal de audio 8 de la señal de audio 8'. Sin embargo, puede ser ventajoso utilizar un generador común o central 17 para calcular la representación 8 de la señal de audio 8' una vez y proporcionar la representación de la señal de audio, tal como la autocorrelación de la señal de audio 8', al reductor de ruido de fondo 14 y el predictor 18. Por lo tanto, el predictor puede recibir la representación de la señal de audio 8 y la representación de la señal de audio de ruido de fondo reducido 16, por ejemplo, la autocorrelación de la señal de audio y la autocorrelación de la señal de audio de ruido de fondo reducido, respectivamente, y determinar, con base en las señales entrantes, el primer conjunto de coeficientes LPC y el segundo conjunto de coeficientes LPC, respectivamente.

En otras palabras, el primer conjunto de coeficientes LPC se puede determinar de la representación de la señal de audio 8 y el segundo conjunto de coeficientes LPC se puede determinar de la representación de la señal de audio de ruido de fondo reducido 16. El predictor puede llevar a cabo el algoritmo Levinson-Durbin para calcular el primer y segundo conjunto de coeficientes LPC de la autocorrelación respectiva.

Además, el codificador comprende un filtro de análisis 22 compuesto de una cascada 24 de filtros en el dominio del tiempo 24a, 24b controlados por el primer conjunto obtenido de coeficientes LPC 20a y el segundo conjunto obtenido de coeficientes LPC 20b. El filtro de análisis puede aplicar la cascada de filtros en el dominio del tiempo, en donde los coeficientes de filtro del primer filtro en el dominio del tiempo 24a son el primer conjunto de coeficientes LPC y los coeficientes de filtro del segundo filtro en el dominio del tiempo 24b son el segundo conjunto de coeficientes LPC a la señal de audio 8' para determinar una señal residual 26. La señal residual puede comprender los componentes de señal de la señal de audio 8' que puede no representarse por un filtro lineal que tiene el primer y/o el segundo conjunto de coeficientes LPC.

De acuerdo con realizaciones, la señal residual se puede proporcionar a un cuantificador 28 configurado para cuantificar y/o codificar la señal residual y/o el segundo conjunto de coeficientes LPC 24b antes de la transmisión. El cuantificador, por ejemplo puede llevar a cabo excitación codificada por transformada (TCX), predicción lineal excitada por código (CELP), o una codificación sin pérdidas tal como, por ejemplo, codificación de entropía.

De acuerdo con una realización adicional, la codificación de la señal residual se puede llevar a cabo en un transmisor 30 como una alternativa a la codificación en el cuantificador 28. Por lo tanto, el transmisor, por ejemplo, lleva a cabo excitación codificada por transformada (TCX), predicción lineal excitada por código (CELP), o una codificación sin pérdidas tal como, por ejemplo codificación de entropía para codificar la señal residual. Además, el transmisor se puede configurar para transmitir el segundo conjunto de coeficientes LPC. Un receptor opcional es el decodificador 6. Por lo tanto, el transmisor 30 puede recibir la señal residual 26 o la señal residual cuantificada 26'. De acuerdo con una realización, el transmisor puede codificar la señal residual o la señal residual cuantificada, al menos si la señal residual cuantificada no se ha codificado ya, en el cuantificador. Después de codificación opcional de la señal residual o alternativamente la señal residual cuantificada, la señal respectiva provista al transmisor se transmite como una señal residual codificada 32 o como una señal residual cuantificada y codificada 32'. Además, el transmisor puede recibir el segundo conjunto de coeficientes LPC 20b', opcionalmente codificar los mismos, por ejemplo, con el mismo método de codificación como se utiliza para codificar la señal residual, y transmitir además el segundo conjunto codificado de coeficientes LPC 20b', por ejemplo al decodificador 6, sin transmitir el primer conjunto de coeficientes LPC. En otras palabras, el primer conjunto de coeficientes LPC 20a no se necesita transmitir.

El decodificador 6 puede recibir además la señal residual codificada 32 o alternativamente la señal residual cuantificada, codificada 32' y además de una de las señales residuales 32 o 32', el segundo conjunto codificado de coeficientes LPC 20b'. El decodificador puede decodificar las señales recibidas individuales y proporcionar la señal residual de codificada 26 a un filtro de síntesis. El filtro de síntesis puede ser el inverso de un filtro FIR (respuesta finita al impulso) predictivo, lineal que tiene el segundo conjunto de coeficientes LPC como coeficientes de filtro. En otras palabras, un filtro que tiene el segundo conjunto de coeficientes LPC se invierte para formar el filtro de síntesis del decodificador 6. La salida del filtro de síntesis y por lo tanto la salida del codificador es la señal de audio

decodificada 8''.

De acuerdo con realizaciones, el estimador de ruido de fondo puede estimar una autocorrelación 12 del ruido de fondo de la señal de audio como una representación del ruido de fondo de la señal de audio. Además, el reductor de ruido de fondo puede generar la representación de la señal de audio de ruido de fondo reducido 16 al restar la autocorrelación del ruido de fondo 12 de una autocorrelación de la señal de audio 8, en donde la autocorrelación estimada 8 de la señal de audio es la representación de la señal de audio y en donde la representación de la señal de audio de ruido de fondo reducido 16 es una autocorrelación de la señal de audio de ruido de fondo reducido.

Las figuras 2 y 3 ambas se refieren a la misma realización, sin embargo, utilizan una notación diferente. Por lo tanto, la figura 2 muestra ilustraciones del enfoque en cascada y el enfoque de mejora/codificación conjuntos donde W_N y W_C representan la contaminación con ruido blanco de las señales ruidosas y limpias, respectivamente, y W_N^{-1} y W_C^{-1} sus inversos correspondientes. Sin embargo, la figura 3 muestra ilustraciones del enfoque en cascada y el enfoque de mejora/codificación conjuntos donde A_y y A_s representan los filtros de contaminación de ruido blanco de las señales ruidosas y limpias, respectivamente, y H_y y H_s son filtros de reconstrucción (o síntesis), sus inversos correspondientes.

Tanto la figura 2a como la figura 3a muestran una parte de mejora y una parte de codificación de la cadena de procesamiento de señal llevando a cabo así una codificación y mejora en cascada. La parte de mejora 34 puede operar en el dominio de la frecuencia, donde los bloques 36a y 36b pueden llevar a cabo una conversión de tiempo-frecuencia utilizando, por ejemplo, una MDCT y una conversión de frecuencia-tiempo utilizando, por ejemplo una IMDCT o cualquier otra transformada adecuada para llevar a cabo la conversión de tiempo-frecuencia y frecuencia-tiempo. Los filtros 38 y 40 pueden llevar a cabo una reducción de ruido de fondo de la señal de audio transformada en frecuencia 42. En la presente, aquellas partes de frecuencia del ruido de fondo se pueden filtrar al reducir su impacto en el espectro de frecuencia de la señal de audio 8'. El convertidor de frecuencia-tiempo 36b por lo tanto puede llevar a cabo la transformada inversa del dominio de la frecuencia al dominio del tiempo. Después de que se llevó a cabo la reducción de ruido de fondo en la parte de mejora 34, la parte de codificación 35 puede llevar a cabo la codificación de la señal de audio con ruido de fondo reducido. Por lo tanto, el filtro de análisis 22' calcula una señal residual 26'' utilizando coeficientes LPC apropiados. La señal residual se puede cuantificar y proporcionar al filtro de síntesis 44, que es en el caso de la figura 2a y la figura 3a el inverso del filtro de análisis 22'. Ya que el filtro de síntesis 42 es el inverso del filtro de análisis 22', en el caso de la figura 2a y la figura 3a, los coeficientes LPC utilizados para determinar la señal residual 26 se transmiten al decodificador para determinar la señal de audio de codificada 8''.

Las figuras 2b y 3b muestran la etapa de codificación 35 sin la reducción de ruido de fondo previamente llevada a cabo. Ya que la etapa de codificación 35 ya se describe con respecto a las figuras 2a y 3a, se omite una descripción adicional para evitar repetir simplemente la descripción.

La figura 2c y la figura 3c se refieren al concepto principal de codificación-mejora conjuntos. Se muestra que el filtro de análisis 22 comprende una cascada de filtros en el dominio del tiempo utilizando filtros A_y y H_s. De forma más precisa, la cascada de filtros en el dominio del tiempo comprende dos veces un filtro de predicción lineal que utiliza el primer conjunto obtenido de coeficientes LPC 20a ($\mathbf{A}_y^2$) y una vez un inverso de un filtro de predicción lineal adicional que utiliza el segundo conjunto obtenido de coeficientes LPC 20b (H_s). Este arreglo de filtros o esta estructura de filtro puede ser llamada como un filtro Wiener. Sin embargo, se debe señalar que un filtro de predicción H_s se cancela con el filtro de análisis A_s. En otras palabras, también se puede aplicar dos veces el filtro A_y (denotado por $\mathbf{A}_y^2$), dos veces el filtro H_s (denotado por $\mathbf{H}_s^2$) y una vez el filtro A_s.

Como ya se describió con respecto a la figura 1, los coeficientes LPC para estos filtros se determinaron, por ejemplo utilizando autocorrelación. Ya que la autocorrelación se puede llevar a cabo en el dominio del tiempo, no se tiene que llevar a cabo ninguna conversión de tiempo-frecuencia para implementar la codificación y mejora conjuntos. Además, este enfoque es ventajoso ya que la cadena de procesamiento adicional de cuantificación que transmite una filtración de síntesis permanece igual cuando se compara con la etapa de codificación 35 descrita con respecto a las figuras 2a y 3a. Sin embargo, se debe señalar que los coeficientes de filtro LPC basados en la señal de ruido de fondo reducida se deben transmitir al decodificador para filtración de síntesis apropiada. Sin embargo, de acuerdo con una realización adicional, en lugar de transmitir los coeficientes LPC, se pueden transmitir coeficientes de filtro ya calculados del filtro 24b (representados por el inverso de los coeficientes de filtro 20b) para evitar una inversión adicional del filtro lineal que tiene los coeficientes LPC para obtener el filtro de síntesis 42, ya que esta inversión ya sea ha llevado a cabo en el codificador. En otras palabras, en lugar de transmitir los coeficientes de filtro 20b, el inverso de la matriz de estos coeficientes de filtro se puede transmitir, evitando así llevar a cabo dos veces la inversión. Además, se tiene que señalar que el filtro del lado de codificador 24b y el filtro de síntesis 42 pueden ser el mismo filtro, aplicado en el codificador y decodificador respectivamente.

En otras palabras, con respecto a la figura 2, los códecs de voz basados en el modelo CELP se basan en un modelo

de producción de voz que asume que la correlación de la señal de voz de entrada s_n se puede modelar por un filtro de producción lineal con coeficientes $\mathbf{a} = [\alpha_0, \alpha_1, \ldots, \alpha_M]^T$ donde *M* es el orden de modelo [16]. El residuo $r_n = a_n * s_n$, que es la parte de la señal de voz que no se puede predecir por el filtro de predicción lineal entonces se cuantifica utilizando cuantificación vectorial.

Permítase que $\mathbf{s}_k = [s_k, s_{k-1}, \ldots, s_{k-M}]^T$ sea un vector de la señal de entrada donde el superíndice ^{T} denota la transpuesta. El residuo entonces se puede expresar como

$$r_k = \mathbf{a}^T \mathbf{s}_k. \quad (1)$$

Dada la matriz de autocorrelación $\boldsymbol{R}_{ss}$ del vector de señal de voz $\boldsymbol{s}_k$

$$\mathbf{R}_{ss} = E\{\mathbf{s}_k \mathbf{s}_k^{\mathrm{T}}\}, \quad (2)$$

una estimación del filtro de producción de orden M puede estar dada como [20]

$$\mathbf{a} = \sigma_e^2 \mathbf{R}_{ss}^{-1} \mathbf{u}, \quad (3)$$

donde $\mathbf{u} = [1, 0, 0, \ldots, 0]^T$ y el error de predicción escalar σ_e^2 elegidos de tal forma que $\alpha_0 = 1$. Obsérvese que el filtro de predicción lineal α_n es un filtro de contaminación de ruido blanco, donde r_k es ruido blanco sin correlacionar. Además, la señal original s_n se puede reconstruir de la señal residual r_n a través de filtración IIR con el predictor α_n. El siguiente paso es cuantificar vectores del residuo $\mathbf{r}_k = [r_{kN}, r_{kN-1}, \ldots, r_{kN-N+1}]^T$ con un cuantificador de vector a $\tilde{\mathbf{r}}_k$, de tal forma que se reduce al mínimo la distorsión perceptual. Permítase que un vector de la señal de salida sea $\mathbf{s}'_k = [s_{kN}, s_{kN-1}, \ldots, s_{k-N+1}]^T$ y $\tilde{\mathbf{s}}'_k$ su contraparte cuantificada, y *W* una matriz de convolución que se aplica a ponderación perceptual en la salida. El problema de optimización perceptual entonces se puede escribir como

$$\min_{\mathbf{r}_k} \|\mathbf{W}(\mathbf{s}'_k - \tilde{\mathbf{s}}'_k)\|^2 = \min_{\mathbf{r}_k} \|\mathbf{WH}(\mathbf{r}_k - \tilde{\mathbf{r}}_k)\|^2, \quad (4)$$

donde **H** es una matriz de convolución que corresponde a la respuesta de impulso del predictor α_n.

El proceso de codificación de voz tipo CELP se representa en la figura 2b. La señal de entrada primero se blanquea con el filtro $A(z) = \sum_{m=0}^{M} \alpha_m z^{-m}$ para obtener la señal residual. Los vectores del residuo entonces se cuantifican en el bloque Q. Finalmente, la estructura de envolvente espectral entonces se reconstruye por filtración IIR que es el doble IR, de $A^{-1}(z)$ para obtener la señal de salida cuantificada $\tilde{s}_k$. Ya que la señal resintetiza se evalúa en el dominio perceptual, este enfoque se conoce como el método de análisis por síntesis.

Filtración Wiener

En mejora de voz de un solo canal, se asume que la señal y_n se adquiere, que es una mezcla aditiva de la señal de voz limpia, deseada s_n y alguna interferencia indeseada v_n, es decir

$$y_n = s_n + v_n. \quad (5)$$

El objetivo del proceso de mejora es estimar la señal de voz limpia s_n, en tanto que es accesible solo a la señal ruidosa y_n y las estimaciones de las matrices de correlación

$$\mathbf{R}_{ss} = E\{\mathbf{s}_k \mathbf{s}_k^T\} \text{ y } \mathbf{R}_{yy} = E\{\mathbf{y}_k \mathbf{y}_k^T\} \quad (6)$$

Donde $\mathbf{y}_k = [y_k, y_{k-1}, \ldots, y_{k-M}]^T$. utilizando una matriz de filtro H, la estimación de la señal de voz limpia $\hat{\mathbf{s}}_n$ se define como

$$\hat{\mathrm{s}}_k = \mathbf{H} \mathrm{y}_k. \quad (7)$$

El filtro óptimo en el sentido de error cuadrático medio, mínimo (MMSE), conocido como el filtro Wiener se puede

obtener fácilmente como [12]

$$\mathbf{H} = \mathbf{R}_{ss}\mathbf{R}_{yy}^{-1}. \quad (8)$$

Por lo general, se aplica filtración Wiener en ventanas superpuestas de la señal de entrada y se reconstruye utilizando el método de superposición y suma [21, 12]. Este enfoque se ilustra en el bloque de mejora de la figura 2a. Sin embargo, conduce a un incremento en retardo algorítmico, que corresponde a la longitud de la superposición entre ventanas. Para evitar este retardo, un objetivo es combinar filtración Wiener con un método basado en predicción lineal.

Para obtener tal conexión, la señal de voz estimada $\hat{\mathbf{s}}_k$ se sustituye en la ecuación 1, por lo cual

$$\begin{aligned} r_k &= \mathbf{a}^T\hat{\mathbf{s}}_k = \mathbf{a}^T\mathbf{H}\mathbf{y}_k = \sigma_e^2\mathbf{u}^T\mathbf{R}_{ss}^{-1}\mathbf{R}_{ss}\mathbf{R}_{yy}^{-1}\mathbf{y}_k \\ &= \sigma_e^2\mathbf{u}^T\mathbf{R}_{yy}^{-1}\mathbf{y}_k = \gamma\mathbf{a}'^T\mathbf{y}_k \end{aligned} \quad (9)$$

donde γ es un coeficiente de modificación de escala y

$$\mathbf{a}' = \hat{\sigma}_e^2\mathbf{R}_{yy}^{-1}\mathbf{u} \quad (10)$$

es el predictor óptimo para la señal ruidosa y_n. En otras palabras, la filtración de la señal ruidosa con a', se obtiene el residuo (modificado en escala) de la señal limpia estimada. La modificación en escala es la relación entre la relación entre los errores residuales esperados de las señales limpia y ruidosa, σ_e^2 y $\hat{\sigma}_e^2$, respectivamente, es decir $\gamma = \sigma_e^2/\hat{\sigma}_e^2$. Esta derivación por lo tanto muestra que la filtración Wiener y la predicción lineal son métodos íntimamente relacionados y en la siguiente sección, esta conexión se utilizará para desarrollar un método conjunto de mejora y codificación.

Incorporación de filtración Wiener en un códec CELP

Un objetivo es combinar filtración Wiener y códecs CELP (descritos en la sección 2 y la sección 2) en un algoritmo conjunto. Al combinar estos algoritmos se puede evitar al retardo de la partición en ventanas de superposición y suma por implementaciones usuales de filtración Wiener, y se reduce la complejidad computacional.

La implementación de la estructura conjunta entonces es sencilla. Se muestra que el residuo de la señal de voz mejorada se puede obtener por la ecuación 9. La señal de voz mejorada por lo tanto se puede reconstruir por filtración IIR del residuo con el modelo predictivo lineal α_n de la señal limpia.

Para cuantificación del residuo, la ecuación 4 se puede modificar al reemplazar la señal limpia $\mathbf{s}_k'$ con la señal estimada $\tilde{\mathbf{s}}_k'$ para obtener

$$\min_{\mathbf{r}_k}\left\|\mathbf{W}(\tilde{\mathbf{s}_k}' - \tilde{\mathbf{s}}_k')\right\|^2 = \min_{\mathbf{r}_k}\|\mathbf{W}\mathbf{H}(\mathbf{r}_k - \tilde{\mathbf{r}}_k)\|^2. \quad (11)$$

En otras palabras, la función objetivo con la señal objetivo mejorada $\tilde{\mathbf{s}}_k'$ permanece igual como si tuviera acceso a la señal de entrada limpia $\mathbf{s}_k'$.

En conclusión, la única modificación a CELP normal es reemplazar el filtro de análisis **a** de la señal limpia con aquel de la señal ruidosa a'. Las partes restantes del algoritmo CELP permanecen sin cambios. El enfoque propuesto se ilustra en la figura 2(c).

Es claro que el método propuesto se pueda aplicar en cualquier códec CELP con cambios mínimos siempre que se desee atenuación de ruido y cuando tenga acceso a una estimación de la autocorrelación de la señal de voz limpia R_{ss}. Si no está disponible una estimación de la autocorrelación de señal de voz limpia, se puede estimar utilizando una estimación de la autocorrelación de la señal de ruido R_w, por $R_{ss} \approx R_{yy} - R_w$ u otras estimaciones comunes.

El método se puede extender fácilmente a escenarios tal como algoritmos multicanal con haz modelado, siempre y

cuando una estimación de la señal limpia sea alcanzable utilizando filtros en el dominio del tiempo.

La ventaja en complejidad computacional del método propuesto se puede caracterizar como sigue. Se señala que en el enfoque convencional se necesita determinar el filtro de matriz **H**, dado por la ecuación 8. La inversión de matriz requerida es de complejidad $\mathcal{O}(M^3)$. Sin embargo, en el enfoque propuesto solo la ecuación 3 se tiene que resolver para la señal ruidosa, que se puede implementar con el algoritmo Levinson-Durbin (o similar) con complejidad $\mathcal{O}(N^2)$.

Predicción lineal excitada por código

En otras palabras con respecto a la figura 3, los códecs de voz basados en el paradigma CELP utilizan un modelo de producción de voz que asume que la correlación, y por lo tanto la envolvente espectral de la señal de voz de entrada s_n se puede modelar por un filtro de predicción lineal con coeficientes $\mathbf{a} = [\alpha_0, \alpha_1, \ldots, \alpha_M]^T$ donde M es el orden de modelo, determinado por el modelo de tubo subyacente [16]. El residuo $r_n = a_n * s_n$, la parte de la señal de voz que no se puede predecir por el filtro de predicción lineal (también llamado como predictor 18), entonces se cuantifica utilizando cuantificación vectorial.

El filtro predictivo lineal $\mathbf{a}_s$, para un cuadro de la señal de entrada S se puede obtener, reduciendo al mínimo

$$\min_{\mathbf{a}_s} \mathcal{E}\left\{\|\mathbf{s}^*\mathbf{a_s}\|^2 - 2\sigma_s^2(\mathbf{u}^*\mathbf{a_s} - 1)\right\}, \quad (12)$$

donde u = [1 0 0 … 0]^T. La solución resulta como:

$$\mathbf{a}_s = \sigma_e^2 \mathbf{R}_{ss}^{-1}\mathbf{u}. \quad (13)$$

Con la definición de la matriz de convolución $\mathbf{A}_s$, que consiste de los coeficientes de filtro α de $\mathbf{a}_s$

$$\mathbf{A}_s = \begin{bmatrix} 1 & 0 & \ldots & & 0 \\ \alpha_1 & \ddots & & & \vdots \\ \alpha_2 & \ddots & 1 & \ddots & \\ \vdots & \ddots & \alpha_1 & 1 & 0 \\ \alpha_M & \ldots & \alpha_2 & \alpha_1 & 1 \end{bmatrix}, \quad (14)$$

la señal residual se puede obtener al multiplicar el cuadro de voz de entrada con la matriz de convolución $\mathbf{A}_s$

$$\mathbf{e}_s = \mathbf{A}_s \cdot \mathbf{s}. \quad (15)$$

La partición en ventanas, se lleva a cabo aquí como en códecs CELP al restar la respuesta de entrada cero de la señal de entrada y reintroduciéndola en la resíntesis [15].

La multiplicación en la ecuación 15 es idéntica a la convolución de la señal de entrada con el filtro de predicción, y por lo tanto corresponde a la filtración FIR. La señal original se puede reconstruir del residuo, por una multiplicación con el filtro de reconstrucción $\mathbf{H}_s$

$$\mathbf{s} = \mathbf{H}_s \cdot \mathbf{e}_s. \quad (16)$$

donde $\mathbf{H}_s$, consiste de la respuesta de impulso $\eta = [1, \eta_1, \ldots, \eta_{N-1}]$ del filtro de predicción

$$\mathbf{H}_s = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \eta_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \eta_{N-1} & \dots & \eta_1 & 1 \\ \vdots & & \vdots & \vdots \end{bmatrix} \quad (17)$$

de tal forma que esta operación corresponde a filtración IIR.

El vector residual se cuantifica aplicando cuantificación vectorial. Por lo tanto, el vector cuantificado $\hat{\mathbf{e}}_s$ se elige, reduciendo al mínimo la distancia perceptual, en el sentido de la normal-2, a la señal limpia reconstruida, deseada:

$$\min_{\hat{\mathbf{e}}_s} \|\mathbf{WH}(\hat{\mathbf{e}}_s - \mathbf{e}_s)\|^2 , \quad (18)$$

donde $\mathbf{e}_s$ es el residuo no cuantificado y **W**(z) = A(0.92z) es el filtro de ponderación perceptual, como se utiliza en el códec de voz AMR-WB [6].

Aplicación de filtración Wiener en un códec CELP

Para la aplicación de mejora de voz de un solo canal, asumiendo que la señal de micrófono adquirida y_n, es una mezcla aditiva de la señal de voz limpia, deseada s_n y alguna interferencia indeseada v_n, de tal forma que $y_n = s_n + v_n$. En el dominio Z, equivalentemente Y(z) = S(z) + V(z).

Al aplicar un filtro Wiener B(z) es posible reconstruir la señal de voz S(z) de la observación ruidosa Y(z) por filtración, de tal forma que la señal de voz estimada es $\hat{S}(z)$:= B(z)Y(z) ≈ S(z). La solución cuadrática media, mínima para el filtro Wiener resulta como [12]

$$B(z) = \frac{|S(z)|^2}{|S(z)|^2 + |V(z)|^2} , \quad (19)$$

dada la suposición que las señales de voz y ruido s_n y v_n, respectivamente, no están correlacionadas.

En un códec de voz, una estimación del espectro de potencia está disponible de la señal ruidosa y_n, en la forma de la respuesta de impulso del modelo predictivo lineal $|A_y(z)|^{-2}$. En otras palabras, $|S(z)|^2 + |V(z)|^2 \approx \gamma\, |A_y(z)|^{-2}$ donde γ es un coeficiente de modificación de escala. El predictor lineal ruidoso se puede calcular de la matriz de autocorrelación $\mathbf{R}_{yy}$ de la señal ruidosa de forma general.

Además, se puede estimar el espectro de potencia de la señal de voz limpia $|S(z)|^2$ o de forma equivalente, la matriz de autocorrelación $\mathbf{R}_{ss}$ de la señal de voz limpia. Los algoritmos de mejora a menudo asumen que la señal de ruido es estacionaria, por lo cual la autocorrelación de la señal de ruido como $\mathbf{R}_w$ se puede estimar de un cuadro no de voz de la señal de entrada. La matriz de autocorrelación de la señal de voz limpia $\mathbf{R}_{ss}$ entonces se puede estimar como $\widehat{\mathbf{R}}_{ss} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$. Aquí es ventajoso tomar las precauciones usuales para asegurar que $\widehat{\mathbf{R}}_{ss}$ permanezca definida, positiva.

Utilizando la matriz de autocorrelación estimada para voz limpia $\widehat{\mathbf{R}}_{ss}$, el predictor lineal correspondiente se puede determinar, cuya respuesta de impulso en el dominio Z es $\hat{A}_S^{-1}(z)$. Por lo tanto, $|S(z)|^2 \approx |\hat{A}_S(z)|^{-2}$ y la ecuación 19 se puede describir como

$$B(z) \approx \frac{|\hat{A}_s(z)|^{-2}}{|A_y(z)|^{-2}} = \frac{|A_y(z)|^2}{|\hat{A}_s(z)|^2} . \quad (20)$$

En otras palabras, por filtración dos veces con los predictores de las señales ruidosa y limpia, en modo FIR e IIR respectivamente, se puede obtener una estimación Wiener de la señal limpia.

Las matrices de convolución se pueden denotar correspondientes a filtración FIR con predictores $\hat{A}_s(z)$ y $A_y(z)$ por $\mathbf{A}_s$ y $\mathbf{A}_y$, respectivamente. De forma similar, permítase que $\mathbf{H}_s$ y $\mathbf{H}_y$ sean las matrices de convolución respectivas que corresponden a filtración predictiva (IIR). Utilizando estas matrices, se puede ilustrar codificación CELP convencional con un diagrama de fijo con en la figura 3b. Aquí, es posible filtrar la señal de entrada s_n con $\mathbf{A}_s$ para obtener el residuo, cuantificarlo y reconstruir la señal cuantificada por filtración con $\mathbf{H}_s$.

El enfoque convencional para combinar mejora con codificación se ilustra en la figura 3a, donde se aplica filtración Wiener como un bloque de preprocesamiento antes de codificación.

Finalmente, en el enfoque propuesto se combina filtración Wiener con códecs de voz tipo CELP. Comparando el enfoque en cascada de la figura 3a con el enfoque conjunto, ilustrado en la figura 3b, es evidente que se puede omitir el esquema de partición en ventanas de suma y superposición (OLA) adicional. Además, el filtro de entrada $\mathbf{A}_s$ en el codificador se cancela con $\mathbf{H}_s$. Por lo tanto, como se muestra en la figura 3c, la señal residual limpia, estimada $\tilde{\mathbf{e}} = \mathbf{A}_y^2\mathbf{H}_s\mathbf{y}$ resulta por filtración la señal de entrada deteriorada **y** con la combinación de filtro $\mathbf{A}_y^2\mathbf{H}_s$. Por lo tanto, la reducción al mínimo de error resulta:

$$\min_{\hat{\mathbf{e}}} \|\mathbf{W}\mathbf{H}_s(\hat{\mathbf{e}} - \tilde{\mathbf{e}})\|^2 . \qquad (21)$$

Por lo tanto, este enfoque reduce al mínimo conjuntamente la distancia entre la estimación limpia y la señal cuantificada, por lo cual es factible una reducción al mínimo conjunta de la interferencia y el ruido de cuantificación en el dominio perceptual.

El desempeño del enfoque conjunto de mejora y codificación de voz se evaluó utilizando tanto mediciones objetivas como subjetivas. Al fin de aislar el desempeño del nuevo método, se utiliza un códec CELP simplificado, donde solo se cuantificó la señal residual, pero no se cuantificó el retardo y la ganancia de la predicción de largo plazo (LTP), la codificación predictiva lineal (LPC) y los factores de ganancia. El residuo se cuantificó utilizando un método iterativo en pares, donde dos pulsos se suman de forma consecutiva al tratarlos en cada posición, como se describe en [17]. Además, para evitar cualquier influencia de algoritmos de estimación, la matriz de correlación de la señal de voz limpia $\mathbf{R}_{ss}$ se asumió que se conocía en todos los escenarios simulados. Con la suposición de que no está correlacionada la señal de voz y ruido, se sostiene que $\mathbf{R}_{ss} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$. En cualquier aplicación práctica la matriz de correlación de ruido $\mathbf{R}_w$ o alternativamente la matriz de correlación de voz limpia $\mathbf{R}_{ss}$ se tiene que estimar de la señal de micrófono adquirida. Un enfoque común es estimar la matriz de correlación de ruido en pausas de voz, asumiendo que la interferencia es estacionara.

El escenario evaluado consistió de una mezcla de la señal de voz limpia, deseada e interferencia aditiva. Se han considerado dos tipos de interferencias: ruido blanco estacionario y un segmento de una grabación de ruido de carro de la librería de mezclas de sonidos Civilisation [18]. Se llevó a cabo cuantificación vectorial del residuo con una velocidad de bits de 2,8 kbit/s y 7,2 kbit/s, que corresponde a una velocidad de bits total de 7,2 kbit/s y 13,2 kbit/s respectivamente para un códec AMR-WB [6]. Se utilizó una velocidad de muestreo de 12,8 kHz para todas las simulaciones.

Se evaluaron las señales mejorada y codificada utilizando tanto mediciones objetivas como subjetivas, por lo tanto, se llevó a cabo una prueba de escucha y se calculó una relación de señal a ruido (SNR) de magnitud perceptual, como se define en la ecuación 23 y la ecuación 22. Esta SNR de magnitud perceptual se utilizó ya que el proceso conjunto de mejora no tiene ninguna influencia en la fase de los filtros, ya que tanto los filtros de síntesis como de reconstrucción están sujetos a la limitación de filtros de fase mínima, según el diseño de filtros de predicción.

Con la definición de la transformada de Fourier como operador $\mathcal{F}(\cdot)$, los valores espectrales absolutos de la señal de referencia limpia, reconstruida y la señal limpia estimada en el dominio perceptual resultan como:

$$S = |\mathcal{F}(\mathbf{W}\mathbf{H}_s\mathbf{e}_k)| \quad \text{y} \quad \hat{S} = |\mathcal{F}(\mathbf{W}\mathbf{H}_s\hat{\mathbf{e}}_k)| . \qquad (22)$$

La definición de la relación de señal a ruido perceptual (PSNR) modificada resulta como:

$$\mathrm{PSNR}_{\mathrm{ABS}} = 10\log_{10}\frac{\|S\|^2}{\|\hat{S} - S\|^2} . \qquad (23)$$

Para la evaluación subjetiva, se utilizaron elementos de voz del conjunto de prueba utilizado para la normalización de USAC [8], contaminados por ruido blanco y de coche, como se describe anteriormente. Se llevó a cabo una prueba de escucha de estimulación múltiple con anclaje y referencia ocultos (MUSHRA) [19] con 14 participantes, utilizando auriculares electrostáticos STAX en un entorno insonoro. Los resultados de la prueba de escucha se ilustran en la figura 6 y las puntuaciones MUSHRA diferenciales en la figura 7, que muestran la media y los intervalos de confianza de 95 %.

Los resultados de prueba MUSHRA absolutos en la figura 6 muestran que la referencia oculta siempre se asignó correctamente a 100 puntos. La mezcla ruidosa original recibió la puntuación media más baja para cada elemento, que indica que todos los métodos de mejora mejoraron la calidad perceptual. Las puntuaciones medias para la velocidad de bits más baja mostraron una mejora estadísticamente significativa de 6.4 puntos MUSHRA para el promedio a través de todos los elementos en comparación con el enfoque en cascada. Para la velocidad de bits mayor, el promedio a través de todos los elementos mostró una mejora, que, sin embargo, no es estadísticamente significativa.

Para obtener una comparación más detallada del método conjunto y el método mejorado previamente, se presentan las puntuaciones MUSHRA diferenciales en la figura 7, donde la diferencia entre el método mejorado previamente y el método conjunto se calcula para cada oyente y elemento. Los resultados diferenciales verifican las puntuaciones MUSHRA absolutas, al mostrar una mejora estadísticamente significativa para la velocidad de bits inferior, mientras que la mejora para la velocidad de bits mayor no es estadísticamente significativa.

En otras palabras, se muestra un método para codificación y mejora conjuntos de voz, que permite reducción al mínimo de ruido de cuantificación e interferencia general. En contraste, los enfoques convencionales aplican mejora y codificación en pasos de procesamiento en cascada. La unión de ambos pasos de procesamiento también es atractiva en términos de complejidad computacional, ya que se pueden omitir operaciones de filtración y partición en ventanas, repetidas.

Los códecs de voz tipo CELP se diseñan para ofrecer un muy bajo retardo y por lo tanto evitan una superposición de ventanas de procesamiento con ventanas de procesamiento futuras. En contraste, los métodos de mejora convencionales, aplicados en el dominio de la frecuencia dependen de partición en ventanas de superposición y suma, que introduce un retardo adicional que corresponde a la duración de superposición. El enfoque conjunto no requiere partición en ventanas de superposición y suma, pero utiliza el esquema de partición en ventanas como se aplica en códecs de voz [15], por lo cual evita el incremento en retardo algorítmico.

Un problema conocido con el método propuesto es que, a diferencia de filtración Wiener espectral, convencional donde la fase de señal se deja intacta, los métodos propuestos aplican filtros en el dominio del tiempo, que modifican la fase. Estas modificaciones de fase se pueden tratar fácilmente por aplicación de filtros all-pass adecuados. Sin embargo, ya que no se ha notado ninguna degradación perceptual atribuida a modificaciones de fase, estos filtros all-pass se omitieron para mantener baja la complejidad computacional. Se señala, sin embargo, que en la evaluación objetiva, se midió SNR de magnitud perceptual, para permitir comparación justa de los métodos. La medición objetiva muestra que el método propuesto es en promedio tres dB mejor que el procesamiento en cascada.

La ventaja de desempeño del método propuesto se confirmó además por resultados de una prueba de escucha MUSHRA, que muestra una mejora promedio de 6,4 puntos. Estos resultados demuestran que la aplicación de codificación y mejora conjuntos es benéfica para el sistema general en términos tanto de calidad como de complejidad computacional, en tanto que se mantiene el bajo retardo algorítmico de códecs de voz CELP.

La figura 8 muestra un diagrama de bloques esquemático de un método 800 para codificar una señal de audio con ruido de fondo reducido utilizando codificación predictiva lineal. El método 800 comprende un paso S802 de estimación de una representación de ruido de fondo de la señal de audio, un paso S804 de generación de un representación de una señal de audio de ruido de fondo reducido al restar la representación del ruido de fondo estimado de la señal de audio de una representación de la señal de audio, un paso S806 de sometimiento de la representación de la señal de audio a análisis de predicción lineal para obtener un primer conjunto de coeficientes de filtro de predicción lineal y para someter la representación de la señal de audio de ruido de fondo reducido a análisis de predicción lineal para obtener un segundo conjunto de coeficientes de filtro de predicción lineal, y un paso S808 de control de una cascada de filtros en el dominio del tiempo por el primer conjunto obtenido de coeficientes LPC y el segundo conjunto obtenido de coeficientes LPC para obtener una señal residual de la señal de audio.

Debe entenderse que, en esta especificación, las señales en líneas son algunas veces nombradas por números de referencia por las líneas o indicadas algunas veces por los mismos números de referencia, que se han atribuido a las líneas. Por lo tanto, la notación es de tal forma que una línea que tiene una cierta señal indica la propia señal. Una línea puede ser una línea física en una implementación alámbrica. En una implementación computarizada, sin

embargo, no existe una línea física, pero la señal representada por la línea se transmite de un módulo de cálculo a otro módulo de cálculo.

Aunque la presente invención se ha descrito en el contexto de diagramas de bloques donde los bloques representan componentes de hardware lógicos o reales, la presente invención también se puede implementar por un método implementado por ordenador. En este último caso, los bloques representan pasos de método correspondientes donde estos pasos representan las funcionalidades llevadas a cabo por bloques de hardware, físicos o lógicos correspondientes.

Aunque algunos aspectos se han descrito en el contexto de un aparato, es claro que estos aspectos también representan una descripción del método correspondiente, donde un bloque o dispositivo corresponde a un paso de método o una característica de un paso de método. De forma análoga, los aspectos descritos en el contexto de un paso de método también representan una descripción de un bloque correspondiente o elemento o características de un aparato correspondiente. Algunos o todos los pasos de método se pueden ejecutar por (o utilizando) un aparato de hardware, como, por ejemplo, un microprocesador, un ordenador programable o un circuito electrónico. En algunas realizaciones, alguno o más de los pasos de método más importantes se pueden ejecutar por este aparato.

La señal transmitida o codificada inventiva se puede almacenar en un medio de almacenamiento digital o se puede transmitir en un medio de transmisión tal como un medio de transmisión inalámbrico o un medio de transmisión alámbrico tal como Internet.

Dependiendo de ciertos requerimientos de implementación, realizaciones de la invención se pueden implementar en hardware o en software. La implementación se puede llevar a cabo utilizando un medio de almacenamiento digital, por ejemplo un disco flexible, un DVD, un Blu-Ray, un CD, una ROM, una PROM, una EPROM, una EEPROM o una memoria flash, que tiene señales de control leíbles electrónicamente almacenadas en el mismo, que cooperan (o son capaces de cooperar) con un sistema informático programable de tal forma que se lleva a cabo el método respectivo. Por lo tanto, el medio de almacenamiento digital puede ser leíble por ordenador.

Algunas realizaciones de acuerdo con la invención comprenden un portador de datos que tiene señales de control electrónicamente leíbles, que son capaces de cooperar con un sistema informático programable, de tal forma que se lleva a cabo uno de los métodos descritos en el presente documento.

En general, las realizaciones de la presente invención se pueden implementar como un producto de programa informático con un código de programa, el código de programa que es operativo para llevar a cabo uno de los métodos cuando el producto de programa informático se ejecuta en un ordenador. El código de programa puede, por ejemplo, almacenarse en un portador leíble por máquina.

Otras realizaciones comprenden el programa informático para llevar a cabo uno de los métodos descritos en la presente, almacenados en un portador leíble por máquina.

En otras palabras, una realización el método inventivo es, por lo tanto, un programa informático que tiene un código de programa para llevar a cabo uno de los métodos descritos en la presente, cuando el programa informático se ejecuta en un ordenador.

Una realización adicional del método inventivo es, por lo tanto, un portador de datos (o un medio de almacenamiento no transitorio tal como un medio de almacenamiento digital, o un medio leíble por ordenador) que comprende, grabado en el mismo, el programa informático para llevar a cabo uno de los métodos descritos en la presente. El portador de datos, el medio de almacenamiento digital o el medio grabado son convencionalmente tangibles y/o no transitorios.

Una realización adicional del método de invención es, por lo tanto, un flujo de datos o una secuencia de señales que representan el programa informático para llevar a cabo uno de los métodos descritos en la presente. El flujo de datos o la secuencia de señales pueden por ejemplo, configurarse para transferirse a través de una conexión de comunicación de datos, por ejemplo, a través de internet.

Una realización adicional comprende un medio de procesamiento, por ejemplo, un ordenador o un dispositivo lógico programable, configurado para, o adaptado para, llevar a cabo uno de los métodos descritos en la presente.

Una realización adicional comprende un ordenador que tiene instalada en la misma el programa informático para llevar a cabo uno de los métodos descritos en la presente.

Una realización adicional de acuerdo con la invención comprende un aparato o un sistema configurado para transferir (por ejemplo, de forma electrónica u óptica) un programa de informático para llevar a cabo uno de los métodos descritos en la presente a un receptor. El receptor, por ejemplo, puede ser un ordenador, un dispositivo

móvil, un dispositivo de memoria o similares. El aparato o sistema puede, por ejemplo, comprender un servidor de archivos para transferir el programa informático al receptor.

En algunas realizaciones, se puede utilizar un dispositivo lógico programable (por ejemplo, un arreglo de compuertas programable en el campo) para llevar a cabo algunas o todas las funcionalidades de los métodos descritos en la presente. En algunas realizaciones, un arreglo de compuertas programables en el campo puede cooperar con un microprocesador a fin de llevar a cabo uno de los métodos descritos en la presente. En general, los métodos se llevan a cabo de manera preferente por cualquier aparato de hardware.

Las realizaciones descritas anteriormente son simplemente ilustrativas para los principios de la presente invención. Se entiende que serán evidentes modificaciones y variaciones de los arreglos y los detalles descritos en la presente para otros expertos en la técnica. Se propone, por lo tanto, que se limite solo por el alcance de las próximas reivindicaciones de patente y no por los detalles específicos presentados a manera de descripción y explicación de las realizaciones en la presente.

Referencias

[1] M. Jeub y P. Vary, “Enhancement of reverberant speech using the CELP postfilter,” in Proc. ICASSP, Abril 2009, pp. 3993–3996.

[2] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, y P. Vary, “Noise reduction for dual-microphone mobile phones exploiting power level differences,” in Proc. ICASSP, Marzo 2012, pp. 1693–1696.

[3] R. Martin, I. Wittke, y P. Jax, “Optimized estimation of spectral parameters for the coding of noisy speech,” in Proc. ICASSP, vol. 3, 2000, pp. 1479–1482 vol.3.

[4] H. Taddei, C. Beaugeant, y M. de Meuleneire, “Noise reduction on speech codec parameters,” in Proc. ICASSP, vol. 1, Mayo 2004, pp. I–497–500 vol.1.

[5] 3GPP, “Mandatory speech CODEC speech processing functions; AMR speech Codec; General description,” 3rd Generation Partnership Project (3GPP), TS 26.071, 12 2009. [En línea]. Disponible: http://www.3gpp.org/ftp/Specs/html-info/26071.htm

[6] -, “Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions,” 3rd Generation Partnership Project (3GPP), TS 26.190, 12 2009. [En línea]. Disponible: http://www.3gpp.org/ftp/Specs/html-info/26190.htm

[7] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, y K. Jarvinen, “The adaptive multirate wideband speech codec (AMR-WB),” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 620–636, Noviembre 2002.

[8] ISO/IEC 23003–3:2012, “MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding,” 2012.

[9] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, y B. Grill, “Unified speech and audio coding scheme for high quality at low bitrates,” in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, Abril 2009, pp. 1–4.

[10] 3GPP, “TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12),” 3rd Generation Partnership Project (3GPP), TS 26.445, 12 2014. [En línea]. Disponible: http://www.3gpp.org/ftp/Specs/html-info/26445.htm

[11] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z.Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, H. Sung, E. Oh, H. Yuan, y C. Zhu, “Overview of the EVS codec architecture,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, Abril 2015, pp. 5698–5702.

[12] J. Benesty, M. Sondhi, y. Huang, Springer Handbook of Speech Processing. Springer, 2008.

[13] T. Bäckström, “Computationally efficient objective function for algebraic codebook optimization in ACELP,” in Proc. Interspeech, Agosto 2013.

[14] -, “Comparison of windowing in speech and audio coding,” in Proc. WASPAA, New Paltz, USA, Octubre

2013.

[15] J. Fischer y T. Bäckström, “Comparison of windowing schemes for speech coding,” in Proc EUSIPCO, 2015.

[16] M. Schroeder y B. Atal, “Code-excited linear prediction (CELP): High-quality speech at very low bit rates,” in Proc. ICASSP. IEEE, 1985, pp. 937–940.

[17] T. Bäckström y C. R. Helmrich, “Decorrelated innovative codebooks for ACELP using factorization of autocorrelation matrix,” in Proc. Interspeech, 2014, pp. 2794–2798.

[18] soundeffects.ch, “Civilisation soundscapes library,” visitado: 23.09.2015. [En línea]. Disponible: https://www.soundeffects.ch/de/geraeusch-archive/soundeffects.ch- produkte/civilisation-soundscapes-d.php

[19] Method for the subjective assessment of intermediate quality levels of coding systems, ITU-R Recommendation BS.1534, 2003. [En línea]. Disponible: http://www.itu.int/rec/R-REC-BS.1534/en.

[20] P. P. Vaidyanathan, \The theory of linear prediction," in Synthesis Lectures on Signal Processing, vol. 2, pp. 1{184. Morgan & Claypool publishers, 2007.

[21] J. Allen, \Short-term spectral analysis, and modification by discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Process., vol. 25, pp. 235{238, 1977.

REIVINDICACIONES

1. Codificador (4) para codificar una señal de audio (8') con ruido de fondo reducido utilizando codificación predictiva lineal, comprendiendo el codificador (4):

un estimador de ruido de fondo (10) configurado para estimar una autocorrelación del ruido de fondo como una representación del ruido de fondo (12) de la señal de audio (8');
un reductor de ruido de fondo (14) configurado para generar una representación de una señal de audio de ruido de fondo reducido (16) restando la autocorrelación del ruido de fondo (12) de la señal de audio (8') desde una autocorrelación de la señal de audio (8) para que la representación de la señal de audio de ruido de fondo reducido (16) sea una autocorrelación de una señal de audio de ruido de fondo de fondo;
un predictor (18) configurado para someter la representación de la señal de audio (8) al análisis de predicción lineal para obtener un primer juego de coeficientes de filtro de predicción lineal, LPC, (20a) y someter la representación de la señal de audio de ruido de fondo reducido (12) al análisis de predicción lineal para obtener un segundo juego de coeficientes filtros de predicción lineal, LPC, (20b); y
un filtro de análisis (22) compuesto por una cascada de filtros de dominio de tiempo (24, 24a, 24b) que es un filtro Wiener y controlado por el primer juego obtenido de coeficientes LPC (20a) y el segundo juego obtenido de coeficientes LPC (20b) para obtener una señal residual (26) de la señal de audio (8'); y
un transmisor (30) configurado para transmitir el segundo juego de coeficientes LPC (20b) y la señal residual (26).

2. Codificador (4) de acuerdo con la reivindicación 1, en donde la cascada de filtros de dominio de tiempo (24) comprende dos veces un filtro de predicción lineal (24a) que usa el primer juego obtenido de coeficientes LPC (20a) y una vez un inverso de un filtro de predicción lineal adicional (24b) que usa el obtenido segundo conjunto de coeficientes LPC (20b).

3. Codificador (4) de acuerdo con la reivindicación 1 o 2, que comprende además un cuantificador (28) configurado para cuantificar y/o codificar la señal residual (26) antes de la transmisión.

4. Codificador (4) de acuerdo con una cualquiera de las reivindicaciones anteriores, que comprende además un cuantificador (28) configurado para cuantificar y/o codificar el segundo juego de coeficientes LPC (20b) antes de la transmisión.

5. Codificador de acuerdo con la reivindicación 3 o 4, en donde el cuantificador está configurado para usar predicción lineal excitada por código, CELP, codificación de entropía, o transformar la excitación codificada, TCX.

6. Sistema (2) que comprende:

el codificador (4) de acuerdo con una cualquiera de las reivindicaciones anteriores;
un decodificador (6) configurado para decodificar la señal de audio codificada.

7. Método (800) para codificar una señal de audio con ruido de fondo reducido utilizando codificación predictiva lineal, comprendiendo el método:

estimar (S802) una autocorrelación del ruido de fondo como una representación del ruido de fondo de la señal de audio;
generar (S804) una representación de una señal de audio de ruido de fondo reducido restando la correlación automática del ruido de fondo de la señal de audio de una autocorrelación de la señal de audio de modo que la representación de la señal de audio de ruido de fondo reducido (16) sea una autocorrelación de una señal de audio de ruido de fondo reducido;
someter (S806) la representación de la señal de audio al análisis de predicción lineal para obtener un primer juego de coeficientes de filtro de predicción lineal, LPC, y someter la representación de la señal de audio de ruido de fondo reducido al análisis de predicción lineal para obtener un segundo juego de coeficientes filtros de predicción lineal, LPC;
controlar (S808) una cascada de filtros de dominio de tiempo que es un filtro Wiener por el primer juego obtenido de coeficientes LPC y el segundo juego obtenido de coeficientes LPC para obtener una señal residual de la señal de audio;
transmitir el segundo juego de coeficientes LPC (20b) y la señal residual (26).

8. Programa informático que comprende instrucciones que, cuando el programa es ejecutado por un ordenador, hace que el ordenador lleve a cabo el método de acuerdo con la reivindicación 7.

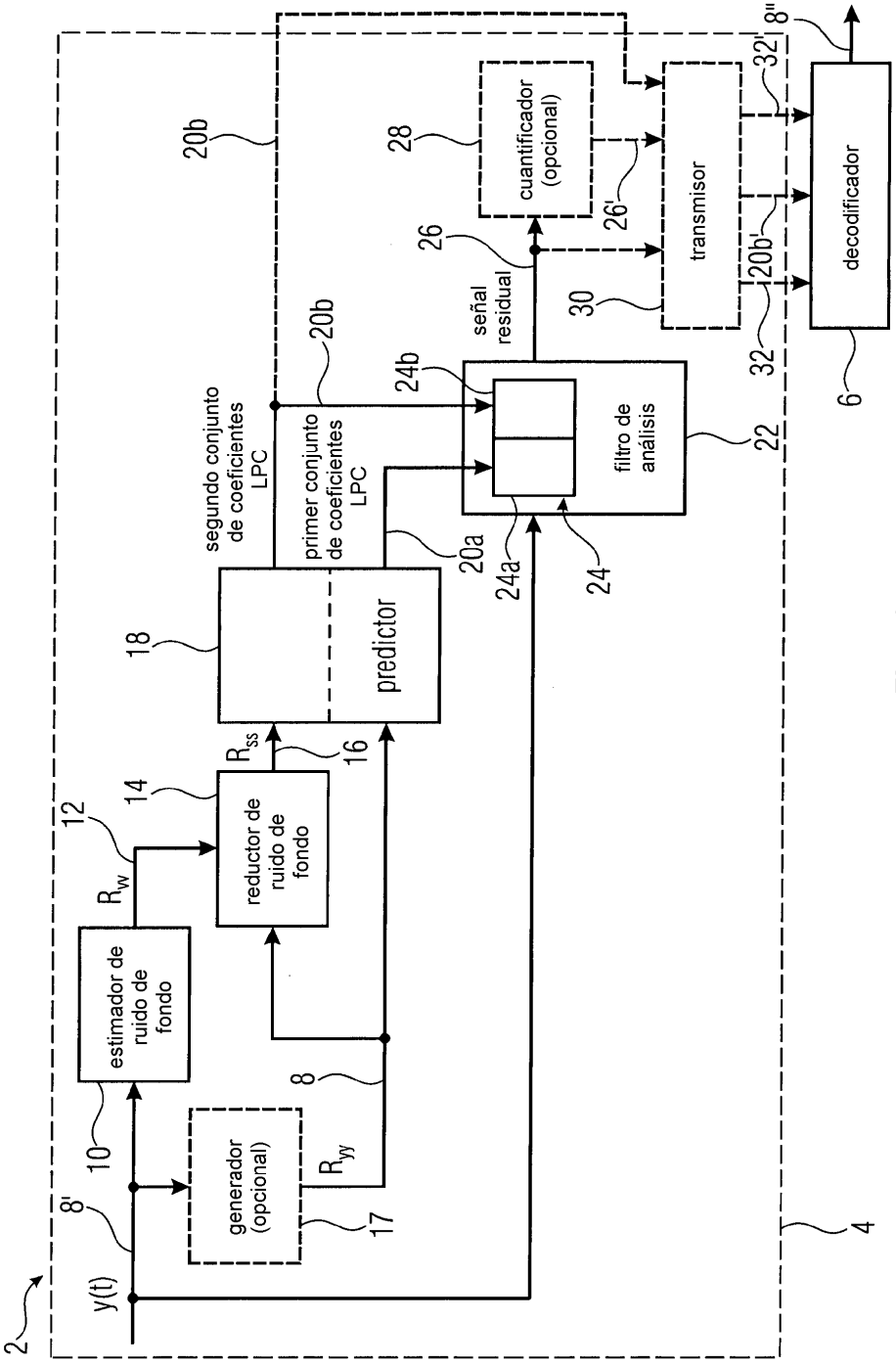

2
y(t)
8'
10
estimador de ruido de fondo
12
R_{vv}
14
reductor de ruido de fondo
R_{ss}
16
18
predictor
20b
segundo conjunto de coeficientes LPC
primer conjunto de coeficientes LPC
20a
24a
24
24b
filtro de análisis
22
señal residual
26
28
cuantificador (opcional)
26'
30
transmisor
32
20b'
32'
8''
decodificador
6
4
8
R_{yy}
generador (opcional)
17

Fig. 1

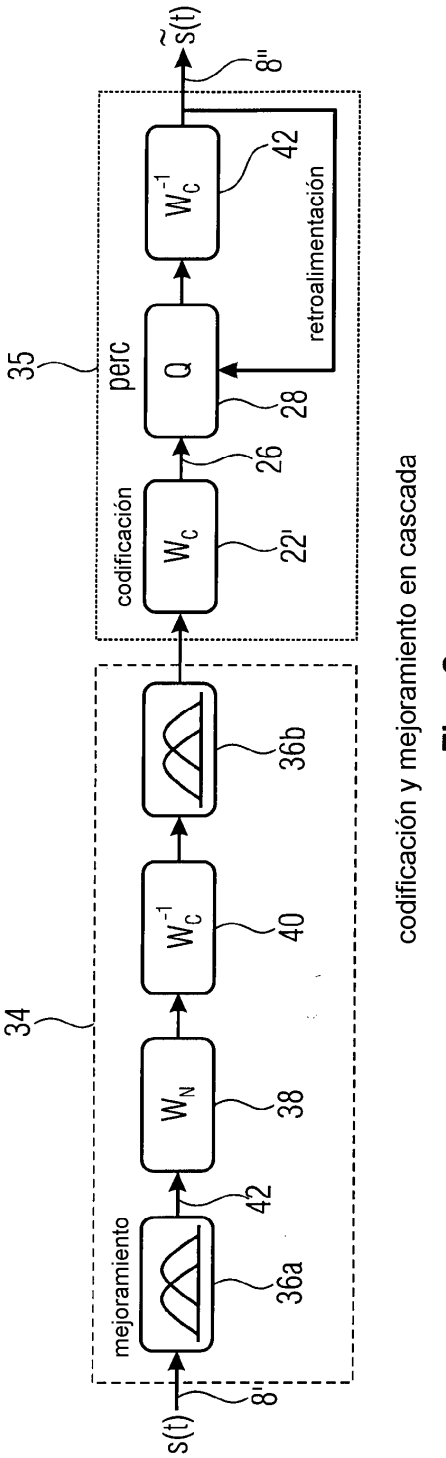


Fig. 2a

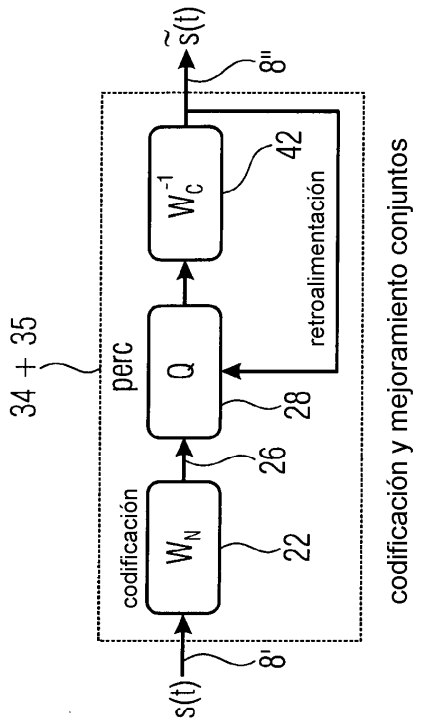


Fig. 2c

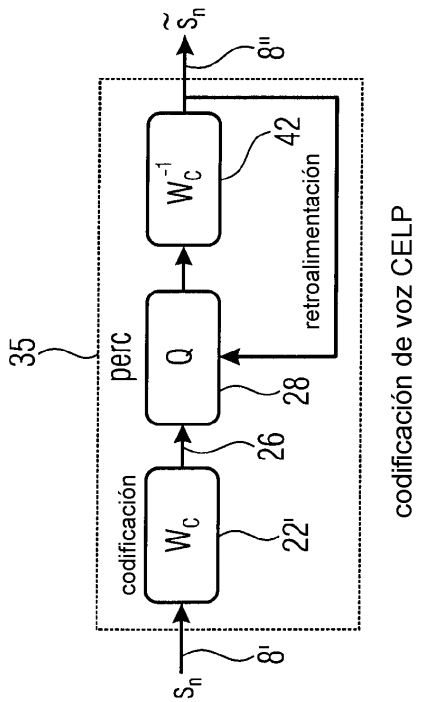
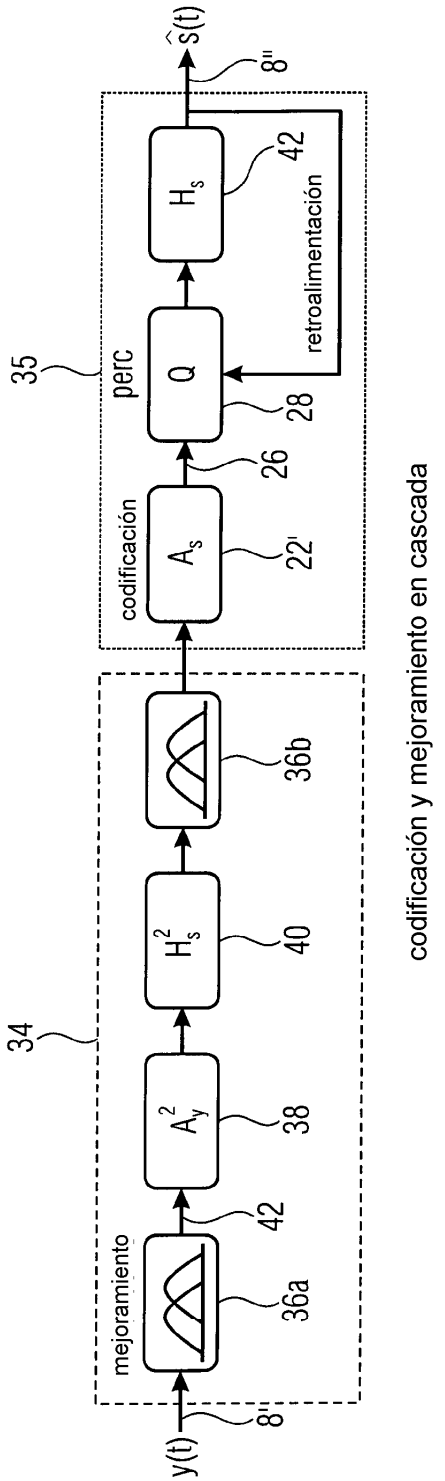
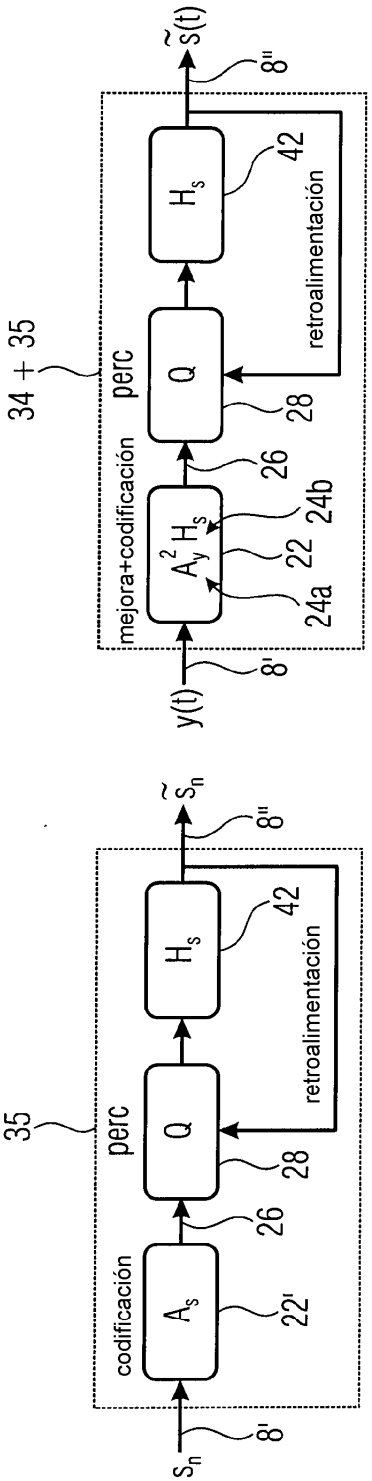


Fig. 2b



codificación y mejoramiento en cascada

Fig. 3a



codificación y mejoramiento conjuntos

Fig. 3c

codificación de voz CELP

Fig. 3b

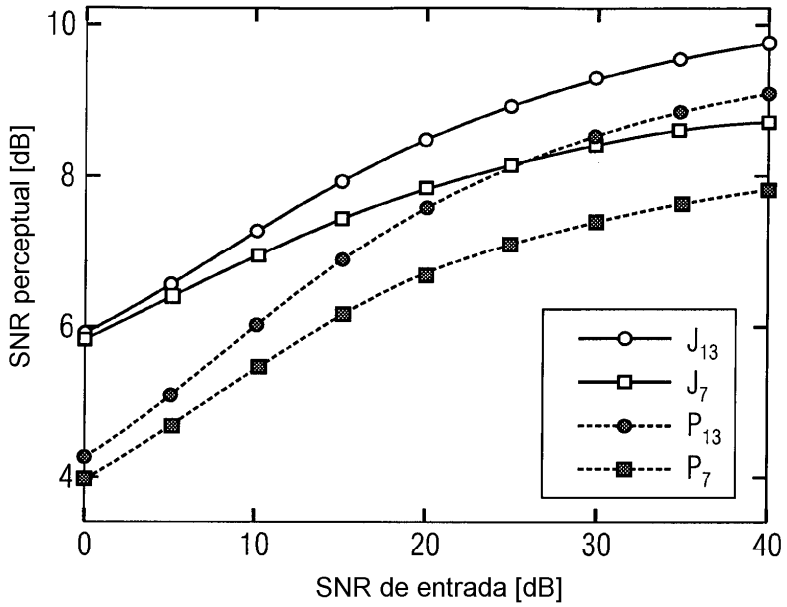

10
8
6
4
SNR perceptual [dB]
0
10
20
30
40
SNR de entrada [dB]
J_{13}
J_7
P_{13}
P_7

Fig. 4

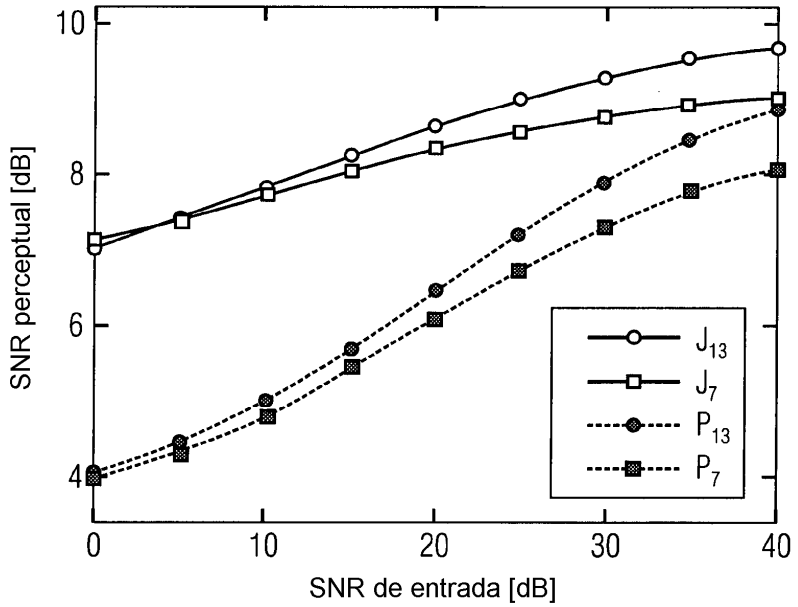

10
8
6
4
SNR perceptual [dB]
0
10
20
30
40
SNR de entrada [dB]
J_{13}
J_7
P_{13}
P_7

Fig. 5

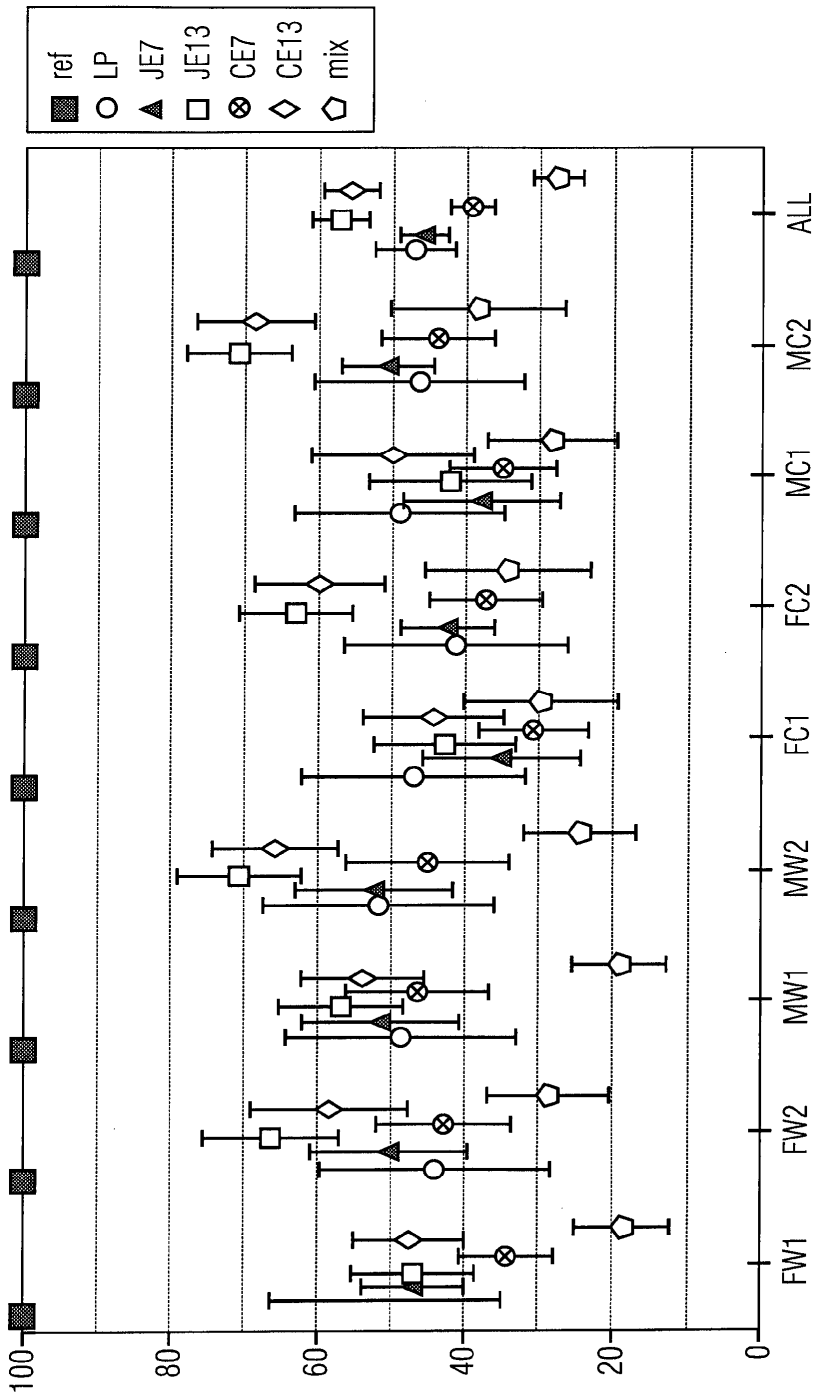
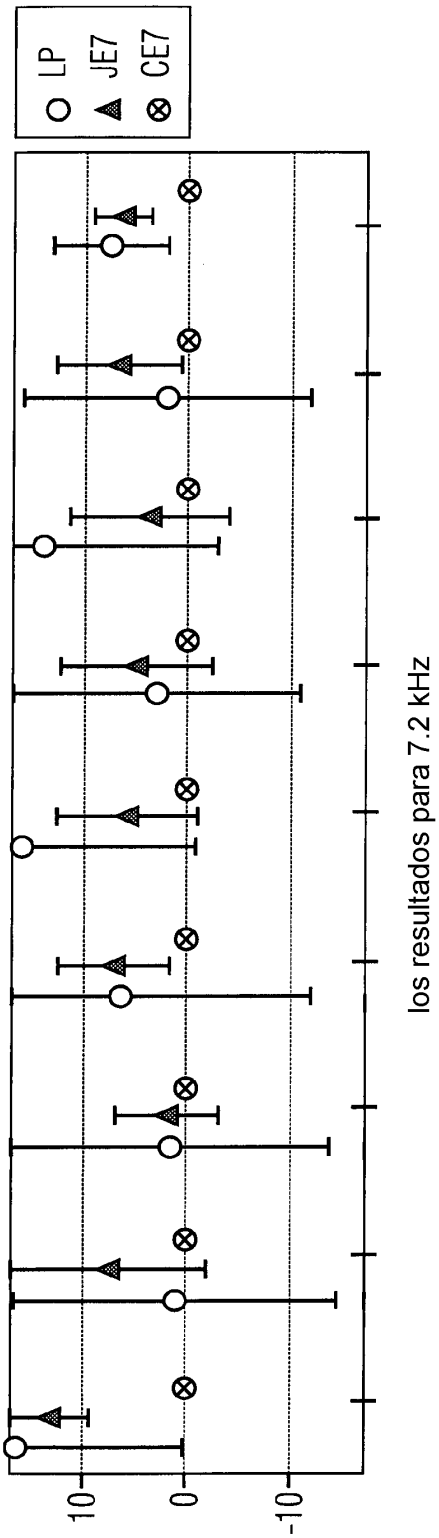
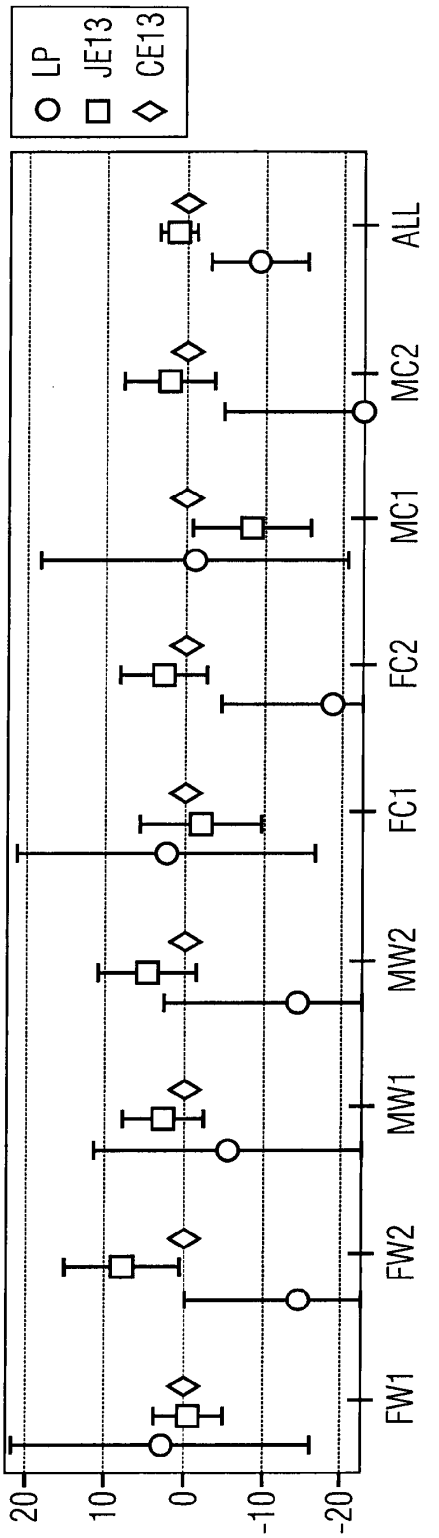

ref
LP
JE7
JE13
CE7
CE13
mix
100
80
60
40
20
0
FW1
FW2
MW1
MW2
FC1
FC2
MC1
MC2
ALL

Fig. 6



los resultados para 7.2 kHz

Fig. 7a



los resultados para 13.2 kHz

Fig. 7b

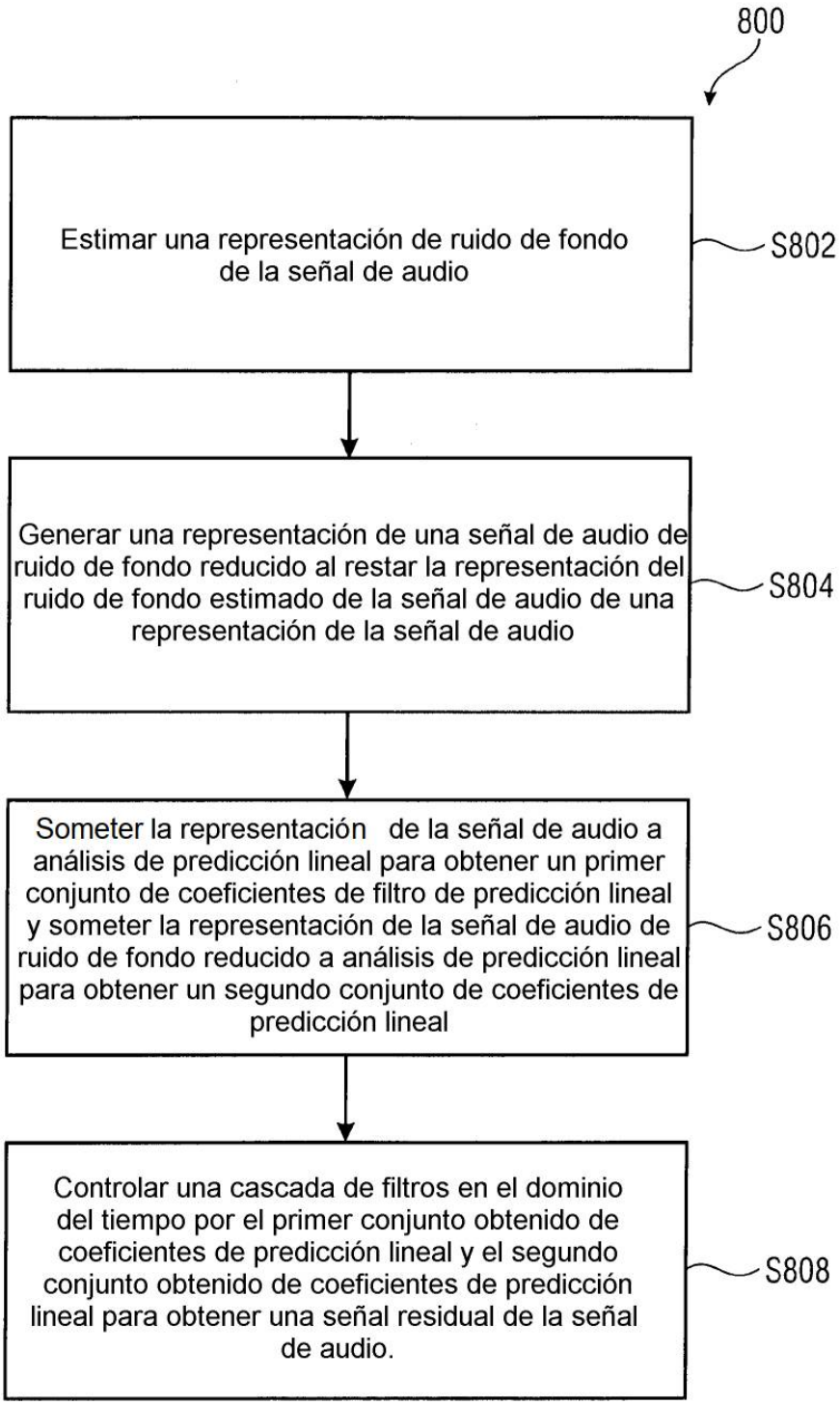

800
Estimar una representación de ruido de fondo de la señal de audio
S802
Generar una representación de una señal de audio de ruido de fondo reducido al restar la representación del ruido de fondo estimado de la señal de audio de una representación de la señal de audio
S804
Someter la representación de la señal de audio a análisis de predicción lineal para obtener un primer conjunto de coeficientes de filtro de predicción lineal y someter la representación de la señal de audio de ruido de fondo reducido a análisis de predicción lineal para obtener un segundo conjunto de coeficientes de predicción lineal
S806
Controlar una cascada de filtros en el dominio del tiempo por el primer conjunto obtenido de coeficientes de predicción lineal y el segundo conjunto obtenido de coeficientes de predicción lineal para obtener una señal residual de la señal de audio.
S808

Fig. 8