(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 928 295**

(51) Int. CI.:

G10L 21/0232 (2013.01)
G10L 25/30 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **14.02.2020 E 20382110 (3)**

(97) Fecha y número de publicación de la concesión europea: **17.08.2022 EP 3866165**

(54) Título: **Método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
16.11.2022

(73) Titular/es:

SYSTEM ONE NOC & DEVELOPMENT SOLUTIONS, S.A. (100.0%)
Josefa Amar y Borbón, 10, Cuarta planta
50001 Zaragoza, ES

(72) Inventor/es:

GALLART MAURI, JAVIER;
GARCIA MORTE, IÑIGO;
RIBAS GONZALEZ, DAYANA;
MIGUEL ARTIAGA, ANTONIO;
ORTEGA GIMÉNEZ, ALFONSO y
LLEIDA SOLANO, EDUARDO

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales

El presente método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales profundas (CNN) es capaz de reducir el efecto de las distorsiones acústicas que se producen en escenarios cotidianos durante una llamada telefónica. Estas distorsiones pueden manifestarse en forma de ruido aditivo, reverberación u otras, y afectar a la inteligibilidad de la voz que se transmite por la línea telefónica, provocando incomodidad entre los interlocutores o incluso el abandono de la conversación. La presente invención es un método de mejora de voz para señales de voz telefónica monocanal con baja latencia. La novedad del método de la presente invención radica en el hecho de que es un método de reducción de ruido que, basándose en el método clásico de la ganancia, utiliza una red neuronal convolucional profunda (CNN) para aprender el estimador de Wiener. A continuación, con esto calcula la ganancia del filtro para mejorar la potencia de la voz frente a la potencia del ruido para cada componente tiempo-frecuencia de la señal. La selección del estimador de la ganancia de Wiener como elemento esencial del método disminuye la vulnerabilidad a errores de estimación, ya que las características de esta medida la hacen muy apropiada para ser estimada mediante enfoques de aprendizaje profundo. El método de la presente invención puede incorporar opcionalmente la capacidad de evaluar la calidad de la señal de voz y, en consecuencia, proceder a aplicar la mejora acústica.

En comparación con estudios previos, la presente invención logra mejorar significativamente la eficacia, permite el procesamiento de ruidos realistas (no estacionarios, mezclados, correlacionados con la voz, etc.) y también mejora en el hecho de que la estimación se realiza de manera no recursiva, lo cual evita que se propaguen errores. Permite dos modos de funcionamiento, uno basado en procesamiento causal, adecuado para aplicaciones que requieran análisis en tiempo real y entornos que requieren causalidad, así como uno hace uso de la señal completa de una manera no causal para aplicaciones con requisitos que lo permitan. Finalmente, se resalta que el método consta de una implementación que implica una baja latencia en el procesamiento.

Campo técnico

La invención pertenece al campo de la tecnología de telecomunicaciones y, más específicamente, a aquellas tecnologías que permiten reducir las distorsiones acústicas en la voz telefónica.

Antecedentes de la invención

Las conversaciones telefónicas que se efectúan en escenarios cotidianos, por ejemplo, en el hogar, una oficina, un parque público, una calle, etc., en su mayoría se ven afectadas por ruido ambiental, efectos reverberantes que se producen habitualmente cuando se utiliza un dispositivo de manos libres en un entorno interior, un micrófono distante, entre otros. Estas distorsiones acústicas se combinan con la voz y se transmiten como un todo a través de la línea telefónica. De esta manera, la inteligibilidad de la voz que llega al otro extremo se ve comprometida según el nivel de afectación de la señal de voz. Niveles moderados de distorsión pueden provocar la incomodidad de los interlocutores involucrados en la conversación. Sin embargo, a medida que aumenta el nivel de afectación de la señal, los interlocutores pueden incluso considerar terminar la llamada. El uso de un método de mejora de voz contribuye a que la repetición de estas situaciones indeseables no afecte a la calidad del servicio y mejor, por lo tanto, la reputación del proveedor de servicios telefónicos.

Dichos métodos son capaces de procesar señales de voz con distorsiones típicas de ambientes reales, ofreciendo una señal de mejor calidad acústica. El método de mejora de voz en el dominio espectral basado en ganancia es un paradigma establecido para reducir ruido en señales de voz monocanal (Philipos C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Nueva York, 2013). El método de la ganancia consiste en obtener una representación de tiempo-frecuencia de la señal de voz y estimar una ganancia dependiente de la frecuencia y que varía en el tiempo, según el nivel de afectación por ruido de cada zona espectral. Esta ganancia se utiliza para modificar los componentes de tiempo-frecuencia de la representación espectral de la señal según el predominio de voz o ruido, lo cual se determina a través de la probabilidad de presencia de voz. La aplicación del filtro de mejora en la representación de tiempo-frecuencia de la señal de voz da lugar a una versión modificada del espectro que se aproxima a la señal de voz limpia. A continuación, un algoritmo de reconstrucción aplica una transformación inversa, de acuerdo con a la utilizada inicialmente, para obtener las muestras de la señal mejorada en el dominio del tiempo.

En el estado de la técnica existe una gran familia de algoritmos estadísticos derivados a partir de este paradigma. Entre ellos es imprescindible mencionar los clásicos de filtrado de Wiener (Norbert Wiener. "Extrapolation, Interpolation, and Smoothing of Stationary Time Series". Nueva York: Wiley. ISBN 978-0-262-73005-1, 1949) y Sustracción Espectral (S. Boll, "Suppression of acoustic noise in speech using spectral subtraction" IEEE Trans. on Acoustic, Speech and Signal Processing, vol. 27, n.º 2, pp. 113-120, 1979), más las numerosas variantes de los mismos; el Estimador de la Amplitud Espectral a Corto Plazo (STSA) (Y. Ephraim y D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator" IEEE Trans. on Acoustic, Speech and Signal Processing, vol. 32, n.º 6, pp. 1109-1121, 1984) y la evolución del mismo, el Estimador de la Amplitud Espectral Logarítmica (en inglés: Logarithmic Spectral Amplitude, LSA) (Y. Ephraim y D. Malah, "Speech

enhancement using minimum-mean square log spectral amplitude estimator" IEEE Trans. on Acoustic, Speech and Signal Processing , vol. 33, n.º 2, pp. 443-445, 1985), que también han sido inspiración para numerosas propuestas, por ejemplo, el Estimador de la Amplitud Espectral Logarítmica Óptimamente Modificado (OMLSA) (I. Cohen y B. Berdugo, "Speech enhancement for non-stationary noise environments", Signal Processing, vol. 81, n.º 11, pp. 2403-2418, 2001*)*, entre otros. Sin embargo, la base estadística de este marco tiene limitaciones en la eficacia de las estimaciones internas del método, especialmente cuando se enfrenta a ambientes de ruido reales que pueden tener mezclas de tipos de ruido, ruidos impulsivos, ruido correlacionado con la voz, etc.

El resurgimiento del aprendizaje automático profundo ha influido en las técnicas clásicas de reducción de ruido. En general, la mejora de voz monocanal basada en redes neuronales profundas se divide en dos tendencias fundamentales: las técnicas basadas en aprendizaje de máscaras o aproximación de máscaras y las técnicas basadas en correlación de características o aproximación de señales. Sin embargo, el método de la ganancia se mantiene como el paradigma subyacente en ambos casos.

Estudios previos en el contexto relacionados con la propuesta, es decir, los métodos de mejora de voz monocanal utilizando redes neuronales profundas (DNN) y los métodos basados en máscaras, difieren ambos por el contexto de aplicación del método que proponen, así como por la esencia de la novedad que plantean.

El estado de la técnica fundamental a la presente invención está compuesto por:

- B.Y. Xia y C.-C. Bao, "Speech enhancement with weighted denoising auto-encoder", en Proc. Interspeech, 2013*:* que propuso estimar el espectro de la señal de voz limpia utilizando una DNN de tipo Autoencoder y, a continuación, continuar con el proceso recursivo de estimaciones para obtener el filtro de mejora de voz. Esta fue una aproximación preliminar del método de ganancia al aprendizaje profundo. Sin embargo, el contexto del método de reducción de ruido en general difería notablemente de esta propuesta. En Xia et al. la DNN solo se utilizó para estimar la señal de voz limpia, que es un paso intermedio en la obtención de la ganancia de Wiener, mientras que en la presente invención, la DNN estima directamente la ganancia de Wiener. A continuación, en Xia et al., se mantuvieron intactos los siguientes elementos que componen el marco del método de ganancia clásico, lo cual les permitía obtener mejoras muy modestas en los resultados.

Además, varios estudios se han basado en el desarrollo de una solución desde el punto de vista del Análisis de Escena Auditiva Computacional (CASA). Estos se han centrado en la estimación de la Máscara Binaria Ideal (IBM) o de la Máscara de Relación Ideal (IRM). La definición de estas máscaras se asemeja al estimador de la ganancia de Wiener, pero no es exactamente igual. Matemáticamente la IRM se define de forma más genérica que la ganancia de Wiener, permitiendo variaciones en la implementación de la misma.

- A. Narayanan y D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition" en IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2013, pp. 7092-7096: la DNN estima una variante que implementa la SNR instantánea comenzando a partir de la IRM, alejándose de la estimación de la ganancia de Wiener. A continuación, la representación de tiempo-frecuencia de la señal modificada se entrega a un sistema de reconocimiento de voz, y la mejora de la señal nunca se reconstruye.

- E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, y D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type", The Journal of the Acoustical Society of America, vol. 138, n.º 3, pp. 1660-1669, 2015: este método estima la raíz cuadrada de la IRM, mientras que en Tobias Goehring, F. Bolner, J.J. Monaghan, B. van Dijk, A. Zarowski y S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users", The Journal of Hearing research, vol. 344, pp. 183-194, 2017", se estima directamente la IRM. A diferencia de la invención, estos métodos funcionan en el entorno de las ayudas de audición, por lo que la señal resultante tras el tratamiento pasa directamente al dispositivo de asistencia auditiva y, por lo tanto, nunca se reconstruye.

Breve descripción de la invención

En un primer aspecto de la invención, se divulga un método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales. El método de la presente invención se aplica a una señal de voz telefónica que se compone de magnitud y fase espectral. El método comprende las siguientes etapas:

- una etapa de preprocesamiento que comprende extraer la magnitud y la fase de la representación espectral de la señal de voz telefónica;
- una etapa de reducción de ruido que comprende aplicar a la magnitud de la representación espectral de la señal de voz telefónica las siguientes etapas:
    - aplicar un estimador espectral;
    - calcular una representación perceptual;
    - aplicar una Red Neuronal Convolucional que, con unas entradas correspondientes al estimador espectral y

a la representación perceptual, genera como salida, una estimación de ganancia de Wiener consistente en una matriz/vector dependiente de la frecuencia y que varía en el tiempo;
○ utilizar la estimación de ganancia de Wiener dentro de la ganancia del filtro de mejora de la siguiente función (función f1):

$$G_{filtro}(t,f) = [\hat{G}_{Wiener}(t,f)\ exp(\frac{1}{2}\int_{v(t,f)}^{\infty}\frac{e^{-t}}{t}dt)]^{p(t,f)}\ G_{mín}^{1-p(t,f)}$$

donde *t* es el segmento de tiempo, *f* el intervalo de frecuencia, $\hat{G}_{Wiener} = DNN(x_t, x_{t-1}, \dots)$ con x_t el vector de parámetros espectrales y perceptuales, $G_{mín}$ es una constante, $p(t,f)$ es la probabilidad de presencia de voz y $v(t,f) = \frac{\hat{G}_{Wiener}}{1-\hat{G}_{Wiener}}$,
y en paralelo, utilizar la estimación de ganancia de Wiener como estimación de la probabilidad de presencia de voz;
○ aplicar la función definida en el artículo anterior (función f1) como un filtro de mejora de voz.

- una etapa de posprocesamiento que comprende fusionar la fase inicial con la magnitud mejorada en la etapa de reducción de ruido.

La Red Neuronal Convolucional se entrena con una función de coste que es el MMSE entre el estimador óptimo de Wiener y la salida de la Red Neuronal Convolucional definida por medio de:

$$F_{coste} = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{F}(G_{Wiener}(t,f) - \hat{G}_{Wiener}(t,f))^2$$

donde $G_{Wiener}(t,f) = \frac{S_{X(t,f)}}{S_{X(t,f)}+S_{N(t,f)}}$ se obtiene de una manera supervisada, siendo $S_{X(t,f)}$ y $S_{N(t,f)}$ respectivamente las estimaciones de las densidades espectrales de potencia de la señal de voz limpia y del ruido.

La Red Neuronal Convolucional puede comprender una capa convolucional que es causal (es decir, la capa convolucional sólo utiliza muestras de señal anteriores a la que está siendo procesada y no posteriores) y tiene baja latencia.

Adicionalmente, el estimador espectral se obtiene utilizando el método de Welch.

Con respecto a la representación perceptual se consideran dos métodos: un banco de filtros de escala de Mel y una representación basada en coeficientes cepstrales en las frecuencias de Mel (MFCC).

Con respecto a la etapa de preprocesamiento para extraer la magnitud y la fase de la señal de voz telefónica, el método de la presente invención puede comprender adicionalmente dividir la señal de voz en segmentos solapados de decenas de milisegundos a los que se aplica una ventana tipo Hamming, una ventana de Hanning u otra ventana equivalente, y posteriormente una transformada de Fourier.

Con respecto a la etapa de preprocesamiento para extraer la magnitud y la fase de la señal de voz telefónica, dicha etapa puede comprender adicionalmente evaluar de forma objetiva la calidad de la señal de voz utilizando una medida de calidad acústica seleccionada de entre SNR (Relación Señal a Ruido), distorsión, POLQA (Análisis de Calidad de la Audición Objetiva de Percepción) u otra equivalente y, dependiendo de este resultado, proceder o no a la mejora acústica.

Con respecto a la etapa de posprocesamiento para fusionar la fase obtenida en la etapa de preprocesamiento con la magnitud obtenida en la etapa de reducción de ruido, dicha etapa comprende adicionalmente aplicar una transformada inversa de Fourier, y posteriormente, un algoritmo de reconstrucción temporal de la señal de voz.

En otro aspecto de la presente invención, se divulga un sistema de mejora de las señales de voz telefónica basado en redes neuronales convolucionales y configurado para realizar el método de la presente invención definido en el primer aspecto de la invención. La presente invención comprende un primer bloque de extracción de señal, un segundo bloque de evaluación de la calidad de la señal, un tercer bloque de mejora de voz y un cuarto bloque de inserción de voz. Por lo tanto, el primer bloque, cuando se inicia una llamada telefónica, procede a extraer la señal de voz de la línea telefónica y la envía a un servidor paralelo de procesamiento. A continuación, el segundo bloque evalúa la calidad acústica del segmento de voz utilizando una medida de calidad acústica predefinida, por ejemplo: SNR, distorsión, POLQA u otra equivalente, y compara con un umbral preestablecido según estudios previos relacionados con la medida de calidad utilizada. De esta manera, el método decide si hay necesidad de someter el segmento a mejora acústica. Si la decisión es positiva, el tercer bloque, implementa el método de mejora de voz de la presente invención en el segmento de voz analizado. Si la decisión fue negativa, se pasa directamente al cuarto bloque de inserción de voz que es responsable de insertar el segmento de la señal de voz en la línea telefónica, salvaguardando la aparición de cortes u otros efectos indeseables que puedan afectar la percepción acústica del interlocutor. A continuación, se repite el proceso descrito y se analiza el siguiente segmento de voz.

Breve descripción de las figuras

Para ayudar a una mejor comprensión de las características de la invención y para complementar esta descripción, las siguientes figuras se adjunta como parte integrante de la misma, por medio de ilustración y no de limitación:

La Figura 1 muestra un diagrama de bloques del sistema donde se inserta el método de mejora de la señal de voz telefónica.

La Figura 2 muestra un diagrama de bloques de la invención que incluye una etapa A de preprocesamiento o parametrización, una etapa B de reducción de ruido donde reside la novedad y etapa C de posprocesamiento o reconstrucción de la señal de voz.

Descripción detallada de una realización ilustrativa

Acrónimos:

FB: Banco de Filtros
MFCC: Coeficientes Cepstrales en las Frecuencias de Mel
DNN: Red Neuronal Profunda
CNN: Red Neuronal Convolucional
MMSE: Error Cuadrático Medio Mínimo
SNR: Relación Señal a Ruido
POLQA: Análisis de Calidad de la Audición Objetiva de Percepción

El método de la presente invención es un método de mejora de voz, el que opcionalmente puede incluir un módulo que analiza la calidad acústica de la señal de voz y decide si necesita someterse al proceso de mejora acústica. Esta decisión se toma por medio de umbrales preestablecidos de la medida de calidad. El operador puede decidir flexibilizar al máximo el umbral, tal que siempre se realice el proceso de mejora, o por el contrario restringirlo al máximo de tal que no se mejora la señal que pasa por la línea telefónica. Estos dos casos de uso pueden responder a aplicaciones específicas o situaciones circunstanciales decididas por el operador. En caso de someter la señal a mejora, a continuación se reinyecta en la línea telefónica, mostrando una mejor calidad acústica y de esta manera continua la trayectoria de la misma hasta el otro extremo.

El método de la presente invención se puede desplegar en un sistema como el mostrado en la Figura 1. El sistema mostrado en la Figura 1 comprende un bloque 1 de extracción de señal, un bloque 2 de evaluación de la señal, un bloque 3 de mejora de voz y un bloque 4 de inserción de voz. Por lo tanto, el sistema de mejora de las señales de voz telefónica basado en redes neuronales convolucionales de la presente invención como se representa en la Figura 1, cuando se inicia una llamada telefónica, procede a extraer la señal **1** de voz de la línea telefónica y la envía a un servidor paralelo de procesamiento. A continuación, el bloque **2** evalúa la calidad acústica del segmento de voz utilizando una medida de calidad acústica predefinida, por ejemplo: SNR, distorsión, POLQA u otra equivalente, y compara con un umbral preestablecido según estudios previos relacionados con dicha medida de calidad. De esta manera, el método decide si hay necesidad de someter el segmento a mejora acústica. Si la decisión es positiva, el siguiente bloque **3**, implementa el método de mejora de voz de la presente invención en el segmento de voz analizado. Si la decisión fue negativa, pasa directamente al bloque de inserción **4** que se encarga de devolver el segmento de voz al flujo telefónico, salvaguardando la aparición de cortes u otros efectos indeseables que puedan afectar la percepción acústica del interlocutor. A continuación, se repite el proceso descrito y se analiza el siguiente segmento de voz.

La presente invención consiste en un método de mejora de voz o reducción de ruido **3** para señales telefónicas monocanal basado en CNN, en la categoría de los métodos basados en máscaras. En general, la presente invención consiste en estimar una máscara (o filtro) que modifica el espectro de la señal de voz observada para generar una versión mejorada de la misma. Esta aprovecha el esquema del método de la ganancia, pero sustituye los bloques de estimación intermedios de SNR a priori, SNR a posteriori, estimación del espectro de ruido y estimación del espectro de la voz limpia, motivado por la tendencia del mismo a introducir errores. En su lugar, el método de la presente invención emplea una estimación basada en CNN capaz de realizar el análisis de cada segmento de tiempo-frecuencia de la señal de una manera no recursiva, evitando de esta forma la propagación de errores típicos de este tipo de estimación. Específicamente, la CNN es responsable de obtener una estimación de la ganancia de Wiener, que se utiliza para generar la máscara o filtro de mejora. Para esto, la CNN implementa un modelo de regresión que aprende el estimador de MMSE de la señal de voz limpia, también conocido como estimador de la ganancia de Wiener. Este término es menos sensible a errores de estimación que los pasos intermedios de estimación antes mencionados, debido al rango dinámico del mismo y a las operaciones del método de aprendizaje de la CNN. De esta manera, el método de mejora de voz asegura que no disminuye la calidad acústica incluso en señales de voz que no están notablemente afectadas por ruido.

La novedad de la presente invención radica en el diseño del método de reducción de ruido basado en CNN, tanto por la esencia del método en sí, como por el contexto de aplicación del mismo. La principal novedad del método consiste en la selección del estimador de ganancia de Wiener como un elemento esencial del método de la presente invención, cuyas características lo hacen apropiado para ser estimado mediante aprendizaje automático, lo cual disminuye la sensibilidad a errores de estimación.

La Figura 2 muestra un diagrama de flujo del método implementado, formado por tres etapas de procesamiento. La etapa **A** realiza un preprocesamiento de la señal de voz ruidosa **10** que es responsable de representarla en el dominio de tiempo-frecuencia. Esta comienza con la segmentación de la señal de voz **10** en segmentos cortos solapados de decenas de milisegundos que conservan las propiedades cuasi estacionarias de la voz, a la que se aplica una ventana de tipo Hamming, ventana de Hanning u otra ventana adecuada para evitar distorsiones **11**. A continuación, se realiza una transformación al dominio de espectro-temporal **12**, que puede implementarse comenzando a partir de una transformada de Fourier u otra transformación equivalente. A continuación, el espectro resultante se divide en magnitud **13** y fase espectral **14**. La magnitud **13** se utiliza como entrada de la etapa **B** de reducción de ruido, mientras que la fase espectral **14** se guarda para la reconstrucción que se implementa en la etapa **C**.

En la etapa **B** de reducción de ruido se concentra la novedad de la propuesta. Esta se encarga de generar un filtro de mejora con el que compensar los efectos del ruido acústico en la magnitud espectral **13** del segmento bajo análisis. La ganancia de dicho filtro **22** depende de la función de ganancia del estimador MMSE de la señal de voz limpia **20** y de la probabilidad de presencia de voz **21**. Para obtener estos elementos se utiliza una CNN que estima la ganancia de Wiener **19** comenzando a partir de aprender la estructura de la voz ruidosa, viendo múltiples ejemplos de espectros de señales de voz y los correspondientes espectros de ruido asociados de forma separada.

La arquitectura de DNN en este caso consiste en una Red Neuronal Convolucional (CNN) **18** con múltiples entradas que se apilan juntas en un vector que incluye una o varias representaciones espectrales, por ejemplo estimaciones espectrales obtenidas por medio del método de Welch **15** u otra representación equivalente, así como una o varias representaciones perceptuales de la señal de voz observada, por ejemplo el banco de filtros de escala Mel (FB) **16**, los coeficientes cepstrales en las frecuencias de Mel (MFCC) **17** u otras representaciones equivalentes. En la presente invención el procesamiento de la capa convolucional está configurado de forma causal, es decir que solo hace uso de la información del pasado, lo cual permite que actúen en tiempo real. La salida de la CNN es una estimación de la ganancia de Wiener **19**, que consiste en una matriz/vector dependiente de la frecuencia y que varía en el tiempo. Esta matriz/vector se utiliza como la ganancia del estimador MMSE de la señal de voz limpia **20** y como estimación de la probabilidad de presencia de voz **21**. Posteriormente, ambas se utilizan para obtener la función del filtro de mejora de voz **22** según la siguiente definición (I. Cohen y B. Berdugo, "Speech enhancement for non-stationary noise environments", Signal Processing , vol. 81, n.º 11, pp. 2403-2418, 2001):

$$G_{filtro}(t,f) = [\hat{G}_{Wiener}(t,f)\, exp(\frac{1}{2}\int_{v(t,f)}^{\infty} \frac{e^{-t}}{t} dt)]^{p(t,f)} G_{mín}^{1-p(t,f)}$$

donde *t* es el segmento de tiempo, *f* el intervalo de frecuencia, $\hat{G}_{Wiener} = DNN(x_t, x_{t-1}, \ldots)$ con x_t el vector de parámetros espectrales y perceptuales en el instante de tiempo t, $G_{mín}$es una constante, $p(t,f)$ es la probabilidad de presencia de voz y $v(t,f) = \frac{\hat{G}_{Wiener}}{1-\hat{G}_{Wiener}}$.

La función del filtro de mejora de voz se define con la misma resolución de tiempo-frecuencia mencionada basada en la ganancia de Wiener y aplicando un tratamiento diferenciado a los segmentos de voz y no-voz. Este criterio se basa en considerar que la afectación del ruido acústico se manifiesta de forma diferente en las zonas de voz y no voz. Finalmente, este filtro es responsable de mejorar el espectro de la señal de voz, por tanto se aplica a la magnitud espectral **13** que resultó de la etapa **A**. Obsérvese que la reducción de ruido se implementa de manera no recursiva, la razón por la cual los errores potenciales que se originan para un cierto segmento de señal no afectarán etapas posteriores de procesamiento.

La red neuronal convolucional de la presente invención necesita ser entrenada. En el presente caso, en la etapa de entrenamiento la función de coste es el error cuadrático medio entre el estimador óptimo de Wiener y la salida de la red:

$$F_{coste} = \frac{1}{T}\sum_{t=1}^{T} \sum_{f=1}^{F} (G_{Wiener}(t,f) - \hat{G}_{Wiener}(t,f))^2.$$

Para hacer esto de manera supervisada, se calcula $G_{Wiener}(t,f) = \frac{S_{X(t,f)}}{S_{X(t,f)}+S_{N(t,f)}}$ que utiliza las estimaciones de las densidades de potencia espectrales de la señal de voz limpia $S_{X(t,f)}$ y del ruido $S_{N(t,f)}$ que dieron lugar a la señal de voz telefónica observada. Este espectro se estima según el método de Welch, que realiza un promedio en los M segmentos de tiempo solapados para obtener una estimación con menor varianza. La red neuronal se entrena con una gran cantidad de señales de voz limpia (cientos de horas de voz) y las correspondientes señales de ruido de las

mismas. Las señales de ruido utilizadas en el entrenamiento cubren una amplia gama de condiciones ruidosas que potencialmente podrían aparecer en los escenarios reales de aplicación, por ejemplo, varios tipos y niveles de ruido o reverberación. Además de las señales de ruido reales, se hacen modificaciones artificiales en la mezcla de voz y ruido de tal forma que se cubre la mayor cantidad de ejemplos vistos por la red en la etapa de aprendizaje de la misma, por ejemplo, cambios de escala, compresión, entre otros.

Finalmente, la etapa **C** de posprocesamiento finaliza el proceso de reducción de ruido obteniendo una señal de voz mejorada **26**. Para esto utiliza la fase espectral **14** que resultó del preprocesamiento de la etapa **A** y la magnitud espectral mejorada **23** resultante del procesamiento de la etapa **B**. Ambos se insertan en un bloque de transformación espectral inversa **24**, empleando el algoritmo de transformación espectral correspondiente al utilizado en la etapa **A**. A continuación, se utiliza un algoritmo de reconstrucción temporal **25**, que tiene en cuenta el solapamiento y enventanado que se utilizó en la segmentación temporal **11** de la etapa **A**. Finalmente se obtiene la forma de onda mejorada de la señal de voz **26**.

REIVINDICACIONES

1. Un método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, comprendiendo el método:

- una etapa de preprocesamiento (A) que comprende extraer la magnitud y la fase de una representación espectral de la señal de voz telefónica;
- una etapa de reducción de ruido (B) que comprende aplicar a la magnitud de la representación espectral de la señal de voz telefónica los siguientes pasos:
  - aplicar un estimador espectral (15);
  - calcular una representación perceptual (16, 17);
  - aplicar una Red Neuronal Convolucional (18) que, con unas entradas correspondientes al estimador espectral (15) y a la representación perceptual (16, 17), genera como salida, una estimación de ganancia de Wiener (19) consistente en una matriz/vector dependiente de la frecuencia y que varía en el tiempo;
  - utilizar la estimación de ganancia de Wiener dentro del filtro de mejora de la función f1:

$$G_{filtro}(t,f) = [\hat{G}_{Wiener}(t,f)\ exp(\frac{1}{2}\int_{v(t,f)}^{\infty}\frac{e^{-t}}{t}dt)]^{p(t,f)}\ G_{mín}^{1-p(t,f)}$$

  donde *t* es el segmento de tiempo, *f* el intervalo de frecuencia, $\hat{G}_{Wiener} = DNN(x_t, x_{t-1}, \ldots)$ con x_t el vector de parámetros espectrales y perceptuales, $G_{mín}$ es una constante, $p(t,f)$ es la probabilidad de presencia de voz y $v(t,f) = \frac{\hat{G}_{Wiener}}{1-\hat{G}_{Wiener}}$;
  y también utilizar la estimación de ganancia de Wiener como una probabilidad de presencia de voz (21);
  - aplicar la función f1 anterior como un filtro de mejora de voz;

- una etapa de posprocesamiento (C) que comprende fusionar la fase inicial con la magnitud mejorada en la etapa de reducción de ruido (B).

2. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, **caracterizado por que** la Red Neuronal Convolucional (18) se entrena con una función de coste que es el error cuadrático medio entre el estimador óptimo de Wiener y la salida de la Red Neuronal Convolucional (18) definida mediante:

$$F_{coste} = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{F}(G_{Wiener}(t,f) - \hat{G}_{Wiener}(t,f))^2$$

donde $G_{Wiener}(t,f) = \frac{S_{X(t,f)}}{S_{X(t,f)}+S_{N(t,f)}}$ se obtiene de una manera supervisada, siendo $S_{X(t,f)}$ y $S_{N(t,f)}$ respectivamente las estimaciones de las densidades espectrales de potencia de la señal de voz limpia y del ruido.

3. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, **caracterizado por que** la etapa de preprocesamiento (A) para extraer la magnitud y la fase de la representación espectral de la señal de voz telefónica comprende adicionalmente dividir la señal de voz en segmentos solapados de decenas de milisegundos a los que se aplica una ventana de Hamming o Hanning, y posteriormente una transformada de Fourier.

4. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, **caracterizado por que** el estimador espectral se calcula mediante el método de Welch.

5. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, **caracterizado por que** la representación perceptual se calcula aplicando un banco de filtros de escala de Mel (16).

6. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, **caracterizado por que** la representación perceptual se realiza con coeficientes cepstrales en las frecuencias de Mel (MFCC) (17).

7. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según las reivindicaciones 1 y 6, **caracterizado por que** la etapa de posprocesamiento (C) para fusionar la fase obtenida en la etapa de preprocesamiento (A) con la magnitud espectral obtenida en la etapa de reducción de ruido (B) comprende adicionalmente aplicar una transformada inversa de Fourier, y posteriormente, una algoritmo de reconstrucción temporal.

8. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 2, donde la Red Neuronal Convolucional (18) comprende al menos una capa convolucional que es causal y tiene baja latencia.

9. El método de mejora de las señales de voz telefónica basado en redes neuronales convolucionales, según la reivindicación 1, donde la etapa de preprocesamiento (A) comprende adicionalmente evaluar de forma objetiva la calidad de la señal de voz (2) utilizando una medida de calidad acústica seleccionada de entre SNR, distorsión y POLQA.

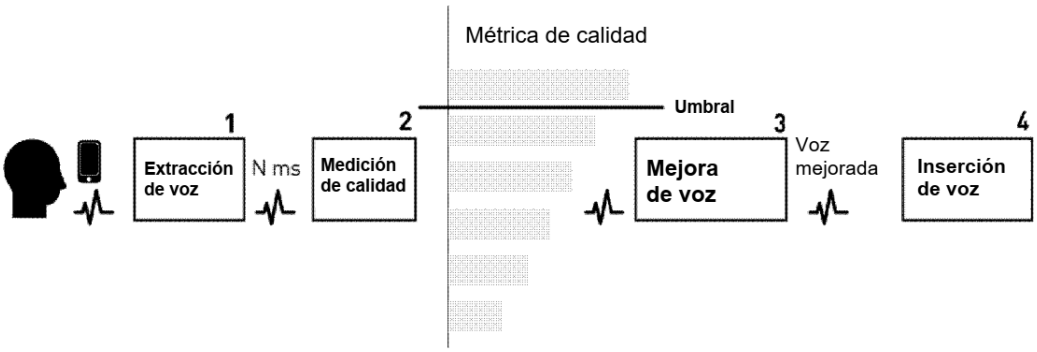
Métrica de calidad
Umbral
1
Extracción de voz
N ms
2
Medición de calidad
3
Mejora de voz
Voz mejorada
4
Inserción de voz

FIG. 1

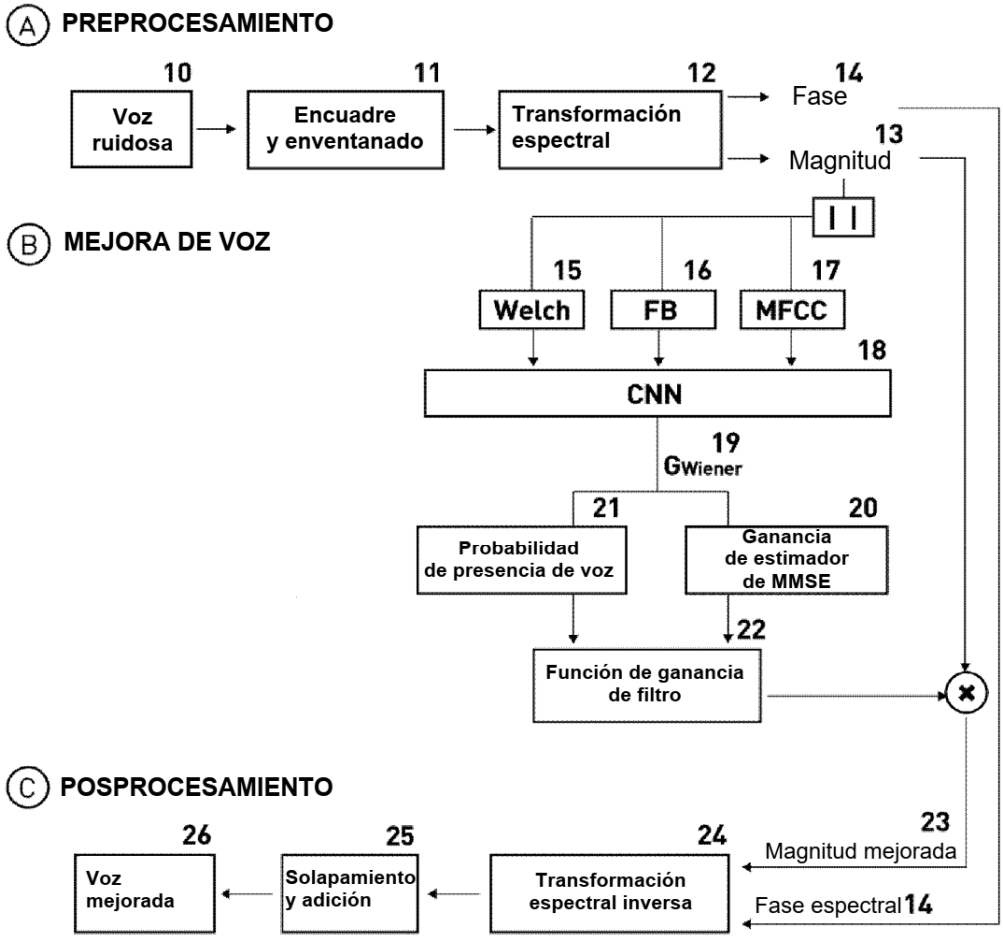
A PREPROCESAMIENTO
10
Voz ruidosa
11
Encuadre y enventanado
12
Transformación espectral
14
Fase
13
Magnitud
B MEJORA DE VOZ
15
Welch
16
FB
17
MFCC
18
CNN
19
GWiener
21
Probabilidad de presencia de voz
20
Ganancia de estimador de MMSE
22
Función de ganancia de filtro
C POSPROCESAMIENTO
26
Voz mejorada
25
Solapamiento y adición
24
Transformación espectral inversa
23
Magnitud mejorada
Fase espectral 14

FIG. 2