(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 654 318**

(51) Int. CI.:

G10L 21/0208 (2013.01)
G10L 21/0264 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **27.07.2007 PCT/NL2007/050378**

(87) Fecha y número de publicación internacional: **05.02.2009 WO09017392**

(96) Fecha de presentación y número de la solicitud europea: **27.07.2007 E 07793879 (3)**

(97) Fecha y número de publicación de la concesión europea: **04.10.2017 EP 2201567**

(54) Título: **Supresión de ruido en señales de voz**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
13.02.2018

(73) Titular/es:

STICHTING VUMC (100.0%)
De Boelelaan 1117
1081 HV Amsterdam, NL

(72) Inventor/es:

DUBBELBOER, FINN y
HOUTGAST, TAMMO

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Supresión de ruido en señales de voz

La invención se refiere a un procedimiento y aparato para procesar señales de voz.

La Patente Estadounidense 5,133,013 describe la supresión del ruido en señales que contienen voz. Como es bien sabido, se puede emplear un filtro Wiener para suprimir el ruido. Un filtro Wiener suprime cada vez más los componentes espectrales cuando contienen relativamente más ruido y menos señal real. Los coeficientes de filtro del filtro Wiener se seleccionan para minimizar la desviación media cuadrática esperada entre la señal filtrada y un componente nocional libre de ruido de la señal de entrada. Esto resulta en un filtro que multiplica cada componente espectral de la señal de entrada con un factor de supresión S/(S+N) que es proporcional a la relación de la densidad espectral esperada S de la señal libre de ruido y la densidad espectral esperada (S+N) de la señal de entrada con ruido a la frecuencia del componente espectral. Sin embargo, para aplicar el filtrado Wiener, se necesita una estimación confiable de la densidad espectral de señal a ruido.

También se conoce el uso de densidades espectrales estimadas dinámicamente en el cálculo del factor de supresión. En este caso, la densidad espectral esperada (S+N) de la señal de entrada con ruido se reemplaza por una densidad I espectral calculada de la señal de entrada en algún intervalo de tiempo, y la densidad espectral S de la señal libre de ruido se determina restando una densidad espectral N esperada del ruido de la densidad I espectral calculada de la señal de entrada.

En efecto, esto da como resultado un filtro no lineal, que pasa idénticamente los componentes espectrales de la señal de entrada con una gran densidad I espectral y atenúa las señales de entrada cuando la densidad I espectral está cerca o por debajo de la densidad espectral N del ruido. La Patente Estadounidense US 5,133,013 utiliza el término “rodilla” para la transición entre señales de paso de forma idéntica y señales de paso atenuadas. La Patente Estadounidense No. 5,133,013 señala que se puede utilizar un filtro no lineal que suprime completamente los componentes espectrales con una densidad espectral debajo de la rodilla. Sin embargo, dicho filtro se rechaza porque introduce una distorsión inaceptable. En cambio, se utiliza una supresión más gradual, que se aproxima a la supresión completa, a lo largo de las líneas de un filtro de Wiener. La Patente Estadounidense US 5,133,013 utiliza una posición predeterminada de la rodilla y adapta la ganancia de la señal de entrada para garantizar que la rodilla se encuentre a aproximadamente al nivel de ruido.

El documento EP 661689 describe un procedimiento de procesamiento de señal de voz telefónica en el que los factores de supresión se seleccionan para los respectivos marcos de tiempo y la señal de voz completa en los marcos de tiempo, o para una parte de frecuencia alta o baja de la señal de voz. El documento EP 661689 propone pasar la señal de voz de manera idéntica cuando su amplitud media está por encima de un primer umbral, y aplicar un factor de supresión cada vez más pequeño, que es inversamente proporcional a la amplitud media cuando la amplitud media está por debajo del primer umbral. El documento EP 661689 menciona que el factor de supresión puede mantenerse constante cuando la amplitud media está por debajo de un segundo umbral, que es más pequeño que el primer umbral. Esto se dice para evitar una supresión de ruido demasiado intensa para pequeños ruidos.

Aunque dichas técnicas reducen la desviación cuadrática media matemáticamente esperada entre la señal filtrada y un componente nocional libre de ruido de la señal de entrada, se ha descubierto que se limita su efecto sobre la inteligibilidad del habla. En algunos casos, la inteligibilidad apenas cambió, a pesar de que mejoró la relación señal a ruido.

Una posible explicación para esto podría ser que los procedimientos conocidos de supresión de ruido introducen artefactos que pueden percibirse como de tipo habla, al tiempo que suprimen el ruido que, en cualquier caso, puede distinguirse por el sistema auditivo humano.

Entre otros, es un objetivo mejorar la inteligibilidad de las señales de voz. El objetivo de la presente invención se alcanza mediante las reivindicaciones independientes. Las realizaciones específicas se definen en las reivindicaciones dependientes.

Un aparato de procesamiento de voz de acuerdo con la reivindicación 1. Aquí se utiliza un factor de ajuste de amplitud con un primer o segundo valor, dependiendo de la intensidad de la señal, con una transición aguda entre el primer y el segundo valor en función de la intensidad de la señal. Por lo tanto, el número de componentes espectrales con factores de ajuste mutuamente diferentes se mantiene en un mínimo, de modo que los errores en las fluctuaciones de intensidad de señal tienen un efecto mínimo. Se ha descubierto que esto aumenta la inteligibilidad.

Estos y otros objetivos y aspectos ventajosos se harán evidentes a partir de una descripción de realizaciones a modo de ejemplo, utilizando las siguientes figuras.

La figura 1 muestra un aparato de procesamiento de voz

La figura 2 muestra una función de ganancia

La figura 3 muestra un selector de factor

La figura 1 muestra un aparato de procesamiento de voz, que comprende un micrófono 10, un filtro 11, un selector 14 de factor y un dispositivo 19 de salida. El filtro 11 comprende un analizador 12 de frecuencia, un multiplicador 16 y un sintetizador 18. El micrófono 10 tiene una salida acoplada a una entrada del analizador 12 de frecuencia. El selector 14 de factor tiene una entrada acoplada a una salida del analizador 12 de frecuencia. El multiplicador 16 tiene una primera entrada acoplada a la salida del analizador 12 de frecuencia y una segunda entrada acoplada a una salida del selector 14 de factor. El multiplicador 16 tiene una salida acoplada al sintetizador 18, que tiene una salida acoplada al dispositivo 19 de salida.

En funcionamiento, el micrófono 10 capta una señal de voz que puede contener ruido adicional. El analizador 12 de frecuencia analiza la señal de voz en una pluralidad de componentes para bandas de frecuencia respectivas. Se puede utilizar procesamiento digital, la señal de voz se digitaliza antes del análisis real. El análisis de frecuencia se puede realizar tomando muestras de señales de voz digitalizadas para una ventana de tiempo en la señal de voz y calculando su transformada de Fourier. El multiplicador 16 multiplica los componentes cada uno por un factor respectivo. El multiplicador 16 puede configurarse para realizar las multiplicaciones sucesivamente para diferentes frecuencias en los resultados de la transformada de Fourier para la ventana, por ejemplo. El sintetizador 18 reensambla los componentes de señal multiplicados y el dispositivo 19 de salida emite la señal reensamblada para que la utilice un oyente humano.

El selector 14 de factor selecciona los factores utilizados por el multiplicador 16. En una realización, el selector 14 de factor selecciona el factor para cada componente basándose en el valor absoluto del componente, utilizando un factor de uno si el valor absoluto excede un umbral T y un valor F que es menor que uno si el valor absoluto no supera el umbral.

La figura 2 ilustra el factor que se selecciona mediante el selector 14 de factor como una función del valor absoluto del componente como una línea continua. Como referencia, un factor típico en función del valor absoluto según un filtro Wiener se muestra como una línea discontinua. Puede observarse que la relación utilizada por el selector 14 de factor asegura que se mantenga la intensidad relativa de diferentes componentes de señal por debajo del umbral. En particular, la intensidad relativa de estos componentes no es sensible al ruido, ya que no depende de las estimaciones de la amplitud de la señal. Además, las variaciones temporales del factor para un componente espectral, debido a las fluctuaciones en la intensidad de señal estimada en el componente espectral se evitan para intensidades de señal pequeñas. Por lo tanto, se minimiza la introducción de artefactos de tipo habla, como la modulación de ruido. La intensidad relativa de diferentes componentes de señal por encima del umbral también se conserva, pero estas resistencias ya eran menos sensibles al ruido en las amplitudes de señal estimadas. Solo se afecta la intensidad relativa de los componentes con amplitudes en diferentes lados del umbral.

Además, se puede observar que esta relación entre el factor y el valor absoluto del componente presenta una discontinuidad en el umbral T. Aunque tal discontinuidad puede presentar algunos artefactos, se ha descubierto que, a los fines de la inteligibilidad, es más eficaz para aceptar esto que para introducir diferencias de factor sensibles al ruido entre diferentes componentes espectrales mediante el uso de una transición más gradual. Para la inteligibilidad, es más efectivo minimizar el número de cambios de amplitud relativa entre los diferentes componentes.

La figura 3 muestra una realización del selector 14 de factor. En esta realización, el selector de factor comprende un detector de amplitud 30, un promediador 32, un detector 34 de nivel de ruido, un fijador de fijador de umbral 36 y una unidad de suministro de factor 38. El detector de amplitud 30 tiene una entrada para recibir las señales componentes del analizador de frecuencia (no mostrado). El promediador 32 tiene una entrada acoplada a una salida del detector de amplitud 30 y una salida acoplada al fijador de umbral 36. El fijador de umbral 36 tiene una salida acoplada a una entrada de control de selección de la unidad de suministro de factor 38, que tiene una salida acoplada a la segunda entrada del multiplicador (no mostrada). La unidad de suministro de factor 38 está configurada para suministrar un factor de uno o F dependiente del resultado del umbral. El detector 34 de nivel de ruido está acoplado entre el detector de amplitud y el fijador de umbral 36.

El promediador 32 calcula promedios para cada componente espectral en puntos de tiempo respectivos, promediando puntos de tiempo cercanos y frecuencias cercanas. En un ejemplo, en el que el analizador de frecuencia emite componentes espectrales para los marcos de tiempo respectivos, el promedio puede tomarse sobre los cuadrados absolutos de los componentes espectrales para las frecuencias N1 más cercanas a cada lado de la frecuencia para la que se calcula el promedio y esa frecuencia en sí misma. De manera similar, el promedio puede tomarse sobre los componentes para 2*N2 que preceden a los marcos de tiempo, o N2 marcos precedentes y N2 marcos siguientes. Este promedio se puede calcular como un promedio continuo, utilizando el promedio calculado para el marco de tiempo anterior.

El detector 34 de nivel de ruido determina el nivel de umbral para la amplitud de señal media a partir de una estimación del nivel de ruido. En una realización, el detector de ruido detecta marcos temporales en los que se presenta ruido, pero no habla y calcula amplitudes medias del ruido para componentes espectrales respectivos en esos marcos de tiempo de

una manera similar a la que el promediador 32 calcula las amplitudes de señal promedio de los componentes espectrales. Los detectores de voz/ruido son conocidos per se. En esta realización, el umbral para cada componente espectral se establece como un factor multiplicado por el ruido promedio calculado para el componente espectral. En una realización, esto tiene el efecto de comparar un umbral T independiente de la frecuencia con una cantidad calculada

$$(<|Y|^2> - <|N|^2>)/<|N|^2>$$

donde los paréntesis denotan promediado (no necesariamente sobre la misma ventana promedia para Y y N), $|Y|^2$ denota la amplitud al cuadrado de los componentes espectrales de la señal y $|N|^2$ denota la amplitud al cuadrado de la señal en marcos de tiempo donde se ha detectado que el habla está ausente.

Como puede observarse, esta técnica requiere la selección de solo un número limitado de parámetros de diseño: el umbral T, el factor F y los números de componentes espectrales N1, N2 utilizados para promediar la amplitud de la señal. Estos parámetros pueden elegirse libremente. Por ejemplo, estos parámetros pueden establecerse experimentalmente, escuchando la voz producida utilizando valores de parámetros específicos y variando los valores de los parámetros para optimizar la inteligibilidad. En un experimento, se obtuvo una inteligibilidad mejorada cuando el umbral T se estableció en 1, F se estableció en 0,5 y N1 se estableció en 1. El resultado podría optimizarse variando N2. Se encontró que se produjo un óptimo pronunciado para N2 a aproximadamente 9.

Sorprendentemente, se encontró que el valor de T para una inteligibilidad óptima variaba con el valor seleccionado para N2. Cuando N2 aumenta, la potencia del ruido se acerca cada vez más a su valor esperado, con el efecto de que se reduce el riesgo de supresión involuntaria del habla. De acuerdo con los anterior, T puede establecerse más bajo. Se encontró que el valor óptimo de T varía con el logaritmo de N2. Se encontró una relación experimental aproximadamente de acuerdo con

$$T = 10^{10} \log 9/N2$$

Sin embargo, incluso sin seleccionar dichos valores óptimos, se encontró un aumento de la inteligibilidad, tanto para las personas con audición normal como para las personas con defectos auditivos. El factor F se puede establecer más bajo o más alto, por ejemplo, en cualquier parte del rango de 0,1 a 0,8 y se pueden utilizar valores mayores de N1. Preferiblemente, se utiliza un factor distinto de cero, para evitar que los componentes espectrales con ruido fuerte y algún componente de voz se supriman por completo. Por lo tanto, no se evita que cerebro haga la recuperación contextual del componente del habla.

Aunque se ha mostrado una realización específica a modo de ejemplo, debe tenerse en cuenta que en la práctica son posibles muchas variaciones. Por ejemplo, el filtro 11 puede implementarse de diferentes maneras. En lugar de análisis y síntesis con multiplicación intermedia, se puede utilizar una convolución temporal, utilizando coeficientes de filtro determinados a partir de los factores de ajuste espectral. En lugar de análisis por transformación de Fourier, se puede utilizar un banco de filtros con filtros para las respectivas bandas de frecuencia. En lugar de multiplicar componentes espectrales (es decir, números complejos que tienen una amplitud y una fase), las amplitudes de los componentes espectrales pueden extraerse, multiplicarse con los factores y recombinarse con la fase. En lugar de las amplitudes, los cuadrados de las amplitudes se pueden multiplicar con factores modificados correspondientemente. El fijador de umbral 36 puede calcular el umbral a partir de la intensidad del ruido, o de forma equivalente, la intensidad del ruido y la intensidad de la señal puede usarse para calcular una relación señal a ruido, que posteriormente se compara con un umbral.

El filtro 11 y el selector 14 de factor pueden implementarse por medio de un circuito de computadora programable tal como un circuito de procesador de señal programable, programado con un programa que hace que el ordenador realice las funciones descritas. Alternativamente, todo o parte del filtro 11 y el selector de factores 14 pueden implementarse como circuitos de hardware dedicados, diseñados para realizar las funciones descritas.

REIVINDICACIONES

1. Un aparato de procesamiento de voz que comprende

- un filtro (11) configurado para ajustar una señal de voz de entrada con un factor de ajuste;

- un selector (14) de factor para seleccionar el factor de ajuste dependiente de la señal de voz de entrada, el selector (14) de factor se configura para establecer el factor a un primer valor distinto de cero cuando un promedio de intensidad está por encima de un valor umbral caracterizado porque el filtro (11) se configura para ajustar un envolvente espectral de la señal de voz de entrada, el factor de ajuste depende de la frecuencia, se configura el selector (14) de factor para seleccionar el factor de ajuste para los respectivos componentes espectrales cada uno dependiente de la señal de voz de entrada, el selector (14) de factor se configura para establecer el factor en el primer valor o en un segundo valor distinto de cero, cuando un promedio de intensidad para el componente espectral está por encima y por debajo de un valor umbral respectivamente, siendo el segundo valor menor que el primer valor .

2. Un aparato de procesamiento de voz según la reivindicación 1, en el que el filtro (11) se configura para calcular conjuntos de componentes espectrales para una serie de marcos de tiempo y para calcular componentes espectrales ajustados en los que los componentes espectrales han sido ajustados por los factores de ajuste. el selector (14) de factor que comprende un promediador configurado para calcular el promedio de la intensidad para el componente espectral para cada trama de tiempo promediando sobre una pluralidad de los marcos de tiempo adyacentes al marco de tiempo para el cual se calcula el promedio de la intensidad.

3. Un aparato de procesamiento de voz de acuerdo con la reivindicación 1 o 2, en el que el selector (14) de factor comprende un detector de nivel de ruido, que configura el selector (14) de factor para establecer el umbral en proporción a un nivel de ruido detectado.

4. Un aparato de procesamiento de voz de acuerdo con la reivindicación 2, en el que el selector (14) de factor comprende un detector de nivel de ruido, el selector (14) de factor se configura para establecer el umbral en proporción a un nivel de ruido detectado, con un factor de proporcionalidad aproximadamente igual a 10x10log 9/N2.

5. Un aparato de procesamiento de voz de acuerdo con una cualquiera de las reivindicaciones precedentes, en el que el promedio de intensidad es un promedio de cuadrados de amplitudes de los componentes espectrales.

6. Un procedimiento de procesamiento de una señal de voz, el procedimiento comprende

- ajustar una señal de voz con un factor de ajuste;

- seleccionar el factor de ajuste dependiente de la señal de voz de entrada, el factor de ajuste se establece en un primer valor distinto de cero, cuando un promedio de intensidad está por encima de un valor umbral, caracterizado porque el factor de ajuste depende de la frecuencia, una envolvente espectral del señal de voz que se ajusta con el factor de ajuste dependiente de la frecuencia, y porque el factor de ajuste se establece en el primer o segundo valor distinto de cero, cuando un promedio de resistencia para el componente espectral es superior e inferior a un valor umbral respectivamente, el segundo valor es más pequeño que el primer valor.

7. Un producto de programa informático, que comprende un programa de instrucciones para una computadora programable, que, cuando es ejecutado por el ordenador, hace que el ordenador realice el procedimiento de la reivindicación 6.

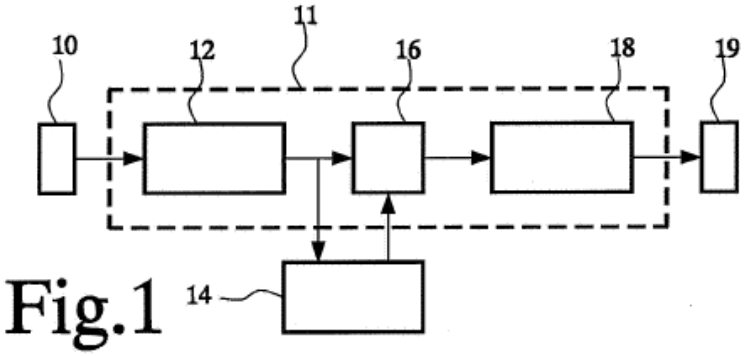

Fig.1

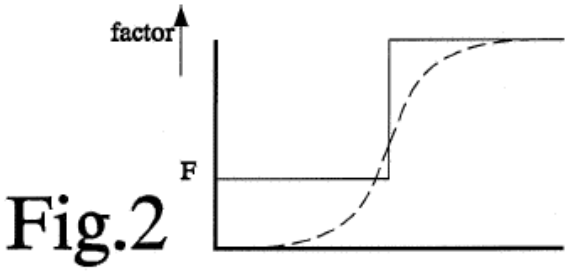

Fig.2

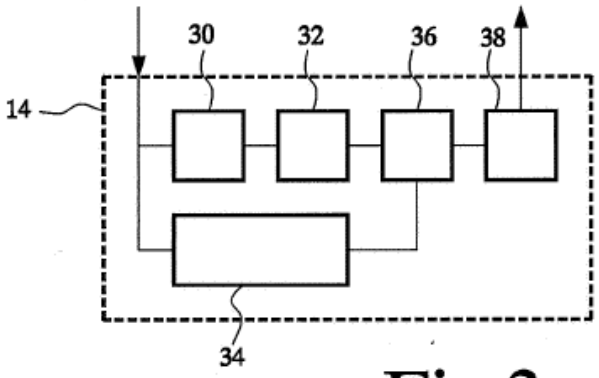

Fig.3