(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 547 731**

(51) Int. CI.:

G10L 17/02 (2013.01)
G10L 17/06 (2013.01)
G10L 17/20 (2013.01)
G10L 25/15 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **03.11.2010 E 10823678 (7)**

(97) Fecha y número de publicación de la concesión europea: **24.06.2015 EP 2482277**

(54) Título: **Procedimiento para identificar a un hablante usando ecualización de formante**

(30) Prioridad:

24.09.2009 RU 2009136387

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
08.10.2015

(73) Titular/es:

SPEECH TECHNOLOGY CENTER LIMITED (100.0%)
4 Krasutskogo street, lit. A
St. Petersburg, 196084, RU

(72) Inventor/es:

KOVAL, SERGEY LVOVICH

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

Observaciones :

Véase nota informativa (Remarks) en el folleto original publicado por la Oficina Europea de Patentes

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Procedimiento para identificar a un hablante usando ecualización de formante

Campo técnico

La invención se refiere al campo de reconocimiento de voz de hablante, particularmente a procedimientos automáticos, automatizados y expertos para la identificación de un hablante a partir de grabaciones de audio de habla oral espontánea. Los procedimientos están destinados a, pero no se limitan a, examen forense.

Se conocen investigaciones prácticas, particularmente el examen forense y la comparación de grabaciones de audio de habla oral para identificar un hablante, en algunos casos para adoptar decisiones difíciles que presentan obstáculos a la evaluación experta. Dichos obstáculos pueden incluir una duración corta y una baja calidad de las grabaciones de audio examinadas, estados psico-fisiológicos diferentes de los hablantes en las grabaciones de audio comparadas, contenidos y lenguajes del habla diferentes, tipos y niveles de ruido y distorsión diferentes del canal de audio, etc.

Por lo tanto, es deseable proporcionar procedimientos para la identificación de un hablante a partir de grabaciones de audio de habla oral espontánea, en los que los procedimientos sean aplicables a, pero no se limiten a, examen forense y tengan en cuenta los problemas indicados anteriormente, En particular, es deseable proporcionar procedimientos que permitan la identificación de un hablante a partir de grabaciones de audio cortas de habla oral grabadas en diversos canales de grabación de audio, con altos niveles de ruido y distorsión, así como a partir de grabaciones de audio con hablantes en estados pisco-fisiológicos diferentes, con contenidos verbales diferentes del habla oral, o habla oral en idiomas diferentes.

Técnica anterior

Se conoce un procedimiento de identificación de un hablante a partir de grabaciones de audio, en el que las características del hablante son extraídas a partir de ciertas frases de tipo único pronunciadas por el hablante (DE 2431458, 2/05/1976).

Este procedimiento comprende el filtrado de una señal de voz a través de un peine de 24 filtros de paso de banda, la detección de la señal, el suavizado de la señal y la introducción de la señal a través de un convertidor analógico-digital y un conmutador a un dispositivo de procesamiento digital, en el que las características distintivas asociadas con el espectro de voz integral son reconocidas y almacenadas automáticamente.

Sin embargo, este procedimiento no puede ser usado con grabaciones de audio de habla oral obtenidas en un entorno de alta distorsión y ruido ya que el procedimiento no puede proporcionar un número suficiente de características distintivas. Además, no se ha probado que este procedimiento sea suficientemente fiable para la identificación, ya que requiere el uso de grabaciones de audio que comprenden contenido verbal idéntico tanto para un hablante verificado como para un hablante desconocido.

Se conoce un procedimiento de identificación de una grabación de audio individual mediante una entrada de voz que es comparada con una firma de voz almacenada anteriormente de ese individuo. La identificación se basa en la selección y la comparación de palabras clave de tipo único a partir de las grabaciones objeto de análisis (US 3.466.394).

Este procedimiento consiste en someter una señal de voz a un análisis espectral a corto plazo y, a continuación, determinar los contornos del espectro y el tono de voz fundamental en función del tiempo. Los contornos resultantes son considerados como características distintivas. La identificación de un hablante se basa en una comparación de los contornos de grabación de audio de las grabaciones de audio obtenidas a partir de un hablante verificado y un hablante desconocido.

El punto débil de este procedimiento es que el resultado de la identificación depende de la calidad de grabaciones de audio obtenidas en un entorno de alta distorsión y ruido. Además, este procedimiento tiene un alto porcentaje de fallos de identificación, ya que requiere grabaciones de audio de un hablante verificado y un hablante desconocido con las mismas palabras.

Se conoce un procedimiento de identificación de un hablante basado en un análisis espectral-banda-temporal del habla oral espontánea (G.S. Ramishvili, G.B. Chikoidze Forensic examination of audio records and identification of a speaker. Tbilisi: "Meznie reba", 1991, p. 265).

Para eliminar la dependencia de los resultados de identificación de la semántica del habla, los elementos de habla sonora son seleccionados de la señal de voz, y sus valores de energía para los formantes superiores son promediados a lo largo de su vida en cada uno de los 24 filtros espectrales. El tono fundamental es determinado en base a la selección de la componente fundamental de la señal en el espectro. Se determina también el ritmo del

habla.

Los parámetros indicados anteriormente son usados como características distintivas.

Sin embargo, este procedimiento falla con grabaciones de audio obtenidas en entornos de alta distorsión de sonido en un canal de grabación de ruido y diferentes estados de los hablantes debido a que las características distintivas no pueden ser seleccionadas de manera fiable.

Se conocen un dispositivo y un procedimiento para el reconocimiento de un hablante, basados en la construcción y la comparación de modelos puramente estadísticos para características de señal cepstral de habla de hablantes conocidos y desconocidos (US 6.411.930). El reconocimiento del hablante se realiza mediante el uso de modelos Gaussianos discriminativos mixtos.

Este procedimiento, al igual que la mayoría de los enfoques puramente estadísticos para el reconocimiento del hablante, falla para mensajes de habla muy cortos (de 1 a 10 segundos), así como en las situaciones en las que los estados del hablante y/o los canales de grabación de las grabaciones de audio poseen propiedades fuertemente diferentes, o los hablantes están en estados emocionales diferentes.

Se conoce un procedimiento para el reconocimiento del hablante mediante el uso de sólo un enfoque estocástico (US 5.995.927).

Según el procedimiento, el reconocimiento del hablante se realiza mediante la construcción y la comparación de matrices de covarianza de descripción de características de una señal de voz de entrada y señales de voz de referencia de hablantes conocidos.

Este procedimiento falla también para mensajes de habla cortos (5 segundos o menos), y es muy sensible a una reducción significativa de la potencia de la señal en segmentos particulares del rango de frecuencias del habla debida al ruido ambiente, así como a una mala calidad de los micrófonos y los canales para la transmisión y la grabación de sonido.

Se conoce un procedimiento adaptable al hablante para el reconocimiento de palabras aisladas (RU 2047912). El procedimiento se basa en el muestreo de una señal de voz de entrada, sucesivamente la pre-enfatización de la señal de voz, la segmentación posterior de la señal de voz, la codificación de los segmentos con elementos discretos, el cálculo del espectro de energía, la medición de las frecuencias de los formantes y la determinación de las amplitudes y las energías en diferentes bandas de frecuencia de la señal de voz, la clasificación de los eventos y los estados articulatorios, la definición y la clasificación de las muestras de palabras, el cálculo de los intervalos entre las muestras de palabras para actualizar la palabra reconocida, el reconocimiento o el fallo en el reconocimiento de la palabra, y la adición al diccionario de muestras durante la adaptación al hablante. La señal de voz de entrada es pre-enfatizada en el dominio del tiempo mediante suavización/diferenciación. El espectro de energía es cuantificado en partes discretas en función de la varianza del ruido del canal de comunicación. Las frecuencias de los formantes son determinadas buscando el máximo global del espectro logarítmico y restando una función predeterminada dependiente de la frecuencia del espectro indicado anteriormente. Durante la clasificación de eventos y estados articulatorios, se determinan las proporciones de fuentes de excitación periódicas y de ruido en comparación con el valor umbral de los coeficientes de autocorrelación de una secuencia de pulsos de onda cuadrada en múltiples bandas de frecuencia. El principio y el final de los movimientos articulatorios y sus procesos acústicos correspondientes se determinan contra el valor umbral de la función de probabilidad a partir de los coeficientes de autocorrelación, las frecuencias de los formantes y las energías en las bandas de frecuencia determinadas. La señal de voz es dividida en intervalos de entre el principio y el final de los procesos acústicos correspondientes a movimientos articulatorios específicos, y de manera secuencial, empezando con vocales. Un segmento es reconocido sólo cuando sus tipos de transición de límite izquierdo y derecho coinciden entre sí, mientras que la segmentación se termina cuando se reconocen pausas entre palabras en los segmentos de tiempo izquierdo y derecho. Las muestras de palabras se forman como matrices con valores de probabilidad de características binarias, y el reconocimiento falla cuando la diferencia de intervalo normalizado entre la actualización desconocida y las los dos muestras más próximas que pertenecen a palabras diferentes es menor que el valor umbral establecido.

La desventaja de este procedimiento conocido adaptable al hablante para el reconocimiento de palabras aisladas es su mala distinción cuando intenta reconocer hablantes por medio de habla espontánea, ya que en la mayoría de los casos el procedimiento no distingue entre hablantes del mismo sexo que transfieren un mensaje verbal con el mismo contenido verbal.

Se conoce un sistema de seguridad basado en el reconocimiento de voz (US 5.265.191), que requiere que tanto el entrenador como el hablante desconocido repitan al menos un mensaje de voz. El sistema compara las representaciones paramétricas de los mensajes de voz repetidos realizados por el hablante desconocido y por el hablante conocido y establece la identidad de los hablantes comparados sólo si cada mensaje pronunciado por el

hablante desconocido es suficientemente cercano al mensaje realizado por el entrenador, indicando un fallo si sus representaciones difieren fuertemente entre sí.

El punto débil de este sistema es su pobre resistencia a los ruidos variables (ruidos de vehículos y urbanos, locales industriales), así como el requisito obligatorio de que ambos hablantes pronuncien un mismo mensaje de voz.

Se conoce un procedimiento para la identificación automática de un hablante mediante las peculiaridades de la pronunciación de la frase de contraseña (RU 2161826). El procedimiento comprende dividir la señal de voz en zonas sonoras y definir intervalos de tiempo dentro de las zonas indicadas anteriormente, en el máximo de intensidad de la señal de voz, así como al principio de la primera zona sonora y al final de la última zona sonora. Para los intervalos de tiempo definidos, se determinan los parámetros de la señal de voz y se comparan con las muestras formadas teniendo en cuenta las esperanzas matemáticas y las variaciones aceptables de los parámetros. Al hacerlo, se definen intervalos de tiempo al final de la primera zona, al principio de la última zona sonora, y al principio y al final de las otras zonas; y la duración de los intervalos de tiempo se establece como un múltiplo del periodo de tono fundamental de la señal de voz. Se determinan los coeficientes de correlación de los parámetros de la señal de voz, teniendo en cuenta los coeficientes de correlación cuando se forman las muestras. La identificación de un hablante se basa en los parámetros de señal de voz obtenidos y en las características estadísticas correspondientes.

La desventaja de este procedimiento conocido es su pobre resistencia al ruido, ya que requiere la determinación de la posición exacta de los límites del periodo de tono de voz fundamental en la señal de voz de entrada, lo cual frecuentemente no es posible con interferencias acústicas y electromagnéticas (ruido de oficina y de la calle, interferencia del canal de voz, etc.). Además, los hablantes deben pronunciar las mismas contraseñas de voz, lo cual no puede conseguirse siempre en la práctica.

Se conoce un dispositivo de verificación de hablante basado en la medida de la distancia al "vecino más cercano" (US 5.339.385), que incluye una pantalla, un generador de indicaciones aleatorias, una unidad de reconocimiento de voz, un verificador de hablante, un teclado y un procesador de señal principal, en el que la entrada al procesador de señal principal es la entrada del dispositivo y su salida está conectada a la primera entrada de la unidad de reconocimiento de voz y la primera entrada al verificador de hablante. La primera salida del generador de indicaciones está conectada a la segunda entrada de la unidad de reconocimiento de voz, cuya salida está conectada a la pantalla. El teclado está conectado a la tercera entrada de la unidad de reconocimiento de voz y la tercera entrada del verificador de hablante, cuya salida es la salida del dispositivo. Para establecer las similitudes o las diferencias en las contraseñas de voz pronunciadas, el verificador de habla divide la señal de voz de entrada en tramas de análisis específicas, calcula los vectores de habla no paramétricos para cada trama de análisis y además determina la proximidad de las descripciones de la señal de voz obtenidas de esta manera de las pronunciaciones comparadas en base a la distancia euclidiana al vecino más cercano.

La desventaja de este dispositivo de verificación de hablante es su pobre resistencia al ruido en entornos de oficina y urbanos debido al uso de vectores de habla no paramétricos y las métricas euclidianas en la determinación del grado de similitud/diferencia en las pronunciaciones de una contraseña de voz, así como una baja fiabilidad del reconocimiento (gran porcentaje de fallos falsos) debido al uso de contraseñas de voz con diferente orden de palabras. Dicha baja fiabilidad es causada por la inevitable variabilidad individual en la pronunciación de las mismas palabras en contextos diferentes, incluso por el mismo hablante. Además, es difícil asegurar la pronunciación del contenido verbal propuesto por ambos hablantes comparados.

Se conoce un procedimiento para el reconocimiento de un hablante (US 6.389.392), en el que el procedimiento comprende comparar la señal de voz de entrada obtenida a partir de un hablante desconocido con muestras de voz de hablantes conocidos anteriormente, de entre los cuales al menos uno está representado por dos o más muestras. Los segmentos de señal de entrada sucesivos se comparan con segmentos de muestras sucesivos para obtener una medida de la proximidad del segmento de señal de voz de entrada y el segmento de señal de voz de muestra comparados. Para cada muestra de un hablante conocido con al menos dos muestras, se forma un resultado de comparación compuesto de la muestra y la señal de voz de entrada en base a la selección, para cada segmento de la señal de voz de entrada, del segmento más cercano de la muestra comparada en términos de la medida de proximidad usada. A continuación, el hablante desconocido es reconocido en base a los resultados de comparación compuestos de la señal de voz de entrada y la muestra.

Este procedimiento conocido de reconocimiento de hablante está limitado en la aplicación práctica, ya que el requisito de que al menos dos muestras para cada mensaje verbal sean pronunciadas por un hablante conocido no siempre es factible en el entorno real. Además, este procedimiento no garantiza una alta fiabilidad de reconocimiento de hablante en el entorno de ruido acústico de una oficina, calle o vehículo reales, o con estados emocionales diferentes de los hablantes, ya que la descripción de señal de voz paramétrica segmento a segmento está sujeta a una fuerte influencia del ruido acústico aditivo y la variabilidad natural del habla. Además, la baja fiabilidad del procedimiento en un entorno de alto ruido surge del hecho de que debe encontrarse el segmento de

muestra más cercano en términos de la medida de proximidad usada para cada segmento de la señal de voz de entrada, lo que conlleva un gran número de segmentos de puro ruido correspondientes a segmentos de pausas de voz tanto en la muestra como en la señal de voz de entrada.

Se conoce un procedimiento para el reconocimiento del hablante a partir de grabaciones de audio de habla oral espontánea (RU 2107950). El procedimiento se basa en el uso de un análisis espectral-banda-temporal de señales de voz, la determinación de las características distintivas del habla de un individuo y su comparación con muestras usando las características integrales acústicas como características distintivas. Las características integrales acústicas son estimaciones de los parámetros de la distribución estadística de los componentes del rango actual y el período de tono primario y los histogramas de distribución de frecuencias. Los parámetros son medidos sobre grabaciones de audio con contextos tanto espontáneos como fijos. Entre las características integrales acústicas, las características más informativas para un hablante determinado, no influenciadas por el ruido y la distorsión presentes en las grabaciones de audio son seleccionadas durante el transcurso del re-entrenamiento adaptativo. También se usan características lingüísticas (las fijas o espontáneas), detectadas por un experto durante el transcurso del análisis auditivo de las grabaciones de audio usando un banco automatizado que comprende muestras de voz de referencia de dialectos, acentos y defectos del habla oral.

Este procedimiento pierde su fiabilidad para hablantes con grabaciones de audio cortas, que hablan diferentes idiomas o que están en estados psico-fisiológicos sustancialmente diferentes debido al empleo de un enfoque integral, promediando las características de la señal de voz y el análisis lingüístico.

Se conoce un procedimiento para el reconocimiento de hablante (RU 2230375) con el mayor número de características comunes con el procedimiento reivindicado y seleccionado como la técnica anterior más próxima. El procedimiento del documento RU 2230375 incluye una comparación segmento a segmento de la señal de voz de entrada con muestras de contraseñas de voz pronunciadas por hablantes conocidos, y una evaluación de la similitud entre una primera grabación de audio del hablante, y una segunda grabación de audio o muestra haciendo coincidir las frecuencias de los formantes en los fragmentos de referencia de la señal de voz, en el que los fragmentos para la comparación son seleccionados de entre la primera grabación de audio y la segunda grabación de audio o muestra.

El procedimiento conocido identifica los vectores de formantes de los segmentos consecutivos y las características estadísticas de los espectros de potencia de la señal de voz de entrada y de las señales de voz de muestra. Los vectores de formantes y las características estadísticas se comparan adicionalmente con los vectores de formantes de los segmentos sucesivos de cada muestra y con las características estadísticas del espectro de potencia de la señal de voz de muestra, respectivamente, para formar las métricas de comparación compuestas para las señales de entrada y las señalas de muestra. Se usa un módulo ponderado de una diferencia de frecuencia del vector de formantes como una medida de la proximidad del vector de formantes en los segmentos.

Para calcular una métrica de comparación compuesta para la señal de entrada y la muestra, cada segmento de señal de voz de entrada recibe el segmento de muestra más cercano en términos de la medida de proximidad correspondiente con el mismo número de formantes. Las métricas compuestas incluyen un promedio ponderado de la medida de proximidad entre un segmento de señal de voz de entrada determinado y el segmento de muestra más cercano para todos los segmentos de señal de voz de entrada usados, así como un coeficiente de correlación cruzada de las características estadísticas de los espectros de potencia de la señal de voz de entrada y la muestra. El reconocimiento de hablante se basa en el resultado de la comparación de la señal de voz de entrada y la muestra usando las métricas compuestas.

Este procedimiento no proporciona un reconocimiento de hablante fiable cuando la estructura fónica de la señal de voz de entrada difiere fuertemente de la estructura fónica de las muestras de señal de voz (por ejemplo, mensajes cortos, idiomas diferentes de la señal de entrada y las muestras), así como en caso de diferencias significativas en las propiedades de los canales de grabación y diferencias en el estado psico-fisiológico de los hablantes en las grabaciones de audio comparadas. Estas deficiencias surgen, para empezar, debido al uso de las características estadísticas del espectro de potencia como un componente de las métricas compuestas, ya que las características estadísticas dependen en gran medida de las propiedades del canal de grabación, el estado del hablante y la estructura fónica del mensaje, y debido a la medida segmental de la proximidad en la forma de un promedio ponderado para todos los segmentos usados de la señal de voz procesada, lo que conduce a promediar los errores de comparación de segmentos y a subestimar la influencia de las grandes desviaciones entre segmentos, indicando la diferencia entre los hablantes incluso cuando se observa una pequeña diferencia media de segmentos.

Otros ejemplos de la técnica anterior se encuentran en los documentos 2005/171774 A1 (4/08/2005), TANABIAN M M ET AL: "Automatic speaker recognition with formant trajectory tracking using CART and neural networks", IEEE CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING, 2005, SAKATOON, SK, CANADÁ, MAYO 1-4, 2005, páginas 1.225-1.228, y EP 0 825 587 A2 (25/02/1998).

Sumario de la invención

Un objetivo de la presente invención, tal como se define en las reivindicaciones adjuntas, es proporcionar un procedimiento para la identificación de un hablante a partir de grabaciones de audio de habla oral espontánea, en el que se seleccionan fragmentos de la señal de voz para la comparación, y se usan características de individualización y procedimientos de comparación de las mismas para reconocer el hablante, permitiendo una identificación fiable del hablante en la mayoría de los casos prácticos, en particular, tanto para grabaciones de audio largas y cortas a ser comparadas, para grabaciones de audio grabadas en canales diferentes que tienen un alto nivel de ruido y distorsión, para grabaciones de habla oral espontánea de hablantes en estados psico-fisiológicos diferentes y/o en idiomas diferentes, con el fin de aplicar más ampliamente el procedimiento, incluyendo su uso en un examen forense.

Este objetivo se consigue proporcionando un procedimiento para la identificación del hablante a partir de grabaciones de audio de habla oral, que comprende la evaluación de la similitud entre una primera grabación de audio del hablante y una segunda grabación de audio o muestra, por una coincidencia de las frecuencias de los formantes en los fragmentos de referencia de la señal de voz, seleccionados para la comparación a partir de la primera grabación de audio y de la segunda grabación de audio o muestra. La invención se caracteriza por la selección de fragmentos de referencia de señales de voz a partir de la primera grabación de audio y de la segunda grabación de audio, de manera que los fragmentos de referencia comprendan trayectorias de los formantes de al menos tres frecuencias de los formantes; la comparación entre sí de los fragmentos de referencia seleccionados a partir de las grabaciones de audio primera y segunda que comprenden al menos dos frecuencias de los formantes coincidentes; la evaluación de la similitud de los fragmentos de referencia comparados a partir de la coincidencia de las otras frecuencias de los formantes, en el que similitud de las grabaciones de audio es determinada, en general, a partir de la evaluación global de todos los fragmentos de referencia comparados.

Este procedimiento se basa en la evidencia experimental de que las resonancias de baja frecuencia del tracto vocal (formantes) de un individuo cambian su frecuencia de una manera mutuamente concordante con el cambio de articulación. En particular, esto es cierto para las tres primeras frecuencias 3 - 5 de resonancia, en función de la longitud del tracto vocal. Si se cambia la frecuencia de al menos uno de los cuatro primeros formantes, entonces, en la mayoría de los casos prácticos, es seguro que uno o más de los otros formantes de baja frecuencia del hablante cambian también su frecuencia. Este cambio es causado por la relación acústica entre las resonancias del tracto vocal y la posibilidad anatómicamente restringida de realizar ningún cambio en el área transversal del tracto vocal del hablante. Las bases teóricas de este hecho pueden encontrarse en la literatura de investigación (G. Fant, Acoustic theory of speech production. - Moscú: Nauka, 1964; V.N. Sorokin, Theory of speech production. - M.: Radio and Communication, 1985).

El procedimiento de la presente invención implica el uso de los parámetros de resonancia acústica del tracto vocal como características de individualización de los hablantes cuando se comparan las grabaciones de audio. Los parámetros de resonancia se definen como frecuencias características en las trayectorias de los formantes que determinan la calidad percibida de cada grabación de audio de voz.

Estas características son resistentes al ruido para las grabaciones de audio grabadas en un entorno de alto ruido y distorsión de la señal.

Se ha encontrado experimentalmente que la presencia de tres o más trayectorias de los formantes dentro de un espectrograma de un fragmento de referencia, en la mayoría de los casos, permite una determinación inequívoca de los valores característicos de las frecuencias de los formantes y, mediante su comparación, la identificación fiable de los hablantes.

Para proporcionar una resistencia al ruido adicional de las características definidas en un entorno de ruido aditivo de banda ancha, los valores de frecuencia de los formantes para cada fragmento de referencia seleccionado se calculan como los valores medios para los intervalos de tiempo de durabilidad fija en los que las frecuencias de los formantes son relativamente invariables.

En este punto, se usan los valores de frecuencia de los formantes fijos específicos relacionados con un intervalo de tiempo predeterminado para la comparación en cada fragmento de referencia de la señal de voz. Los valores de frecuencia de los formantes fijos se denominan vectores de formantes.

Para garantizar la fiabilidad de la identificación, es razonable comparar fragmentos de referencia en los que los valores de frecuencia de los dos primeros formantes coinciden, en los que los valores de frecuencia están dentro de la variación estándar de las frecuencias de los formantes para el tipo de sonido seleccionado a partir de un conjunto fijo de sonidos similares a vocales.

Al comparar grabaciones de audio, los fragmentos de referencia se comparan unos con otros si corresponden a segmentos de señal de voz en las grabaciones de audio que actualizan eventos articulatorios comparables, es

decir, si los sonidos pronunciados tienen estructuras de resonancia de señal altamente perceptibles con dos o más frecuencias de resonancia coincidentes (formantes), independientemente de los fonemas particulares en el mensaje de voz al que corresponden. Estos segmentos se encuentran en fragmentos de voz en los que los hablantes pronuncian fonemas tanto idénticos como diferentes.

Dichos fragmentos comparados según los valores de los formantes se conocen, según los expertos, como fragmentos de formantes iguales.

Para obtener una solución de identificación global general, se seleccionan al menos dos fragmentos de señal de voz de referencia en las grabaciones de audio, de manera que los fragmentos seleccionados estén relacionados con los sonidos de una articulación de diferencia máxima con los valores de frecuencia máximo y mínimo de los formantes primero y segundo para una grabación de audio determinada.

Para garantizar una alta fiabilidad de la identificación, los fragmentos de referencia se comparan para diversos sonidos articulados de manera tan diferente como sea posible, es decir, para las realizaciones más diferentes de la geometría del tracto vocal.

La comparación de los fragmentos de voz de formantes iguales, que corresponden a sonidos diferentes, no necesariamente idénticos, permite la identificación del hablante en grabaciones de audio con contenido fónico sustancialmente diferente, en particular, grabaciones de audio largas y cortas, grabaciones de audio realizadas por hablantes en estados psico-fisiológicos diferentes, así como hablantes de los mismos idiomas o idiomas diferentes.

Para garantizar una alta fiabilidad de la identificación del hablante con la señal de voz altamente distorsionada debido a una respuesta a la frecuencia sustancialmente irregular de los canales de audio que es diferente para las grabaciones de audio comparadas, es ventajoso, antes de calcular los valores de frecuencia de los formantes, someter el espectro de potencia de la señal de voz de cada grabación de audio a filtrado inverso. En el transcurso del filtrado inverso, se calcula el promedio temporal para cada componente de frecuencia del espectro de potencia, al menos para algunos fragmentos de la grabación de audio y, a continuación, el valor original del espectro de potencia de la señal de grabación de audio para cada componente del espectro de frecuencias es dividido por su valor medio inverso (invertido).

La división de espectro puede ser sustituida también por el filtrado inverso tomando el logaritmo de los espectros y restando el logaritmo de la potencia media del espectro para cada componente de frecuencia del logaritmo de la potencia de espectro de la señal de grabación de audio original.

La invención se describe más detalladamente a continuación mediante una realización ejemplar con referencia a los dibujos adjuntos.

Breve descripción de los dibujos

Figura 1 - un espectrograma ejemplar con las trayectorias de los formantes de la sílaba [te] superpuestas.

Figura 2-3 - comparaciones ejemplares de espectrogramas con trayectorias de los formantes de hablantes coincidentes superpuestas para los sonidos de tipo [e].

Figura 4 - una comparación ejemplar de espectrogramas con trayectorias de los formantes de hablantes coincidentes superpuestas para los sonidos de tipo [e].

Figura 5 - una comparación ejemplar de espectrogramas con trayectorias de los formantes de hablantes no coincidentes superpuestas para los sonidos de tipo [e].

Figura 6 - una comparación ejemplar de tres espectrogramas de una señal de voz de un mismo hablante (fragmento de referencia con la sílaba [na]) antes y después del filtrado inverso.

Figura 7 - una comparación ejemplar de los espectros de potencia media de la señal de voz para los tres espectrogramas en la Figura 6.

Figura 8 - una comparación ejemplar de los espectros de los vectores de formantes, definidos en los espectrogramas mostrados en la Figura 6.

Descripción detallada de la invención

El procedimiento reivindicado comprende comparar al menos dos grabaciones de audio de habla oral espontánea. En particular, para el examen forense, una de las grabaciones de audio puede ser una grabación de audio de un hablante a ser verificado, mientras que la otra de las grabaciones de audio puede ser una grabación de audio de una persona desconocida (una grabación de muestra). El propósito de dicho examen es establecer la identidad o

diferencia de las personas cuyas voces fueron grabadas en las grabaciones a ser comparadas.

Las grabaciones de audio comparadas son convertidas a formato digital y sus imágenes digitales son almacenadas en una memoria de PC como archivos de señal de audio digital.

Las imágenes digitales son sometidas a análisis espectral usando un PC, según los procedimientos de análisis espectral de señal de aceptación general (S.L. Marple, Digital spectral analysis and its applications, Mir, 1990), sus espectrogramas dinámicos son representados en un gráfico y son usados para calcular las trayectorias de los formantes como líneas de modificación de tiempo secuencial de las frecuencias resonantes del tracto vocal, representadas en los espectrogramas como máximos locales del espectro. Las frecuencias de los formantes son corregidas, si es necesario, colocándolas sobre un espectrograma, rectificando las diferencias evidentes en el movimiento de los formantes.

Los fragmentos de referencia a partir de los cuales deben compararse las frecuencias de los formantes, se seleccionan de entre los espectrogramas de las grabaciones de audio primera y segunda comparadas.

Las propias frecuencias de los formantes están influenciadas por muchos factores aleatorios en el momento de pronunciar ciertos sonidos, lo que explica su "desaparición" de algunos segmentos, y "vibración" de segmento a segmento. Además, debido a que la geometría del tracto vocal cambia constantemente en el transcurso del suministro de voz, las frecuencias de los formantes siguen moviéndose suavemente de un valor a otro, formando las trayectorias de los formantes. Al mismo tiempo, la comparación de las frecuencias de los formantes requiere la selección de ciertos valores fijos de las mismas para cada uno de los fragmentos de referencia de la señal de voz.

Para garantizar que se cumpla esta condición, las frecuencias de los formantes usadas para comparar los fragmentos de referencia y las grabaciones de audio se seleccionan en esta realización del procedimiento de identificación de la siguiente manera.

Al principio, los fragmentos de referencia que cumplen los dos criterios siguientes se seleccionan de entre los espectrogramas de las grabaciones de audio primera y segunda:

(1) - el fragmento de referencia contiene las trayectorias de los formantes de tres o más formantes;

(2) - los valores de frecuencia de los dos primeros formantes de las trayectorias de los formantes indicadas anteriormente están dentro de la variación estándar de las frecuencias de los formantes para uno de los tipos de sonido predeterminados a partir de un conjunto fijo de sonidos similares a vocales.

Se ha encontrado experimentalmente que el rango de frecuencias del canal telefónico requiere, por regla general, cuatro trayectorias de los formantes en el fragmento de referencia seleccionado para los hablantes masculinos y tres para los hablantes femeninos.

A continuación, se seleccionan los vectores de formantes en cada fragmento de referencia seleccionado para la posterior comparación de las frecuencias de los formantes; los valores de las frecuencias de los formantes se calculan como un promedio sobre intervalos de tiempo de duración fija en los que las frecuencias de los formantes son relativamente invariables.

El uso de vectores de formantes para comparar las frecuencias de los formantes dentro de cada fragmento de referencia permite seleccionar de entre el conjunto de valores formantes "vibrantes" y, a veces de “desaparición”, aquellos segmentos en los que las frecuencias de los formantes interpoladas y suavizadas en el tiempo y en la frecuencia forman segmentos de valores relativamente estables que son adecuados para una detección fiable y una comparación posterior. Esto proporciona una resistencia al ruido adicional de las características de individualización en un entorno de ruido aditivo de banda ancha.

Se ha encontrado experimentalmente que la presencia de tres o más trayectorias de los formantes dentro del espectrograma de un fragmento de referencia en la mayoría de los casos permite la determinación inequívoca de los valores de frecuencias de los formantes para al menos un vector de formantes dentro del fragmento de referencia. Esto proporciona una identificación fiable de los hablantes en el transcurso de una comparación adicional.

La duración del fragmento de referencia está determinada por los límites del segmento de la señal de voz para los cuales las trayectorias de los formantes se determinan de manera inequívoca para esos formantes, que a continuación se usan para determinar los vectores de formantes. Los segmentos de señal de voz largos que son significativamente heterogéneos en su estructura fónica, adecuados para su uso como fragmentos de referencia, se dividen en diversos fragmentos de referencia, cada uno de cuyos dos primeros formantes, por regla general, no superan los límites típicos de la variación de la frecuencia de los formantes para un tipo de fonema vocal en el idioma correspondiente.

Un ejemplo de dicho espectrograma con las trayectorias de los formantes superpuestas se muestra en la Figura 1.

El espectrograma mostrado en la Figura 1 es el espectrograma de la sílaba [te] pronunciada por un hablante masculino. El eje horizontal representa el tiempo en segundos. El eje vertical representa la frecuencia en Hz. La intensidad de ennegrecimiento del espectrograma representa la potencia de la señal en el punto de tiempo y de frecuencia correspondiente. Las líneas negras delgadas representan las frecuencias de los formantes seleccionadas automáticamente que forman las trayectorias de los formantes.

Las líneas verticales marcan los límites del fragmento de referencia señalado y el vector de formantes seleccionado para la comparación. Un vector de formantes es un intervalo estrecho dentro de un fragmento de referencia dentro del cual el valor de frecuencia del formante es relativamente invariable.

Los cursores horizontales marcan la posición de las frecuencias de los formantes. El espectrograma en la Figura 1 los muestra como aproximadamente iguales a 430, 1,345, 2,505, 3,485 Hz.

Una vez seleccionados los fragmentos de referencia para las señales de voz comparadas a partir de las grabaciones de audio primera y segunda y una vez definidos los vectores de formantes dentro de dichos fragmentos de referencia, para cada fragmento de referencia y vector de formantes de la primera grabación de audio, un fragmento de referencia y un vector de formantes correspondiente son seleccionados para la comparación a partir de la segunda grabación de audio, de manera que los vectores a comparar tengan valores de frecuencia de los formantes coincidentes para al menos dos formantes. Si esto es imposible, los fragmentos de referencia se consideran no comparables y no se comparan más.

Dicha elección de los fragmentos de referencia y los vectores de formantes a ser comparados corresponde a la selección y la comparación de los segmentos de señal de voz en las grabaciones de audio primera y segunda, donde se actualizan los eventos articulatorios comparables, es decir, los hablantes pronuncian sonidos con una estructura de resonancia altamente discernible del espectro de la señal y las frecuencias de dos o más resonancias (formantes) coinciden independientemente del fonema pronunciado por los hablantes.

Tal como se ha indicado anteriormente, dichos fragmentos comparables en términos de algunos valores de frecuencia de formante se denominan en la literatura experta como fragmentos de formantes iguales. Estos fragmentos de formantes iguales se encuentran en segmentos de voz en los que los hablantes pronuncian fonemas tanto idénticos como diferentes. Por ejemplo, puede haber un fragmento de pronunciación de vocal estática seleccionado como un fragmento de referencia para la comparación en la primera grabación de audio y, en la segunda grabación de audio, un fragmento puede ser seleccionado con una transición rápida desde un fonema a otro con un segmento en el que al menos dos valores de frecuencia de formantes coinciden con los valores de frecuencia de los formantes correspondientes de la vocal estática en la primera grabación de audio.

La coincidencia o no coincidencia de las frecuencias de los formantes frecuentemente sigue el procedimiento de umbral conocido. Los valores umbral de desviación aceptable dependen de la calidad de la señal grabada en una grabación de audio particular (la relación señal/ruido, el tipo y la intensidad de ruido y distorsión, el estado físico y psico-fisiológico del hablante), y se determinan en base a la variabilidad de la frecuencia formante natural para cada hablante dentro de una grabación de audio determinada, para tipos de sonido determinados, y para cada formante por separado. Esta variabilidad y los umbrales correspondientes se determinan, por ejemplo, buscando y comparando entre sí fragmentos de referencia y vectores de formantes dentro de la grabación de audio, es decir, para señales de voz de cada una de las grabaciones de audio.

Entre los segmentos de formantes iguales, se comparan las frecuencias de los formantes de los restantes, es decir, no iguales. La coincidencia o no coincidencia de los valores de estas frecuencias determina la coincidencia o no coincidencia entre los fragmentos de referencia comparados (ciertos fragmentos de voz). Por ejemplo, en los casos típicos prácticos, una desviación del 3% se considera aceptable, y una desviación del10% se considera inaceptable.

Si las frecuencias de los formantes del vector coinciden, los hablantes se consideran idénticos para este tipo de sonido. Esta situación demuestra que los hablantes comparados tienen una geometría de tracto vocal idéntica caracterizada por resonancias acústicas al pronunciar este tipo de sonido en los períodos de tiempo comparados.

La decisión de coincidencia o no coincidencia para los fragmentos de formantes iguales de las grabaciones de audio se adopta para cada fragmento de referencia seleccionado de la primera grabación de audio. La adopción de una decisión informada requiere la comparación de varios fragmentos de referencia (normalmente 3-5) para cada tipo de sonido con frecuencias de los formantes del fragmento de referencia correspondientes a un tipo de sonido determinado.

Si, para el fragmento de referencia seleccionado de la primera grabación de audio, se encuentra un fragmento de formantes iguales en la segunda grabación de audio de manera que las frecuencias de los formantes de los vectores de formantes correspondientes del fragmento encontrado no coinciden con las del fragmento de referencia

en la primera grabación de audio, debe realizarse una búsqueda inversa de un fragmento de formantes iguales con las mismas frecuencias de los formantes en la primera grabación de audio. Sólo un fallo al encontrar un vector coincidente puede servir como base para adoptar la decisión de falta de coincidencia de los vectores de formantes en las dos grabaciones de audio para este tipo de sonido y, por lo tanto, la diferencia de hablantes para este tipo de fragmento. Esta situación demuestra que los hablantes comparados tienen diferentes geometrías de tracto vocal para articulaciones comparables caracterizadas por resonancias acústicas al pronunciar este tipo de sonido en los períodos de tiempo comparados.

La identificación del hablante usando una selección de fragmentos de formantes iguales que corresponden a diferentes sonidos, no necesariamente idénticos, permite el reconocimiento del hablante en casos con contenidos fónicos que difieren fuertemente de las grabaciones de audio. El enfoque indicado anteriormente asegura, sobre todo, la identificación del hablante a partir de grabaciones de audio cortas, grabaciones de audio realizadas por hablantes en estados psico-fisiológicos diferentes, así como hablantes de los mismos o diferentes idiomas.

El presente inventor ha demostrado experimentalmente que para garantizar una alta fiabilidad de la identificación del hablante masculino en la mayoría de los casos prácticos se requiere seleccionar frecuencias iguales para tres formantes y comparar el valor del cuarto. Las voces de mujeres y de niños requieren también seleccionar tres formantes iguales, pero debido a que esto no es posible en muchos casos prácticos se permite seleccionar dos frecuencias de formantes iguales y comparar el valor del tercero.

La adopción de una decisión de identificación global general requiere la búsqueda y la comparación de fragmentos comparables de formantes iguales en las grabaciones de audio para obtener una selección representativa de los tipos de articulación más diferentes, es decir, para la máxima diferencia entre las geometrías de tracto vocal de los hablantes comparados. Los estudios teóricos demuestran que estos tipos de articulación (L.V. Bondarko, L.A. Verbitskaya, M.V. Gordina Basics of General Phonetics. - San Petersburgo: SPSU 2004, la teoría de G. Fant Acoustic theory of speech production. -Moscú: Nauka, 1964; V.N. Sorokin. Theory of speech production. - M.: Radio and communication. 1985) corresponden a las vocales en la parte superior del denominado "triángulo fonético", es decir, sonidos de vocales con valores mínimo y máximo de frecuencias de los formantes primero y segundo para todo el rango de cambios de las dos primeras frecuencias de los formantes.

Por ejemplo, un conjunto representativo del idioma ruso incluye típicamente fragmentos de referencia de sonidos con frecuencias de formantes cercanas a los valores medios de las vocales del tipo [Λ], [o], [u], [e], [i]. Para el idioma Inglés, las vocales de tipo A, O, U, E, I.

Por ejemplo, en la mayoría de los idiomas de diferentes tipos, las frecuencias de los formantes de los sonidos en un conjunto representativo de fragmentos de referencia para hablantes masculinos deberían ser de los siguientes valores típicos con una desviación de aproximadamente el +/- 20%.

Tipo de sonido	1 (U)	2 (O)	3 (A)	4 (E)	5 (I)	6 (ϵ)	7 (¥)	8 (ɜ)
Primer formante F1, Hz	300	430	700	420	300	500	350	500
Segundo formante F2, Hz	700	1.000	1.300	1.800	2.100	1.650	1.350	1.500

De manera ventajosa, se seleccionan fragmentos de referencia similares en términos de frecuencias de los formantes para todos los tipos de sonido en la tabla. En este caso, preferiblemente se seleccionan varios fragmentos de referencia para cada tipo de sonido. Se ha encontrado experimentalmente que para garantizar una decisión de identificación fiable para cada tipo de sonido es suficiente encontrar 3-4 fragmentos de referencia en cada una de las grabaciones de audio comparadas. En general, la adopción de una decisión final de identificación global requiere 4-8 tipos diferentes de sonidos. De esta manera, es factible usar 12-32 fragmentos de referencia en las grabaciones de audio comparadas, dependiendo de la calidad de la señal de voz.

En el caso de grabaciones de audio cortas, el conjunto de sonidos mínimo requerido contiene fragmentos de referencia con frecuencias de los formantes cercanas de al menos tres tipos diferentes de sonidos de la tabla.

El procedimiento de selección de formantes iguales y de comparación de formantes descrito anteriormente se aplica para los fragmentos de referencia de cada tipo de sonido (normalmente, al menos, 3), y se adopta una decisión de identificación de coincidencia/no coincidencia de formantes particular para cada tipo de sonido. En base a una pluralidad de decisiones sobre los fragmentos de referencia de formantes iguales para cada tipo de sonido,

se adopta una decisión de coincidencia de frecuencia de los formantes para un tipo de sonido determinado al que corresponden los fragmentos de referencia comparados. La decisión es positiva si todos los fragmentos de formantes iguales de este tipo demuestran una coincidencia de frecuencia de los formantes para todos los formantes considerados. La decisión puede ser probabilística si la calidad de la grabación de audio es deficiente y no permite determinar la posición de los formantes con precisión suficiente. La decisión puede ser incierta si la calidad y la cantidad del material de grabación de audio de voz son insuficientes para proporcionar un reconocimiento fiable de los fragmentos y los vectores de formantes en el mismo.

Las decisiones de identificación individuales para tipos de sonido diferentes se combinan en una decisión de identificación global general sobre la identidad o diferencia de los hablantes cuyos mensajes de voz están grabados en los archivos de audio comparados.

Una decisión global definitivamente positiva se realiza normalmente en el caso de decisiones individuales de identificación positiva para al menos cinco sonidos diferentes con generalmente al menos 15 fragmentos de referencia comparados y en ausencia de decisiones de identificación individuales negativas que descarten la identidad del hablante. Una decisión de identificación negativa se realiza si hay al menos una decisión de identificación particular definitivamente negativa para un tipo de sonido. La decisión puede ser probabilística si la calidad de la grabación de audio es deficiente y no permite distinguir un número suficiente de fragmentos de referencia, ni determinar con suficiente precisión la posición de los formantes.

La implementación del procedimiento se demuestra mediante los siguientes ejemplos que comparan espectrogramas con trayectorias de los formantes superpuestas para hablantes coincidentes o no coincidentes.

Las Figuras 2-4 muestran comparaciones ejemplares para espectrogramas con trayectorias de los formantes superpuestas para frecuencias de los formantes coincidentes para los sonidos de tipo [e] (Figuras 2,3), para los sonidos de tipo [e] (Figura 4) y para vectores de formantes no coincidentes para sonidos de tipo [e] (Figura 5).

Con referencia a la Figura 2, se muestran espectrogramas con trayectorias de los formantes superpuestas para sonido de tipo [e] de la sílaba [te] (en la parte izquierda de la Figura) y la palabra ['mogut]/pueden/ (en la parte derecha de la Figura) pronunciadas por el mismo hablante masculino. El eje horizontal representa el tiempo en segundos. El eje vertical representa la frecuencia en Hz. La intensidad de ennegrecimiento en el espectrograma representa la potencia de la señal en el punto de tiempo y de frecuencia correspondiente. Las líneas negras delgadas representan las frecuencias de los formantes seleccionadas automáticamente que forman las trayectorias de los formantes. Las líneas verticales marcan los límites de los fragmentos de referencia seleccionados para la comparación y los vectores de formantes en los mismos. Los cursores horizontales marcan la posición coincidente de las frecuencias de los formantes. Son aproximadamente iguales a 430, 1,345, 2,505 y 3,485 Hz. Diferentes fonemas tienen frecuencias coincidentes en los vectores de formantes marcados (el centro de la tónica [e] y el final de la post-tónica [U]).

Con referencia a la Figura 3, en la misma se muestran espectrogramas con trayectorias de los formantes superpuestas para sonidos de tipo [e] de la sílaba [te] (en la parte izquierda de la Figura) para hablantes coincidentes y la sección central de la sílaba [suda] de la palabra [gosu'darstvo]/estado/ (en la parte derecha de la Figura) pronunciadas por el mismo hablante masculino. El eje horizontal representa el tiempo en segundos. El eje vertical representa la frecuencia en Hz. La intensidad de ennegrecimiento en el espectrograma representa la potencia de la señal en el punto de tiempo y de frecuencia correspondiente. Las líneas negras delgadas representan las frecuencias de los formantes seleccionadas automáticamente que forman las trayectorias de los formantes. Las líneas verticales marcan los límites de los fragmentos de referencia seleccionados para la comparación y los vectores de formantes en los mismos. Los cursores horizontales marcan la posición coincidente de las frecuencias de los formantes. Son aproximadamente iguales a 430, 1,345, 2,505 y 3,485 Hz. Diferentes fonemas tienen frecuencias coincidentes en los vectores de formantes marcados (la mitad de la tónica [e] y el inicio de la pre-tónica [U]).

Con referencia a la Figura 4, en la misma se muestran espectrogramas con trayectorias de los formantes superpuestas pronunciadas por hablantes coincidentes para sonidos de tipo [e] de la sílaba [re] de la palabra [inte'resny]/interesante/ (en la parte izquierda de la Figura) y la sílaba [she] de la palabra [re'shenije]/decisión/ (en la parte derecha de la Figura) pronunciadas por el mismo hablante masculino. El eje horizontal representa el tiempo en segundos. El eje vertical representa la frecuencia en Hz. La intensidad de ennegrecimiento en el espectrograma representa la potencia de la señal en el punto de tiempo y de frecuencia correspondiente.

Las líneas negras delgadas representan las frecuencias de los formantes seleccionadas automáticamente que forman las trayectorias de los formantes. Las líneas verticales marcan los límites de los fragmentos de referencia seleccionados para la comparación y los vectores de formantes en los mismos. Los cursores horizontales marcan la posición coincidente de las frecuencias de los formantes para los vectores de formantes del fonema [e] tónico. Son aproximadamente iguales a 340, 1,850, 2,430 y 3,505 Hz.

Con referencia a la Figura 5, en la misma se muestran espectrogramas con trayectorias de los formantes superpuestas pronunciadas por hablantes no coincidentes para los sonidos de tipo [e] de la sílaba [re] de la palabra [inte'resny]/interesante/ (en la parte izquierda de la Figura) y la sílaba [de] de la palabra [utverzh'denie]/declaración/ (en la parte derecha de la Figura) pronunciadas por hablantes masculinos diferentes. El eje horizontal representa el tiempo en segundos. El eje vertical representa la frecuencia en Hz. La intensidad de ennegrecimiento en el espectrograma representa la potencia de la señal en el punto de tiempo y de frecuencia correspondiente. Las líneas negras delgadas representan las frecuencias de los formantes seleccionadas automáticamente que forman las trayectorias de los formantes. Las líneas verticales marcan los límites de los fragmentos de referencia seleccionados para la comparación y los vectores de formantes en los mismos. Los cursores horizontales marcan la posición coincidente de tres valores de frecuencia de los formantes. Son aproximadamente iguales a 340, 1,850 y 3,505 Hz. En lugar del tercer formante de 2,430 Hz del hablante del espectrograma izquierdo, el hablante del espectrograma derecho tiene el tercer formante en la región de 2,800 Hz.

Para garantizar una alta fiabilidad de la identificación del hablante con la señal de voz gravemente distorsionada debido a una respuesta de frecuencia sustancialmente irregular de los canales de audio que es diferente para las grabaciones de audio comparadas, se propone una realización del procedimiento de identificación del hablante, en la que la realización implica el filtrado inverso de las señales de voz de las grabaciones de audio comparadas antes de la ejecución de las etapas anteriores.

Antes de calcular los espectrogramas y los valores de frecuencia de los formantes, el espectro de potencia de la señal de cada grabación de audio es sometido a filtrado inverso. En este punto, se calcula el espectro promediado en el tiempo para el espectrograma en su conjunto o sus partes particulares y, a continuación, para cada frecuencia, el valor original de la potencia de espectro de la señal de grabación de audio es dividido por el valor de potencia inversa del espectro promediado para una frecuencia determinada. El valor de potencia inversa es el valor de potencia de espectro obtenido dividiendo uno por este valor.

El procedimiento de filtrado inverso puede ser realizado no mediante una división, sino calculando un logaritmo y restando los espectros correspondientes. En este caso, antes de calcular los espectrogramas y las frecuencias de los formantes, el espectro de la señal de cada grabación de audio es transformado a un logaritmo del espectro, el logaritmo del espectro promediado en el tiempo es calculado para la grabación de audio en su conjunto o sus partes individuales, y a continuación el logaritmo de la potencia del espectro promediado para cada frecuencia es restado del logaritmo de la potencia del espectro de señal original de la grabación de audio procesada.

Las Figuras 6-8 ilustran un ejemplo del uso de filtrado inverso para determinar los valores de las frecuencias de los formantes para el vector de formantes seleccionado.

La Figura 6 muestra tres espectrogramas de la misma señal de voz (fragmento de referencia con la sílaba [na]):

- el espectrograma del cuadro de la izquierda corresponde a la señal de voz original, grabada a través de un micrófono de alta sensibilidad;
- el espectrograma del cuadro central corresponde a la misma señal de voz, grabada a través de un canal telefónico de baja calidad;
- el espectrograma del cuadro de la derecha corresponde a la misma señal de voz, grabada a través de un canal telefónico de baja calidad, después de haber sido sometida a filtrado inverso según la realización propuesta del procedimiento.

En cada espectrograma en la Figura 6, el eje horizontal representa el tiempo en segundos desde el comienzo de la grabación de audio, el eje vertical representa la frecuencia en Hz. La intensidad del ennegrecimiento en el espectrograma es proporcional a la potencia de espectro de la señal en el punto de tiempo y de frecuencia correspondiente. Los cursores horizontales marcan los cinco valores de frecuencia de resonancia de baja frecuencia del tracto vocal del hablante para el vector de formantes marcado por los cursores verticales, en la grabación de audio original.

En la grabación de audio central, los óvalos con los números 2 y 3 marcan las zonas del espectro donde no hay rastro de los formantes F1 y F3, presentes en la grabación de audio original.

El óvalo marcado con el número 1 muestra una frecuencia de formante falsa, que sólo aparece en el espectrograma 2 debido a la influencia del canal telefónico de baja calidad.

La Figura 7 muestra: la curva (1) – el espectro de potencia media de la señal de voz original para toda la grabación de audio, cuyo espectrograma se muestra en el cuadro de la izquierda de la Figura 6; curva (2) – el espectro de potencia media de la señal de voz original para toda la grabación de audio, cuyo espectrograma se muestra en el cuadro central de la Figura 6; curva (3) – el espectro de potencia media de la señal de voz original para toda la

grabación de audio, cuyo espectrograma se muestra en el cuadro derecho de la Figura 6.

La Figura 8 muestra los espectros promediados y las frecuencias de los formantes marcadas para el vector de formantes, resaltadas en los espectrogramas en la Figura 6. La curva (1) representa el espectro medio del vector de formantes para el espectrograma original, mostrado en el cuadro de la izquierda en la Figura 6, la curva (2) es para la misma grabación de audio, pasada a través de un canal telefónico de baja calidad (corresponde al cuadro central en la Figura 6), la curva (3) es para la grabación de audio, pasada a través de un canal telefónico de baja calidad y sometida a filtrado inverso (cuadro de la derecha en la Figura 6). Los cursores verticales marcan los formantes F1, F2, F3, F4 y F5, que coinciden para la grabación de audio original (cuadro de la izquierda en la Figura 6) y la grabación de audio después del filtrado inverso (cuadro de la derecha en la Figura 6). Los formantes F1 y F2 no se muestran en la curva 2. En la curva 2, el formante WF3 falso adicional es claramente visible, ausente de este vector de formantes en la grabación de audio original.

De esta manera, en el segmento del espectro de señal de voz, grabado a través de un canal telefónico de baja calidad, en el que no hay rastro de los formantes F1 y F3 (óvalos marcados con los números 2 y 3 en el espectrograma del cuadro central en la Fig. 6), la Figura 8 muestra estos formantes F1 y F3 como presentes en la grabación de audio original (curva 1) y en la grabación de audio después del filtrado inverso (curva 3). No están presentes en la grabación de audio del canal telefónico de baja calidad (curva 2 en la Figura 8). Al mismo tiempo, la grabación 2 de audio muestra un formante WF3 falso.

Estos ejemplos muestran que el filtrado inverso propuesto permite restaurar la estructura de formantes original de una grabación de audio, modificada cuando la grabación de audio pasa a través de canales de comunicación de baja calidad.

Los presentes inventores han determinado experimentalmente que el filtrado inverso realizado tanto mediante división como mediante el cálculo de logaritmos y restando los espectros correspondientes proporciona prácticamente los mismos resultados.

El procedimiento propuesto puede ser implementado usando el equipo existente.

REIVINDICACIONES

1. Un procedimiento para la identificación de un hablante a partir de una grabación de audio de habla oral, en el que el procedimiento comprende:

evaluar la similitud entre una primera grabación de audio de un hablante y una segunda grabación de audio o muestra por una coincidencia de las frecuencias de los formantes en los fragmentos de referencia de la señal de voz seleccionada para la comparación a partir de la primera grabación de audio y la segunda grabación de audio, y comprende además:

seleccionar fragmentos de referencia de la señal de voz desde la primera grabación de audio y la segunda grabación de audio de manera que los fragmentos de referencia comprendan trayectorias de los formantes de al menos tres frecuencias de los formantes;

para cada fragmento de referencia de la primera grabación de audio, seleccionar un fragmento de referencia de la segunda grabación de audio para la comparación, de manera que los fragmentos de referencia a ser comparados tengan valores de frecuencia de formantes coincidentes para al menos dos frecuencias de los formantes;

evaluar la similitud de dichos fragmentos de referencia a ser comparados mediante la comparación de las frecuencias de los formantes restantes, en el que dichos fragmentos de referencia a ser comparados coinciden si sus frecuencias de los formantes restantes coinciden dentro de una desviación predeterminada,

en el que la similitud de las grabaciones de audio es reconocida si todos los fragmentos de referencia comparados coinciden.

2. Procedimiento según la reivindicación 1, en el que las frecuencias de los formantes en cada uno de los fragmentos de referencia seleccionados se calculan como valores medios durante los intervalos de tiempo de duración fija en los que las frecuencias de los formantes son relativamente invariables.

3. Procedimiento según la reivindicación 1, en el que se seleccionan y se comparan fragmentos de referencia en los que los valores de frecuencia para los dos primeros formantes coinciden, en el que los valores de frecuencia para los dos primeros formantes están dentro de un límite predeterminado de variabilidad típica de los valores de frecuencia de los formantes para el tipo correspondiente de fonemas de vocal en un idioma determinado.

4. Procedimiento según la reivindicación 1, en el que al menos dos fragmentos de referencia de la señal de voz son seleccionados para la comparación a partir de las grabaciones de audio, en el que los al menos dos fragmentos de referencia están relacionados con sonidos de una articulación de diferencia máxima con los valores de frecuencia máximo y mínimo para una grabación de audio determinada.

5. Procedimiento según cualquiera de las reivindicaciones 1-4, en el que,

antes de calcular los valores de las frecuencias de los formantes, el espectro de potencia de la señal de voz para cada grabación de audio es sometido a filtrado inverso,

en el que se calcula el promedio temporal para cada componente de frecuencia del espectro de potencia, al menos para fragmentos particulares de la grabación de audio, y a continuación

el valor original del espectro de potencia de la señal de la grabación de audio es dividido por su valor inverso medio para cada componente de frecuencia del espectro.

6. Procedimiento según cualquiera de las reivindicaciones 1-4, en el que,

antes de calcular los valores de las frecuencias de los formantes, el espectro de potencia de la señal de voz para cada grabación de audio es sometido a filtrado inverso,

en el que se calcula el promedio temporal para cada componente de frecuencia del espectro de potencia, al menos para fragmentos individuales de la grabación de audio, y a continuación

los espectros son sometidos a una etapa de cálculo de logaritmo, y el valor de logaritmo promedio del espectro de potencia de la señal de la grabación de audio para cada componente de frecuencia es restado de su logaritmo del valor original.

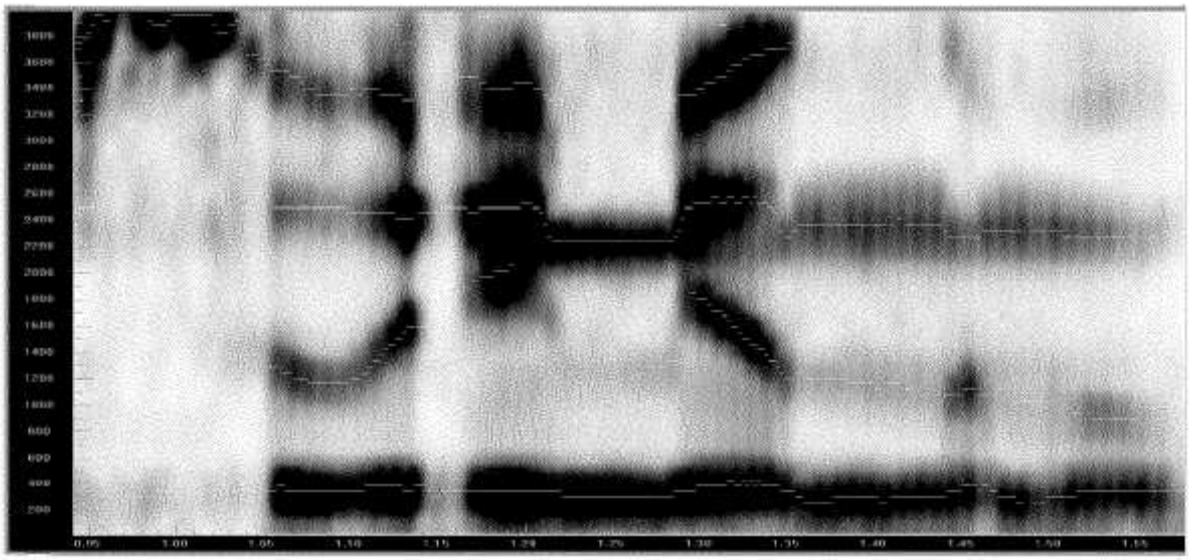

FIG. 1

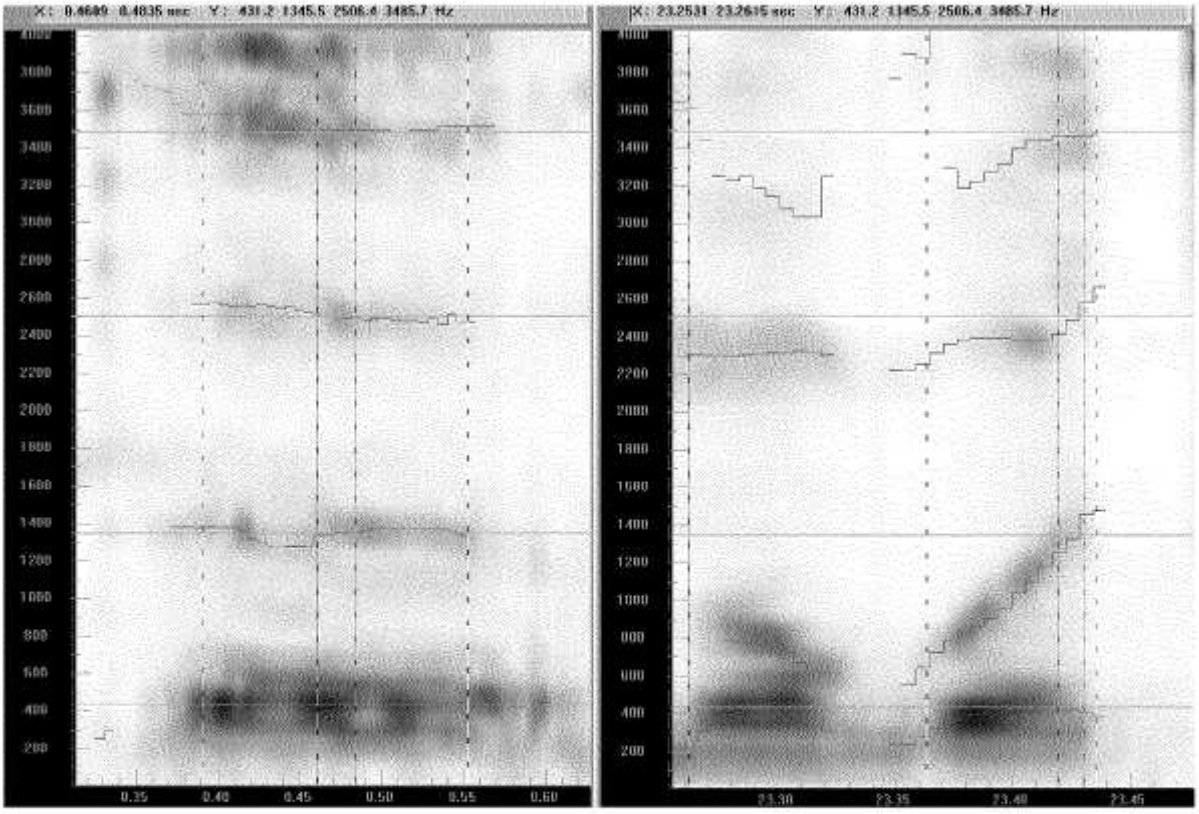

FIG. 2

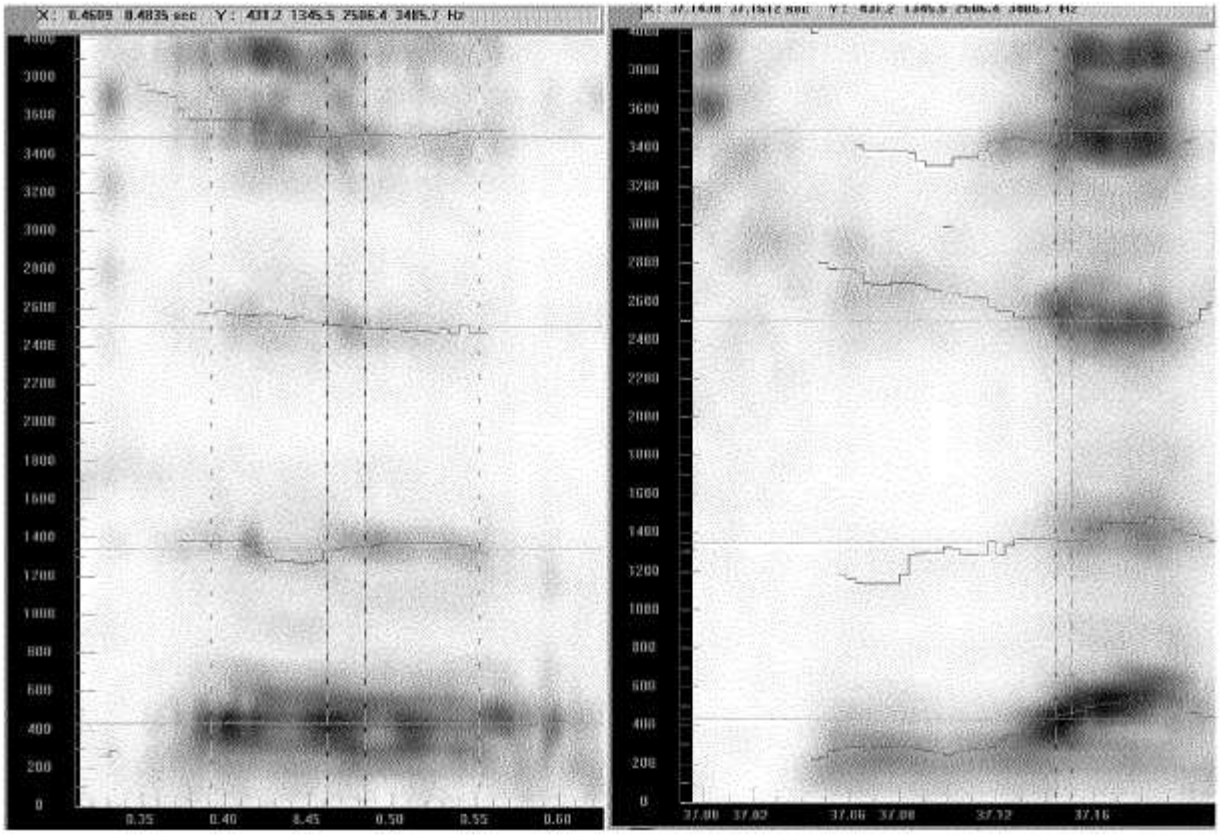

FIG. 3

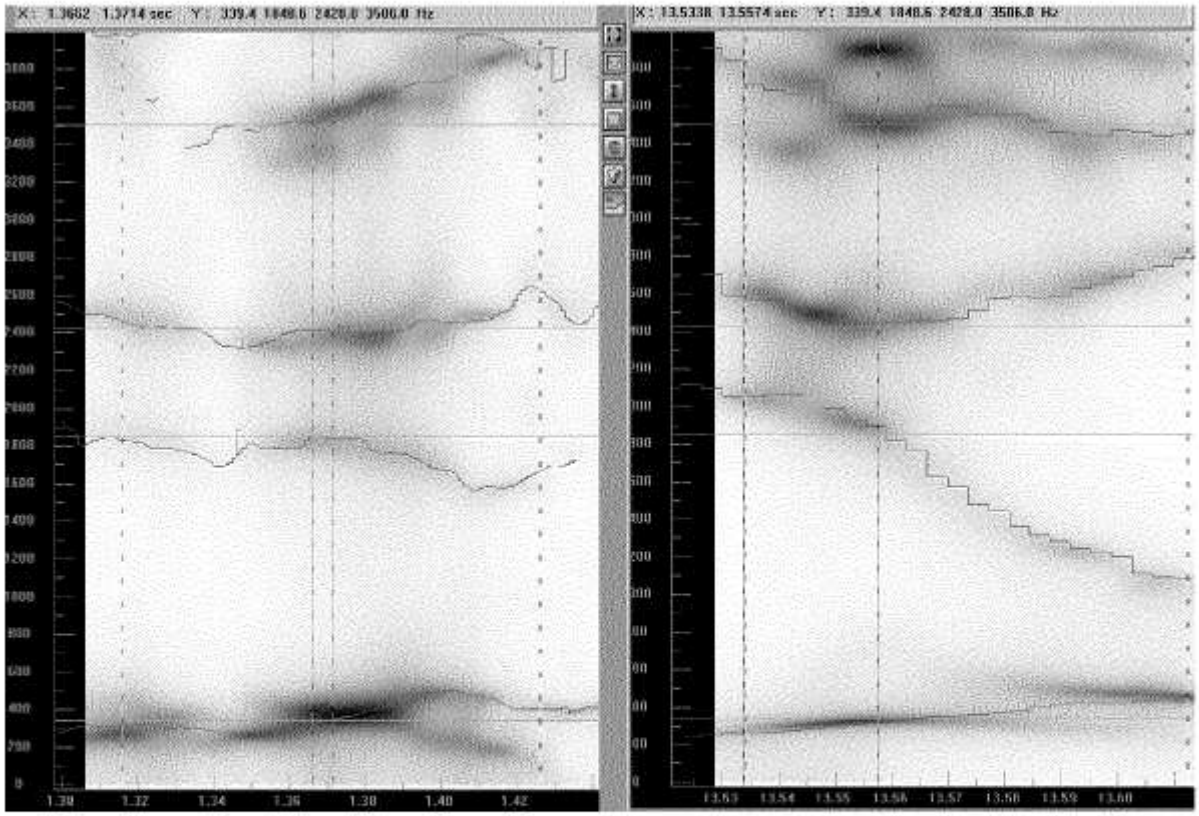

FIG. 4

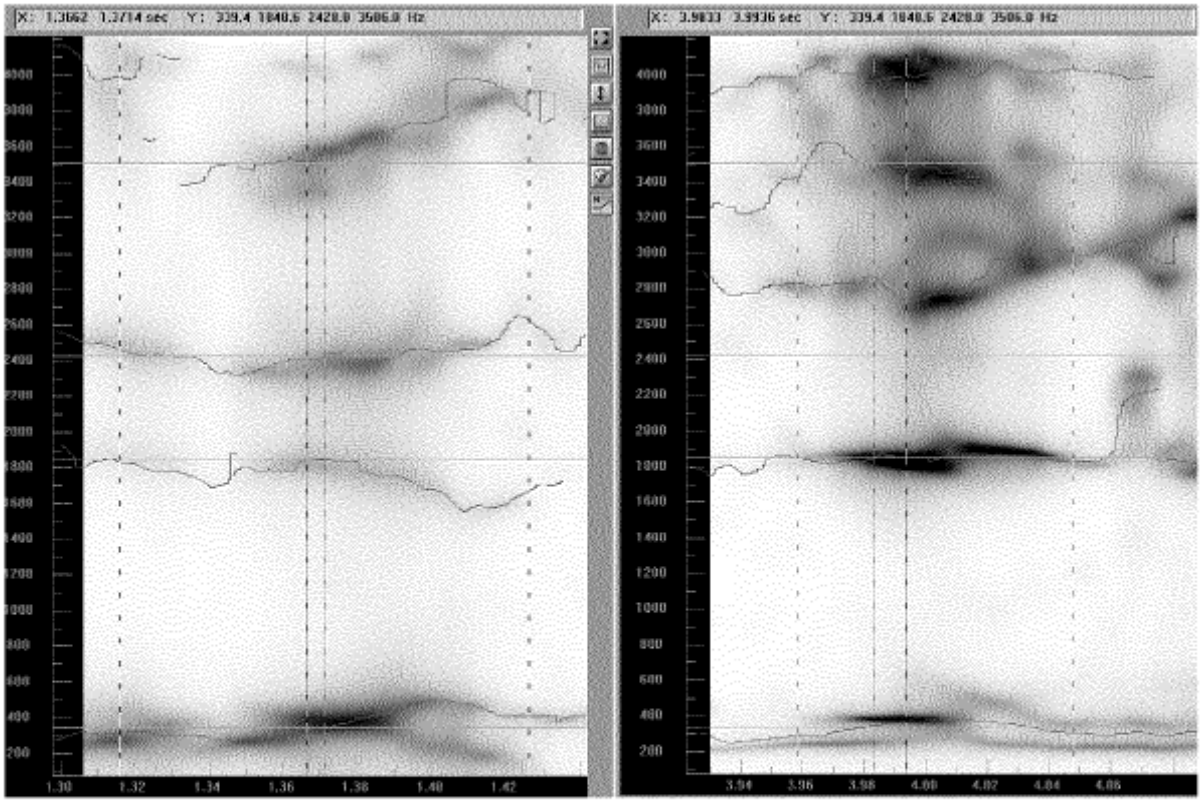

FIG. 5

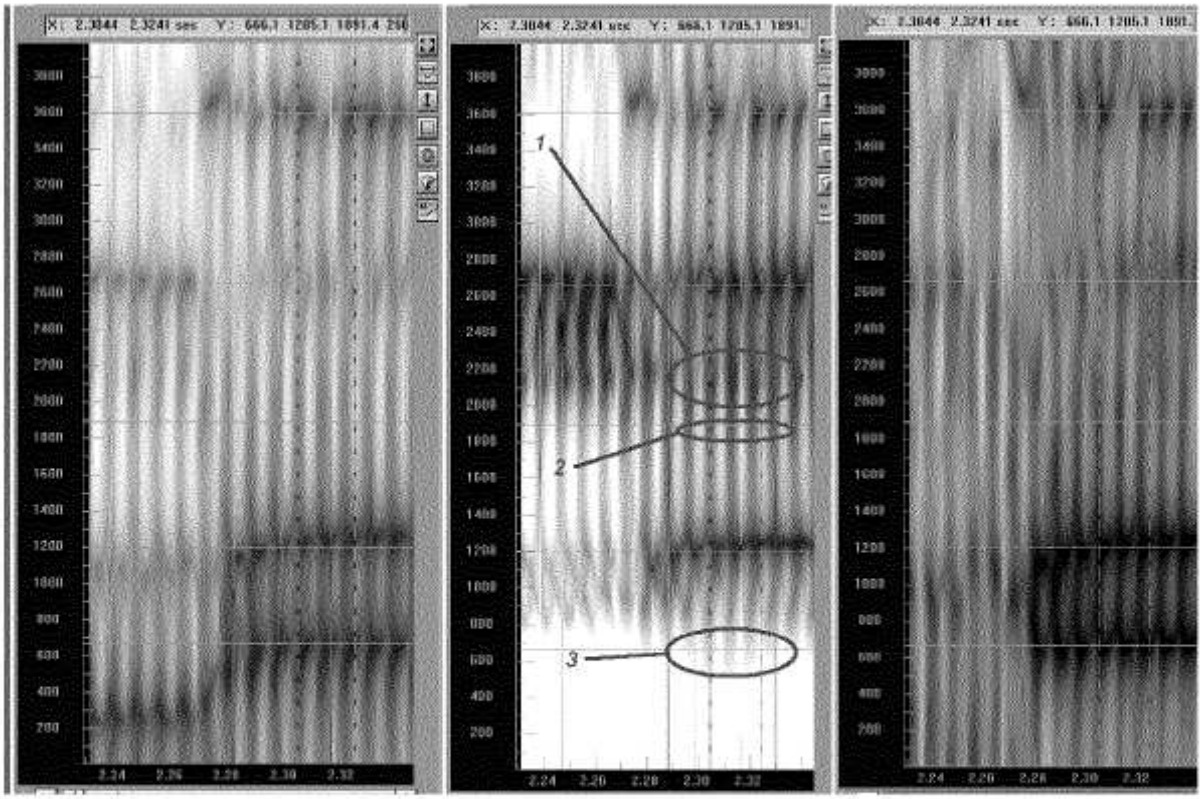

FIG. 6

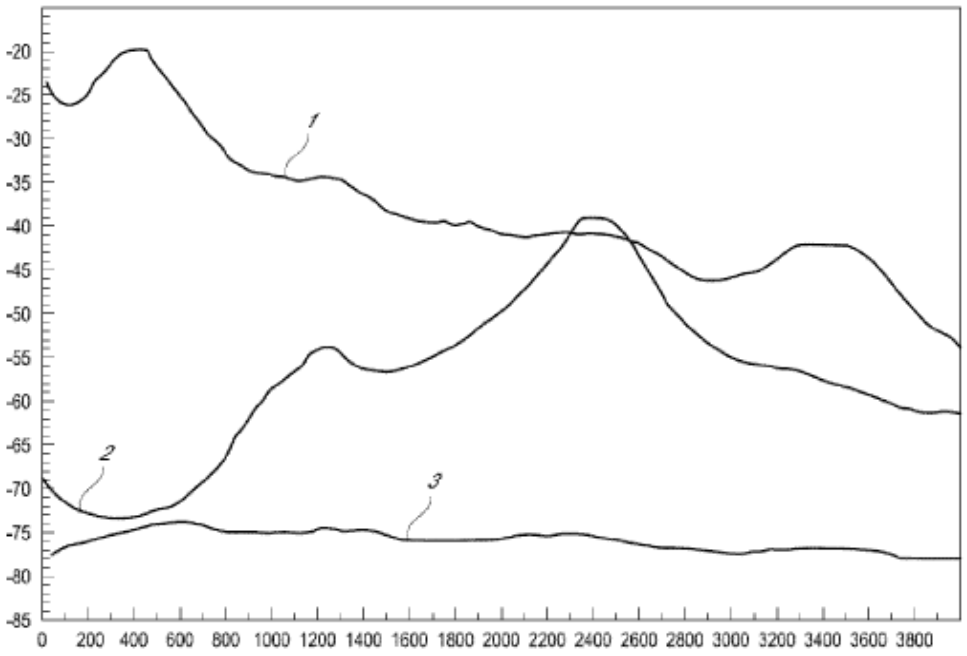
-20
-25
-30
-35
-40
-45
-50
-55
-60
-65
-70
-75
-80
-85
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800
1
2
3

FIG 7

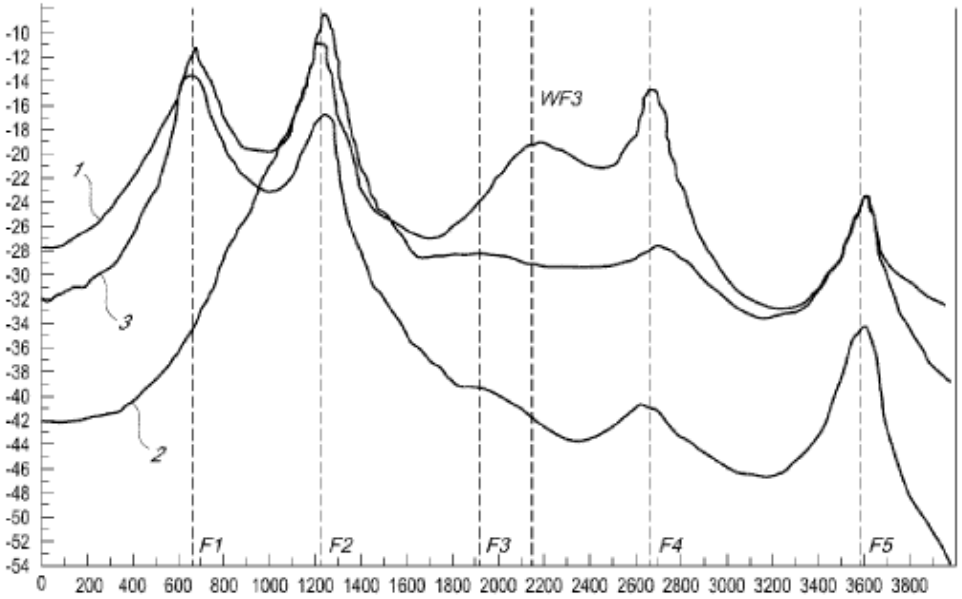
-10
-12
-14
-16
-18
-20
-22
-24
-26
-28
-30
-32
-34
-36
-38
-40
-42
-44
-46
-48
-50
-52
-54
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800
1
2
3
WF3
F1
F2
F3
F4
F5

FIG 8