(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 554 622**

(51) Int. CI.:

H04R 1/40 (2006.01)
H04M 1/60 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **01.12.2010 E 10785422 (6)**

(97) Fecha y número de publicación de la concesión europea: **23.09.2015 EP 2508009**

(54) Título: **Dispositivo y método para capturar y procesar la voz**

(30) Prioridad:

02.12.2009 CH 18482009

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.12.2015

(73) Titular/es:

VEOVOX SA (100.0%)
Chemin des Roches 10 CP 508
1009 Pully, CH

(72) Inventor/es:

LISSEK, HERVÉ;
MARTIN, PHILIPPE;
CARMONA, JORGE;
IMHASLY, MICHEL;
MILLAR, IAN;
FALOURD, XAVIER;
MARMAROLI, PATRICK y
MAÎTRE, GILBERT

(74) Agente/Representante:

CURELL AGUILÁ, Mireia

DESCRIPCIÓN

Dispositivo y método para capturar y procesar la voz.

Campo de la invención

La presente invención se refiere a un dispositivo y un método para capturar y procesar la voz, especialmente en entornos con ruido. La invención se refiere, entre otros aspectos, a un dispositivo móvil que se puede utilizar en entornos con ruido, tales como, sin carácter limitativo, restaurantes, para capturar y procesar la voz, y para llevar a cabo un reconocimiento del habla.

Técnica anterior

Aunque recientemente ha mejorado la velocidad de reconocimiento de los algoritmos de reconocimiento del habla, la misma sigue siendo baja en condiciones difíciles, especialmente cuando la relación señal/ruido es insuficiente. Por este motivo, la captura y el reconocimiento de la voz en entorno con ruido siguen siendo dificultosos o poco fiables.

No obstante, existe una necesidad de dispositivos con capacidad de llevar a cabo un reconocimiento fiable del habla incluso en entornos con mucho ruido, tales como (sin carácter limitativo) bares y restaurantes. Por ejemplo, sería útil disponer de un dispositivo con capacidad de capturar y reconocer la voz de un camarero en un restaurante, y utilizar este dispositivo para anotar, reconocer y transmitir pedidos por voz.

El documento US7110963 divulga un sistema de reconocimiento del habla para posibilitar que un camarero en un restaurante transmita órdenes a la cocina. Se usa una aplicación de software de reconocimiento del habla para controlar el procesado y el flujo de datos durante operaciones de anotación de pedidos y para recibir información de pedidos desde el servidor en tiempo real durante la interacción con el cliente.

El documento US-A1-2002/0007315 divulga otro sistema de pedidos activado por voz para un restaurante de comida rápida, en donde los pedidos se introducen en una caja registradora de un punto de venta y se convierten a mensajes de voz para el manipulador de alimentos del restaurante. En el punto de venta se utiliza un circuito de conversión de habla-a-texto para introducir órdenes por voz.

Las soluciones antes mencionadas son útiles y permiten una transmisión más rápida y natural de pedidos entre el restaurante y la cocina. No obstante, en muchos restaurantes con un nivel de ruido elevado o incluso medio, la fiabilidad del reconocimiento del habla es insatisfactoria; la relación señal/ruido no es suficiente para que los algoritmos existentes de reconocimiento del habla tengan un rendimiento fiable.

Se ha observado que la calidad y la directividad del micrófono tienen una importancia primordial para capturar una señal de voz de buena calidad. El documento US-B2-7120477 (Microsoft Corporation) describe un dispositivo informático móvil personal que tiene un micrófono montado en una antena y con detección del habla. La antena comprende un micrófono posicionado en su extremo distal y está adaptada para orientarse hacia un usuario, reduciendo así la distancia entre la boca del usuario del micrófono mientras el dispositivo es sostenido con la palma de la mano del usuario. La reducción de esta distancia ayuda a aumentar la relación señal/ruido de las señales de habla proporcionadas por el micrófono. No obstante, esta solución sigue siendo insuficiente para entornos con mucho ruido.

En el documento EP694833 se divulga otro módulo para capturar voz en entornos con ruido. Este documento describe una primera matriz de micrófonos orientables según el haz, para capturar voz, y una matriz de micrófonos adicional orientable según el haz para reconocer fuentes adicionales de datos de audio y fuentes de ruido y/o interferencia. La finalidad en este caso es localizar el hablante (fuente de audio) con un algoritmo de triangulación, y controlar un sistema de accionamiento mecánico para dirigir el foco de una cámara de vídeo hacia el hablante.

Las dos matrices de micrófonos son bidimensionales y ocupan por lo tanto una gran superficie; por lo tanto no es posible montar las matrices sobre un haz lineal al mismo tiempo que manteniendo una distancia suficiente entre los micrófonos. Por parte, el post-procesado de las señales de audio entregadas por dos matrices multidimensionales de micrófonos es complicada y requiere una gran cantidad de circuitería o poder de procesado, un consumo de energía elevado, y con frecuencia dará como resultado un filtrado no deseado de la señal de salida.

El documento US 2005074129 divulga un dispositivo acústico que está provisto de un primer o primeros y un segundo o segundos elementos acústicos para generar una primera señal que incluye en su mayor parte audio no deseado y que está desprovista sustancialmente de audio deseado, y una segunda señal que incluye audio deseado así como no deseado respectivamente. El primer o primeros elementos acústicos están diseñados y dispuestos para generar un haz cardiode con un cero en una dirección de origen del audio deseado. El segundo o segundos elementos acústicos están diseñados y dispuestos para generar un haz complementario que incluye el audio deseado. Un sistema está provisto de un módulo lógico de procesado de señales apropiado para recuperar el audio

deseado utilizando la primera y segunda señales. El módulo lógico de procesado de señales puede poner en práctica técnicas de tipo cancelación de eco o técnicas de separación ciega de señales.

El documento US2007165879 divulga unas técnicas para mejorar señales de voz en un sistema de micrófono dual. Según un aspecto de este documento, se dispone de por lo menos dos micrófonos que están posicionados en una matriz pre-configurada. Se reciben dos señales de audio y las mismas se acoplan a un módulo de ajuste que está previsto para controlar la ganancia de cada una de las señales de audio con el fin de reducir al mínimo diferencias de señal entre las dos señales. Se proporciona un módulo de separación para recibir señales de audio adaptadas desde el módulo de ajuste. Se proporciona un módulo de filtrado adaptativo para eliminar el componente de ruido de la señal de audio con el fin de obtener una señal de voz estimada con una relación S/N mayor.

El documento JP2009260418 se refiere a un equipo acústico.

El documento US2008013758 divulga un módulo de micrófono que está conectado externamente a un dispositivo electrónico el cual incluye un primer conector de entrada de sonido. El módulo de micrófono incluye una matriz de micrófonos, un procesador de señal digital, y una primera parte de conexión. La matriz de micrófonos recibe sonido y genera una primera señal eléctrica. El procesador de señal digital recibe la primera señal eléctrica, lleva a cabo una conformación del haz y una supresión de ruido, y obtiene una segunda señal eléctrica. La primera parte de conexión se inserta en el primer conector de entrada de sonido, transmitiendo la segunda señal eléctrica al dispositivo electrónico.

El documento JP2007027939 se refiere a un procesador de señales acústicas.

El documento US2003125959 se refiere a un dispositivo traductor con una matriz plana de micrófonos. Formas de realización de la invención incluyen un dispositivo y un método para traducir palabras pronunciadas en un idioma en una versión gráfica o audible de las palabras en un segundo idioma. Una matriz plana de tres o más micrófonos se puede colocar en un dispositivo portátil, tal como un ordenador de mano o un asistente personal digital. La matriz plana, en combinación con un circuito de procesado de señales, define una dirección de sensibilidad. En un entorno con ruido, se seleccionan las palabras pronunciadas que se originan en la dirección de sensibilidad y se rechazan los otros sonidos. Las palabras pronunciadas son reconocidas y traducidas, y la traducción se muestra en una pantalla de visualización y/o se emite por medio de un altavoz.

El documento JP 4 324324 se refiere a un dispositivo y un método para medir una fuente de sonido en movimiento.

El documento JP 2009 188641 se refiere a un aparato emisor y captador de sonido.

Por lo tanto, una de las finalidades de la presente invención es desarrollar un dispositivo de mano mejorado que pueda capturar y procesar la voz, y generar una señal de voz con una relación de señal/ruido suficiente para aplicaciones fiables de reconocimiento del habla.

Otra de las finalidades de la invención es desarrollar un dispositivo basado en micrófonos con capacidad de mejorar la captación de la voz del usuario al mismo tiempo que reduciendo al mínimo el ruido de fondo y los posibles hablantes parásitos en condiciones difusas.

Las prestaciones del dispositivo deberían cubrir por lo menos el ancho de banda de voz medio, aunque también se deberían ampliar para mejorar el proceso de reconocimiento del habla, a saber [300Hz–6kHz].

Otra de las finalidades es desarrollar un dispositivo con capacidad de extraer información de voz útil (tal como una orden o un pedido en un restaurante) de entre el ruido de fondo el cual puede ser más o menos difuso (sin ángulo de incidencia predominante), más o menos intenso (en términos de niveles de presión de sonido) y con características espectrales variadas (música amplificada, voces individuales, “ruido cóctel”, etcétera).

Otra de las finalidades es desarrollar un dispositivo mejorado para captar voz procedente de la boca del hablante y ruido de otras direcciones, y que se pueda sostener en la palma de la mano del usuario.

Breve sumario de la invención

Según un aspecto de la invención, un dispositivo portátil de captura de voz comprende:

un brazo orientable adaptado para orientarse hacia la boca de un usuario, comprendiendo dicho brazo una primera matriz lineal diferencial de micrófonos, estando la directividad de dicha primera matriz dispuesta para captar voz de la boca de dicho usuario;

una segunda matriz lineal diferencial de micrófonos, estando la directividad de dicha segunda matriz dispuesta para captar ruido de una dirección diferente a la de la boca del usuario;

un circuito de reducción de ruido para proporcionar una señal de voz con ruido reducido, basándose en la salida de dicha primera matriz y en la salida de dicha segunda matriz.

En una forma de realización preferida, el dispositivo es un dispositivo de mano. En otra forma de realización, el dispositivo se puede conectar con otros equipos, incluyendo, sin carácter limitativo, un PC fijo, ordenadores portátiles, estaciones de trabajo, otros dispositivos móviles, tales como teléfonos móviles, y otros dispositivos. En una forma de realización preferida, la primera matriz diferencial se utiliza para capturar ruido de fondo desde una dirección posterior.

Se conocen matrices diferenciales de micrófonos en sí mismas, y se describen por ejemplo en "*Superdirectional Microphone Arrays*" de Elko, G.W., en "*Acoustic Signal Processing for Telecommunication*", J. Benesty y S. Gay (eds.), págs. 181 a 236, *Kluwer Academic Publishers,* 2000. La mayoría de matrices de micrófonos son relativamente voluminosas y no están adaptadas para dispositivos portátiles.

La presente invención se refiere a una disposición específica de matrices lineales, que permite capturar sonido según diferentes direcciones. El usuario puede orientar el brazo hacia su boca, y asegurarse de que la primera dirección está adaptada para capturar la voz del usuario, mientras que la segunda dirección captura esencialmente ruido de fondo. A continuación, el circuito de reducción de ruido puede mejorar la señal de voz eliminando el ruido de fondo, con el uso de, por ejemplo, técnicas de coherencia.

En una forma de realización, la primera matriz de micrófonos captura voz en una primera dirección frontal y ruido de fondo desde una dirección posterior, mientras que la segunda matriz de micrófonos captura ruido de fondo y otras voces procedentes de la derecha y la izquierda.

Otras formas de realización pueden usar más de dos matrices de micrófonos, y/o matrices complejas con el fin de proporcionar un mejor control sobre la directividad del dispositivo. Preferentemente, los micrófonos son, sin carácter limitativo, micrófonos de tipo electret.

En una forma de realización, el brazo tiene forma de L, y comprende una matriz lineal de micrófonos en cada una de las dos ramificaciones. En la invención pueden utilizarse también otras disposiciones, incluyendo micrófonos con una pluralidad de ramificaciones no perpendiculares, micrófonos en forma de U con tres matrices de micrófonos, o disposiciones con pares de micrófonos en ramificaciones diferentes de un brazo común.

Según otro aspecto, posiblemente independiente, de la invención, la señal de voz que proporciona a la salida el micrófono se somete a un post-procesado por medio de un filtro de post-procesado que comprende una pluralidad de capas de procesado de señal, permitiendo la extracción de la voz de entre el ruido, reduciendo el ruido residual, y evaluando la coherencia de la señal resultante con la captación de voz original.

Según otro aspecto, un detector de actividad vocal automático adicional permite una mejora adicional de la señal al eliminar segmentos de tiempo durante los cuales no se detecta la actividad vocal.

Las técnicas de post-procesado de señales de voz son conocidas como tales y han sido descritas, por ejemplo, por Kim KM, Choi YK, Park KS, "A new approach for rustle noise cancelling in pen-type voice recorder", *IEEE Transactions on Consumer Electronics,* vol. 49 (4), págs. 1.118 a 1.124, noviembre de 2003. Otro ejemplo de método de post-procesado ha sido descrito por O. Yilmaz y S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking", *IEEE Transactions on Signal Processing*, vol. 52 (7), págs. 1.830 a 1.847, julio de 2004. La combinación específica de métodos descritos y reivindicados se ha probado resultando ser particularmente eficaz para el fin antes mencionado, y ha resultado ser eficaz para eliminar ruido para una señal de voz capturada con el micrófono específico que se describe y reivindica en esta solicitud, al mismo tiempo que se evitan componentes innecesarios de hardware y software los cuales son requeridos por disposiciones más complejas.

Una ventaja clave del dispositivo que se divulga en la descripción y las reivindicaciones es su capacidad de ajustar la directividad con el fin de capturar y reconocer voz desde una distancia cómoda; el hablante puede hablar a una distancia cómoda (superior a 10 cm, preferentemente superior a 15 cm incluso en condiciones ruidosas, tales como en un restaurante) con respecto al dispositivo de mano.

En una forma de realización preferida, la matriz de micrófonos es suficientemente pequeña para garantizar la ergonomía y portabilidad del sistema, y no supera las dimensiones de los PDA o PC de bolsillo comunes (aproximadamente 150 mm x 70 mm, y en cualquier caso inferior a 180 x 100 mm).

Breve descripción de las figuras

La presente invención se entenderá mejor con la descripción de algunas formas de realización ilustradas por las figuras, en las cuales:

la figura 1 ilustra esquemáticamente un sistema para capturar y transmitir pedidos por voz en un restaurante.

la figura 2 ilustra esquemáticamente un sub-sistema de micrófonos.

la figura 3 es un diagrama que muestra la influencia de μ sobre el patrón de directividad de un sub-sistema de primer orden de micrófonos.

la figura 4 es un diagrama que muestra la dependencia de la sensibilidad de una matriz diferencial de primer orden con respecto al ángulo y la frecuencia.

la figura 5 ilustra esquemáticamente una matriz diferencial de segundo orden.

la figura 6 ilustra un ejemplo de dispositivo que comprende una configuración de micrófonos diferencial y bidimensional (a la izquierda: matriz de radiación transversal (*broadside*); a la derecha: matriz orientable de radiación longitudinal (*endfire*)).

la figura 7 ilustra esquemáticamente la disposición de micrófonos en el dispositivo de la figura 6.

la figura 8 es un diagrama de flujo que ilustra una posibilidad de combinación de varios filtros y métodos utilizados después de la conformación de haz para el post-procesado con el fin de potenciar la voz y/o amortiguar el ruido.

Descripción detallada de las formas de realización preferidas

La siguiente descripción se aporta poniendo énfasis en la forma de realización basada en PC de bolsillo para la anotación de pedidos por voz en restaurantes. No obstante, el dispositivo de la invención también se puede utilizar con otros equipos, incluyendo, sin carácter limitativo, PC fijos, ordenadores portátiles, estaciones de trabajo, otros dispositivos móviles, tales como teléfonos móviles, y otros dispositivos, y para otras aplicaciones diferentes de restaurantes y bares (industria de la hostelería, hospitales, industria del entretenimiento, tiendas de comestibles, laboratorios, etcétera).

En la figura 1 se ilustra un ejemplo de entorno en el cual se pueden usar el método y el dispositivo. En este escenario, un camarero 2 en un bar o restaurante anota un pedido de clientes 3 que están sentados en una mesa. El camarero repite cada pedido y lo pronuncia al micrófono de su dispositivo móvil 1. En esta forma de realización, la señal de voz grabada se somete a un post-procesado localmente, por ejemplo mediante el procesador del dispositivo móvil 1 ó a través de medios de procesado dedicados, con el fin de mejorar la relación de señal/ruido. Este post-procesado también se podría realizar por medio de un ordenador remoto o servidor en otra forma de realización, aunque esto es probable que introduzca un retardo. A continuación, la señal de voz procesada se transmite por vía aérea a un punto de acceso 7, utilizando un protocolo normalizado de comunicaciones inalámbricas, tal como el 802.11, Bluetooth, etcétera. El punto de acceso 7 pertenece a una red de área local 8 (LAN), a la cual están conectados otros diversos equipos, tales como un ordenador personal 5, un servidor 6, etcétera. La señal de voz recibida desde el punto de acceso 7 se convierte en órdenes de texto por medio del servidor 6 el cual ejecuta un algoritmo de reconocimiento de habla. El algoritmo de reconocimiento de habla podría ser ejecutado por el dispositivo móvil si este dispositivo dispone de suficiente poder de procesado; no obstante, esto puede hacer que una actualización de los modelos de habla y de idioma (tales como la lista de órdenes a reconocer, y la gramática asociada) resulte más dificultosa.

En una forma de realización preferida, el reconocimiento del habla depende del hablante y utiliza plantillas dependientes del hablante almacenadas en una base de datos 60. En esta base de datos 60 se almacenan un diccionario y una gramática para limitar el número de palabras o expresiones a reconocer, y para definir algunas reglas que caracterizan el texto hablado por el camarero. Esta gramática se actualiza de manera ventajosa cada vez que se proponen nuevos productos a los clientes 3, por ejemplo cada vez que se modifica el menú del restaurante. La gramática y el diccionario están adaptados ventajosamente para aplicaciones de “mando y control” y/o para la anotación de pedidos en restaurantes.

Para esta aplicación, el algoritmo de reconocimiento del habla se basa ventajosamente en un clasificador estadístico, tal como una red neuronal, combinado con un clasificador basado en plantillas. Los ensayos han demostrado que este escenario proporciona una velocidad de reconocimiento mejorada y una introducción sencilla de palabras o expresiones nuevas en la gramática. La gramática puede incluir unidades de reconocimiento por plantillas de tamaños diferentes (expresión, frase, palabra, fonema). También puede utilizarse una gramática dependiente del usuario.

La gramática y/o el clasificador son preferentemente adaptativos, y aprenden unidades de reconocimiento por plantillas, incorporadas en la entrada pronunciada. Esto permite el aprendizaje en línea de palabras nuevas o de otras plantillas. Se puede usar una retroalimentación del usuario, por ejemplo en el dispositivo del usuario, para introducir o seleccionar el texto equivalente de una plantilla recién aprendida.

Además, la gramática está dispuesta ventajosamente en categorías y sub-categorías independientes; esto hace que mejore la calidad del reconocimiento del habla puesto que el sistema conoce la categoría de la siguiente plantilla que está esperando. Esto también hace que la introducción manual de plantillas nuevas sea más sencilla. Por ejemplo, una categoría de plantillas se puede corresponder con la lista de vinos, y otra categoría con los postres.

El texto reconocido por el sistema de reconocimiento de habla en el servidor 6 se transmite a través de la LAN 8 y sobre el canal inalámbrico de vuelta al dispositivo 1 del camarero, y se visualiza en tiempo real. En otro entorno, el reconocimiento se podría realizar directamente en el dispositivo del camarero. El camarero puede comprobar si el reconocimiento fue correcto, y confirmar o corregir la orden reconocida por el servidor y visualizada por el dispositivo. Esta retroalimentación del usuario se puede usar para adaptar la plantilla dependiente del hablante, la gramática, y/o para añadir nuevas unidades de reconocimiento.

Cuando el nivel de confianza alcanzado por el algoritmo de reconocimiento de habla se encuentra por debajo de un nivel predefinido o cuando hay varias opciones posibles que están muy próximas entre sí, se visualiza un menú con una lista de múltiples elecciones de las entradas más probables, para el camarero, el cual puede seleccionar la orden deseada en este menú, utilizando por ejemplo una pantalla táctil, un lápiz táctil, o cualesquiera otros medios adecuados de entrada incluyendo la voz. El camarero también puede seleccionar otras opciones, por ejemplo para especificar la cantidad de productos pedidos (número o volumen), el tipo (por ejemplo la añada de un vino, las preferencias del cliente en relación con la cocción, etcétera), en función del producto pedido o de si el pedido inicial no fue suficientemente preciso.

Una vez validado por el camarero, este texto, y la respuesta de los camareros al menú de opciones, se visualiza también en un ordenador personal 5 ó se imprime y es leído por el personal del restaurante, con el fin de preparar y servir el pedido solicitado. En otra forma de realización, este texto se pronuncia en la cocina. La lista de productos pedidos se puede almacenar en una base de datos del servidor 6, la cual se puede utilizar posteriormente para prepararle la factura al cliente. En una forma de realización alternativa, la señal de voz grabada se somete a un post-procesado por parte de un ordenador o servidor.

En una forma de realización alternativa, el reconocimiento del habla se lleva a cabo localmente, en el dispositivo 1 del usuario. No obstante, esto requiere dispositivos 1 con un mayor poder de procesado, y una sincronización más dificultosa de los modelos dependientes del hablante en caso de que un usuario utilice varios dispositivos diferentes.

En la figura 6 se ilustra un ejemplo de dispositivo 1 según la invención. El mismo se construye ventajosamente en torno a un PDA (asistente personal digital) convencional, un *netbook* o un dispositivo similar. Comprende:

- una caja adaptada para llevar y manipular el dispositivo en la mano del usuario;
- un módulo de visualización 21 para presentarle visualmente al usuario 2 el texto reconocido, y otro texto o imágenes;
- medios hápticos 22, tales como un teclado numérico, un teclado normal, un botón táctil, una rueda de selección, etcétera;
- una interfaz de comunicaciones (no mostrada), por ejemplo una interfaz WLAN y/o Bluetooth;
- medios de procesado (no mostrados), tales como un microprocesador con una memoria adecuada volátil y no volátil, para procesar el audio de la señal de audio capturada con el micrófono, y para ejecutar otros programas y funciones;
- un brazo orientable en forma de L 23 que incluye varias matrices lineales de micrófonos 24, 25 con diferente separación entre los micrófonos de cada matriz. El uso de una pluralidad de matrices de micrófonos proporciona una mejora de la captación de la voz, y un control de directividad de banda ancha. El brazo está conectado a la caja a través de una conexión giratoria, con el fin de dirigir el tramo más largo de manera precisa hacia la boca del hablante.

El brazo 23 es ventajosamente un accesorio el cual está adaptado para una instalación “posterior” y un montaje semi-permanente en un dispositivo móvil existente. A este brazo se le puede asociar una circuitería electrónica, tal como conversores analógicos-a-digitales, retardos, sumadores, etcétera, y/o procesadores de señales digitales (DSP) o FPGA, para procesar señales de audio obtenidas a la salida de las matrices de micrófonos. Este accesorio (brazo extraíble con circuitería opcional) se puede comercializar por separado con respecto al dispositivo móvil, y se puede instalar posteriormente en un dispositivo móvil existente con el fin de transformarlo en un dispositivo de acuerdo con la invención. La instalación también puede incluir instalación de controladores adecuados y software de aplicación en el dispositivo móvil, para recuperar señales del accesorio, someter a post-procesado dichas señales, enviarlas al servidor remoto o al dispositivo móvil, y visualizar la retroalimentación del servidor. La conexión eléctrica entre el brazo y el dispositivo utiliza preferentemente una interfaz existente del dispositivo móvil, por ejemplo un USB, un RS-232 ó un *socket* privativo, o una conexión inalámbrica.

En otra forma de realización no ilustrada, el brazo con las matrices de micrófonos y la circuitería electrónica asociada está conectado a un dispositivo móvil existente a través de una interfaz inalámbrica, por ejemplo una interfaz Bluetooth o Zigbee. En este caso, el brazo se puede desmontar del dispositivo móvil, y se puede manipular por separado. También es posible dividir el brazo en varias partes, y usar uno de los tramos como lápiz táctil sostenido hacia la boca y conectado (inalámbricamente por cable) a las otras partes y/o al dispositivo móvil. Por otra parte, el brazo, o cada parte del brazo, puede ser un componente completamente pasivo que incluye solamente micrófonos, o una parte "inteligente" que tenga un microprocesador, una FPGA o un procesador de señales. Las diferentes partes se pueden conectar entre sí, y se pueden conectar al dispositivo móvil y/o a un módulo receptor del dispositivo móvil, a través de una interfaz por cable o inalámbrica. Por otra parte, el micrófono o partes del micrófono, y/o el dispositivo móvil, se pueden conectar remotamente desde un módulo de mando a distancia con el fin de controlar la amplificación, la reducción del ruido, la direccionalidad, etcétera. En una forma de realización, el sistema comprende medios de procesado de la señal los cuales están distribuidos entre el brazo, o entre diferentes partes del brazo, y el dispositivo móvil.

En la figura 2 se ilustra un ejemplo de matriz lineal de micrófonos 24. Esta matriz simple comprende dos micrófonos 240, 241 separados por una distancia d. Las señales de salida de un micrófono se suma algebraicamente con un elemento sumador 243 a la señal de salida retardada del otro micrófono distante en d, indicándose como τ_e el retardo aplicado por el elemento de retardo 242. Esta matriz forma un sistema de conformación de haz; una elección adecuada del retardo τ_e mejora la relación señal/ruido y mejora la sensibilidad a señales de audio provenientes según la dirección de la matriz lineal.

Si se considera una señal acústica entrante con un ángulo de incidencia θ (referido al eje del sub-sistema), y suponiendo una señal armónica de frecuencia f[Hz] (o pulsación $\omega=2\cdot\pi\cdot f$), el "retardo acústico" entre los dos micrófonos es $\tau_d = (d\cdot\cos\theta)/c$ [s] (siendo c la celeridad del sonido en el aire) y el voltaje de salida resultante $\underline{U}$ [V] del sub-sistema depende del ángulo de incidencia θ [rad]:

$$\underline{U} = \underline{U}_1 - \underline{U}_2 e^{(-j\omega\tau_e)} = \underline{M}_1 \underline{p}_1 \left(1 - e^{-j\omega(\tau_e + \tau_d \cos\theta)}\right) \cong \underline{M}_1 \underline{p}_1 j\omega(\tau_e + \tau_d \cos\theta) \quad (1)$$

donde $\underline{M}_1$ [V/Pa] es la sensibilidad del primer micrófono, $\underline{p}_1$ [Pa] la presión acústica de una onda plana en el primer micrófono, τ_e [s] el retardo aplicado al segundo micrófono y τ_d el tiempo de propagación desde el primer al segundo micrófono. Con $\tau = \tau_e + \tau_d$ y $\mu = \tau_d/\tau$, se tiene finalmente la sensibilidad $\underline{M}$ del sub-sistema:

$$\underline{M} = \frac{\underline{U}}{\underline{p}} \cong \underline{M}_1 j\omega[(1-\mu) + \mu\cos\theta] \quad (2)$$

que es la característica de un micrófono directivo de primer orden.

A partir de esta ecuación, se observa que la respuesta en frecuencia se corresponde con un filtro paso-alto con una pendiente de +6 dB/octava. Esto significa que la sensibilidad se reduce en la banda de frecuencias bajas, lo cual puede ser desventajoso.

Fijando μ=0,5, se obtiene una directividad cardiode de la matriz de micrófonos, y con μ=1, un micrófono bidireccional. La figura 3 muestra los patrones de directividad característicos para diferentes valores de μ.

La directividad depende grandemente de la frecuencia tal como se ilustra mediante la figura 4. Para garantizar un patrón de directividad constante en el ancho de banda completo de las frecuencias, en las matrices de micrófonos 24, 25 se combinan diferentes pares de redes con diferentes distancias entre micrófonos dentro de los pares, y diferentes limitaciones de frecuencia.

Así, el brazo del micrófono de la invención utiliza varios pares de micrófonos los cuales están dispuestos a lo largo del mismo eje para obtener una matriz más directiva (en el eje de la matriz). Por lo tanto, cada matriz es monodimensional y comprende una pluralidad de pares dispuestos, todos ellos, en una fila.

Combinando dos matrices diferenciales de primer orden y después de introducir un retardo de tiempo adicional, se puede construir una matriz de micrófonos diferencial de segundo orden, genérica. La sensibilidad total de dicho sistema se puede calcular multiplicando las sensibilidades de los subsistemas implicados, lo cual conduce a una directividad mejorada con dos subsistemas en cascada en comparación con solamente uno, aunque con el inconveniente del comportamiento de un filtro pasoalto de 2º orden. Seleccionando las dimensiones de cada sub-sistema, se pueden cubrir anchos de banda de frecuencia más amplios con directividades y sensibilidades constantes, construyendo así matrices diferenciales.

Una matriz diferencial se describe por su orden, es decir, el número de las “etapas” de retardos, tal como se describe en la figura 5 para una matriz de segundo orden 24. En este ejemplo, la matriz comprende N=3 micrófonos dispuestos en cuatro pares: {1;2}, {2;3}, {2;1}, {3;2}. El primer dígito de cada par se refiere al signo “+” y el segundo dígito al signo “-” de los elementos sumadores 242_1 a 242_4 de la figura 5. Las distancias d_i entre micrófonos sucesivos dentro de los pares son variables.

La señal analógica u1(t), ui(t),..., uN(t) en la salida de cada micrófono 240, 241, 244 se convierte en una señal digital mediante conversores analógicos-a-digitales respectivos 245_1, 245_2, 245_3. A continuación, para cada par, una primera etapa de procesado 246 lleva a cabo la suma algebraica digital entre una señal y la señal retardada del otro micrófono del par. A continuación, una segunda etapa de procesado 247 lleva a cabo la suma algebraica entre la señal de un elemento sumador 243 y la salida retardada de otro elemento sumador de la primera etapa. La primera señal digital entregada por esta segunda etapa forma una señal de haz frontal 248, mientras que la otra señal digital entregada por esta segunda etapa forma una señal de haz posterior 249.

En teoría, se pueden combinar tantos pares como se desee, aunque en la práctica, resulta difícil llegar más allá de una matriz de segundo orden. Esto es debido principalmente al hecho de que una matriz diferencial es una matriz diferenciadora (filtro pasoalto) del mismo orden que el orden de la matriz, lo cual significa que las frecuencias bajas de la señal están fuertemente atenuadas y significa también por cierto una reducción de la relación señal/ruido. Existe por lo tanto un compromiso al que se debe llegar en relación con las dimensiones de cada matriz, el orden de la matriz, el ancho de banda de frecuencias que interesa, y el número de canales disponibles para el procesado de las señales.

El brazo del micrófono del dispositivo 1 está dispuesto para detectar sonido no solamente procedente de la dirección útil (dirección de la boca), sino también procedente de por lo menos otra dirección, correspondiente al ruido. Un mejor conocimiento del ruido que proviene de direcciones diferentes permite llevar a cabo la extracción de la señal útil y el rechazo de la señal de ruido, utilizando técnicas de coherencia. Permite también mejorar la eficiencia del post-filtrado subsiguiente.

En una forma de realización, el brazo del micrófono 23 de la presente invención comprende ventajosamente una matriz de micrófonos bidimensional (en lugar de una matriz unidimensional como la que se ha descrito hasta el momento). Esta matriz bidimensional está constituida por dos matrices unidimensionales, tal como se ilustra en la figura 7. Una primera matriz 24 está dispuesta en el primer tramo, el más largo, del brazo en forma de L 23, mientras que la segunda matriz está dispuesta en el otro tramo, más corto, del mismo brazo. Esta segunda matriz de micrófonos transversal se utiliza para mejorar la cancelación del ruido interferente.

Tal como se ha mencionado, este brazo en forma de L es orientable, por rotaciones alrededor del eje de uno de los dos tramos (en este caso el más corto), de manera que el usuario puede ajustar la posición a un punto óptimo (delante de la boca). Cuando el brazo 23 está orientado correctamente, el tramo más largo (en este ejemplo) capta la señal frontal útil procedente de la dirección de la boca del hablante, así como ruido de la parte posterior. El segundo tramo (en este caso el más corto, aunque no de forma necesaria) capta ruido difuso procedente de las direcciones izquierda y derecha.

En el escenario ilustrado, la orientación del segundo tramo permanece esencialmente sin cambios cuando el brazo se hace girar; existe solamente un grado de libertad para orientar el primer tramo en la dirección de la boca del usuario.

En una forma de realización preferida, los dos tramos son perpendiculares entre sí; no obstante son posibles otras disposiciones.

Cada tramo está equipado de por lo menos una matriz diferencial lineal de micrófonos.

En otra forma de realización, el micrófono tiene forma de U y comprende dos tramos conectados mediante un tercer tramo, preferentemente perpendicular, aunque sin carácter limitativo, a los dos primeros tramos.

El dispositivo de la invención puede usar además micrófonos o matrices de micrófonos adicionales, incluyendo micrófonos o matrices de micrófonos no orientables en la caja del dispositivo para capturar ruido de fondo procedente de diferentes direcciones.

Además, micrófonos de tramos diferentes se pueden emparejar para proporcionar una captación adicional del ruido difuso según otras direcciones.

La figura 8 es un diagrama de flujo que ilustra una posibilidad de combinación de varios filtros y métodos utilizados después de la conformación de haz para el post-procesado con el fin de potenciar la voz y/o amortiguar el ruido.

En referencia a la figura 8, en una primera etapa, se aplican métodos de conformación de haz (según se ha descrito anteriormente) para reducir el ruido y controlar la directividad, calculando sumas algebraicas entre señales entregadas por diferentes micrófonos o subsistemas de micrófonos.

Una de las señales entregadas por las matrices de micrófonos, por ejemplo señales del haz, contiene principalmente la voz del usuario mientras que las señales de las otras matrices de micrófonos contienen principalmente ruido. En una forma de realización preferida, el haz frontal contiene la voz del usuario mientras que los haces posterior, izquierdo y derecho contienen señales de fuentes de ruido. Las diferentes señales entregadas por las diferentes matrices en el micrófono se someten a continuación a un post-procesado para entregar una señal de voz con una mejor relación señal/ruido y adecuada como entrada para un software de reconocimiento de habla. Las etapas que suceden a la conformación del haz se podrían llevar a cabo en un orden diferente al ilustrado en la figura 8.

El post-procesado puede incluir una estimación de características espectrales del ruido durante una cierta franja de tiempo. Este módulo puede actuar sobre haces con ruido de fondo y/o el haz con voz. En este último caso, se debe llevar a cabo mientras el usuario no está hablando. El tiempo necesario para obtener la característica espectral del ruido puede variar en función de la aplicación; con el fin de abordar aplicaciones de tiempo real, como la anotación de pedidos en restaurantes, el cálculo de una estimación del ruido debe realizarse cuando el usuario 2 ordena el pedido, es decir, una fracción de segundo antes de que el usuario comience a hablar.

El post-procesado puede incluir un filtro de Wiener que lleva a cabo una resta del espectro de ruido con respecto al espectro del haz de voz, estimados o bien para el haz de voz y/o bien los haces de ruido.

El post-procesado puede incluir una etapa de post-filtrado en la cual el espectro del haz de voz se compara con los espectros de otros haces en diversas frecuencias y se amortigua y/o elimina en estas frecuencias cuando no sea más de k veces mayor que el más alto de los espectros de haces de ruido.

El post-procesado puede incluir un filtrado del haz de voz basándose en una medición, en el dominio espectral, de su coherencia con el micrófono que esté más próximo a la boca del hablante. El espectro del haz de voz se amortigua y/o elimina en estas frecuencias cuando la coherencia con el espectro del micrófono es baja.

En otra forma de realización, la etapa de post-filtrado implica una comparación, en el dominio de la frecuencia, de las cuatro señales entregadas por la matriz de micrófonos (frontal, posterior, izquierda, derecha), calculadas con la fase de conformación de haz y desprovistas de ruido con la fase de reducción de ruido utilizando filtros adaptativos basados en una DUET (Degenerate Unmixing Estimation Technique) modificada. Para cada canal del conformador de haz, estos filtros adaptativos permiten reducir la influencia del ruido en el canal frontal, mediante resta espectral de las señales de los otros tres canales que están captando básicamente ruido.

En otra forma de realización, la etapa de post-filtrado implica un cálculo de coherencia el cual se lleva a cabo entre la señal frontal entregada por el conformador de haz y el resultado del post-filtrado, con el fin de filtrar señales residuales que no provienen del hablante. Dos señales son coherentes si una de ellas es una versión a escala y retardada de la otra.

La etapa de post-filtrado puede implicar un Detector de Actividad Vocal (VAD) para detectar cuándo está hablando el usuario. La detección vocal se lleva a cabo preferentemente mediante análisis de la señal de potencia.

En una forma de realización, el Detector de Actividad Vocal es multicapa y está integrado en el dispositivo (VAD multicapa incorporado). Puede disponer de por lo menos uno de los siguientes medios:

- medios de “Mantenimiento del habla”: permiten evitar o limitar la cancelación del inicio y/o el final del habla y evitar o limitar cortes en el habla que se podrían producir como consecuencia de los umbrales del VAD. Cuando la energía de la señal medida sube por encima de un cierto nivel, el sistema considera que hay presencia de habla. Para evitar que se pierde el inicio del habla que pudiera estar por debajo de los umbrales del VAD, el VAD no cancela un espacio de tiempo parametrizado antes de la detección del habla (el sistema transmite la grabación con un pequeño retardo para permitir la toma de dicha decisión). Para evitar cortes y/o que se pierda el final del habla, los umbrales del VAD se desactivan mientras el sistema está en presencia de habla. Los mismos se activan nuevamente cuando el habla permanece por debajo de los umbrales del VAD durante un espacio de tiempo parametrizado dado.

- medios de estimación de ruido: permiten determinar el nivel de ruido. El tiempo necesario para llevar a cabo la estimación del ruido puede variar en función de la aplicación, aunque ya se puede realizar en una fracción de segundo para abordar aplicaciones de tiempo real como la anotación de pedidos en restaurantes. En este caso, la estimación del ruido se efectúa cuando el usuario 2 ordena el pedido, una fracción de segundo antes de que comience a hablar.

- medios de VAD Relativo: fijan el nivel por encima del cual el sonido se considera como habla de acuerdo con la estimación del ruido. El sonido se considera como habla cuando alcanza n veces el nivel del ruido determinado

con la estimación del ruido. Esto se usa para eliminar el ruido y cualesquiera variaciones de ruido que puedan estar por encima de la estimación de ruido pero permanecen por debajo de n dicha estimación de ruido.

- medios de VAD Absoluto: fijan un nivel absoluto por encima del cual el sonido se considera como habla. Esto se usa para cancelar ruidos pequeños como el toque de una pantalla, música baja o algunos ruidos reducidos de fondo.

La decisión de presencia del habla con medios de VAD relativo y absoluto se realiza sobre la base de un periodo de tiempo breve, típicamente 10 ms, para evitar una interrupción corta del habla cuando se requiera un espacio de tiempo de cancelación mínimo con el fin de tener en cuenta la cancelación. Esta restricción se relaja en situaciones de inicio del habla y final del habla.

El VAD absoluto es especialmente necesario cuando hay poco ruido, ya que el VAD relativo podría no fijar el umbral en un nivel que permita cancelar ciertas variaciones de ruido.

En una forma de realización, el VAD Relativo también se podría calcular de una manera no lineal. Por ejemplo, si el nivel de estimación del ruido es muy alto, como en el caso de un restaurante muy ruidoso, el valor de n podría ser menor que cuando el nivel de estimación del ruido es bajo.

En otra forma de realización, el sistema con VAD incorporado podría incluir múltiples umbrales de VAD absoluto y relativo los cuales se pueden activar y desactivar en función de ciertos criterios de aplicación.

Este dispositivo se puede usar, por ejemplo, para aplicaciones de anotación de pedidos de voz y aplicaciones de reconocimiento del habla en restaurantes, bares, discotecas, hoteles, hospitales, industria del entretenimiento, tiendas de comestibles, etcétera.

REIVINDICACIONES

1. Dispositivo de mano de captura de voz (1), que comprende:

un brazo orientable (23), que comprende un primer tramo y un segundo tramo, presentando dicho primer tramo y dicho segundo tramo unas orientaciones diferentes, comprendiendo dicho primer tramo de dicho brazo (23) una primera matriz lineal diferencial (25) de micrófonos, estando la directividad de dicha primera matriz lineal diferencial (25) dispuesta para una captación mejorada de voz procedente de un usuario;

comprendiendo dicho segundo tramo de dicho brazo una segunda matriz lineal diferencial (24) de micrófonos, estando la directividad de dicha segunda matriz lineal diferencial (24) dispuesta para una captación mejorada de ruido procedente de una dirección diferente a la de dicha voz captada;

un circuito de reducción de ruido para proporcionar una señal de voz con ruido reducido, sobre la base de la salida de dicha primera matriz lineal diferencial (25) y de la salida de dicha segunda matriz lineal diferencial (24).

2. Dispositivo según la reivindicación 1, en el que dicho circuito de reducción de ruido se basa en técnicas de coherencia para eliminar ruido de la salida de dicha primera matriz lineal diferencial (25).

3. Dispositivo según la reivindicación 1, que comprende una conexión giratoria para hacer girar dicho brazo orientable (23) alrededor del eje de uno de dichos tramos.

4. Dispositivo según una de las reivindicaciones 1 a 3, presentando dicho brazo forma de L, estando la primera matriz lineal diferencial (25) dispuesta en un primer tramo y la segunda matriz lineal diferencial (24) en un segundo tramo de dicho brazo con forma de L, pudiendo dicho brazo orientable (23) ser girado alrededor de un eje paralelo a uno de dichos tramos.

5. Dispositivo según una de las reivindicaciones 1 a 3, presentando dicho brazo orientable (23) forma de U y comprendiendo tres matrices de micrófonos.

6. Dispositivo según una de las reivindicaciones 1 a 5, que además comprende:

unos medios de procesado de datos;

un módulo de visualización (21);

una interfaz de comunicaciones inalámbricas;

un filtro de Wiener para reducción del ruido;

un detector de actividad vocal.

7. Dispositivo según la reivindicación 6, siendo el detector de actividad vocal un detector de actividad vocal multicapa incorporado que comprende por lo menos uno de los siguientes medios:

unos medios de estimación de ruido;

unos medios de detector absoluto de actividad vocal;

unos medios de detector relativo de actividad vocal;

unos medios de mantenimiento del habla.

8. Dispositivo según una de las reivindicaciones 1 a 7, conectado funcionalmente a un módulo de reconocimiento del habla dependiente del usuario.

9. Dispositivo según la reivindicación 8, comprendiendo dicho módulo de reconocimiento del habla dependiente del usuario una gramática y un diccionario adaptados para aplicaciones de “mando y control”, y/o para la anotación de pedidos en restaurantes.

10. Método para capturar voz, que comprende:

capturar una señal de voz con una primera matriz lineal diferencial (25) de micrófonos montados en un primer tramo de un brazo orientable (23) de un dispositivo de mano (1), estando dicho brazo orientable (23) dirigido hacia la boca de un usuario (2);

capturar el ruido procedente de por lo menos una dirección diferente a la dirección de dicha señal de voz, usando una segunda matriz lineal diferencial (24) de micrófonos montados en un segundo tramo de dicho brazo orientable (23) de manera que dicho primer y segundo tramos presenten direcciones diferentes;

reducir ruido de dicha señal de voz, usando la salida de dicha primera matriz lineal diferencial (25) y de dicha segunda matriz lineal diferencial (24).

11. Método según la reivindicación 10, en el que la directividad de dicha primera y segunda matrices lineales diferenciales (24, 25) de micrófonos capturan el habla a una distancia con respecto a la boca superior a 15 cm en condición de ruido.

12. Sistema que comprende un asistente digital portátil como dispositivo de mano de captura de voz (1) según cualquiera de las reivindicaciones 1 a 9, siendo dicho brazo orientable (23) un accesorio externo extraíble montado en dicho asistente digital portátil.

13. Sistema que comprende un asistente digital portátil como dispositivo de mano de captura de voz (1) según cualquiera de las reivindicaciones 1 a 9, estando conectado inalámbricamente dicho brazo orientable (23) a dicho asistente digital portátil.

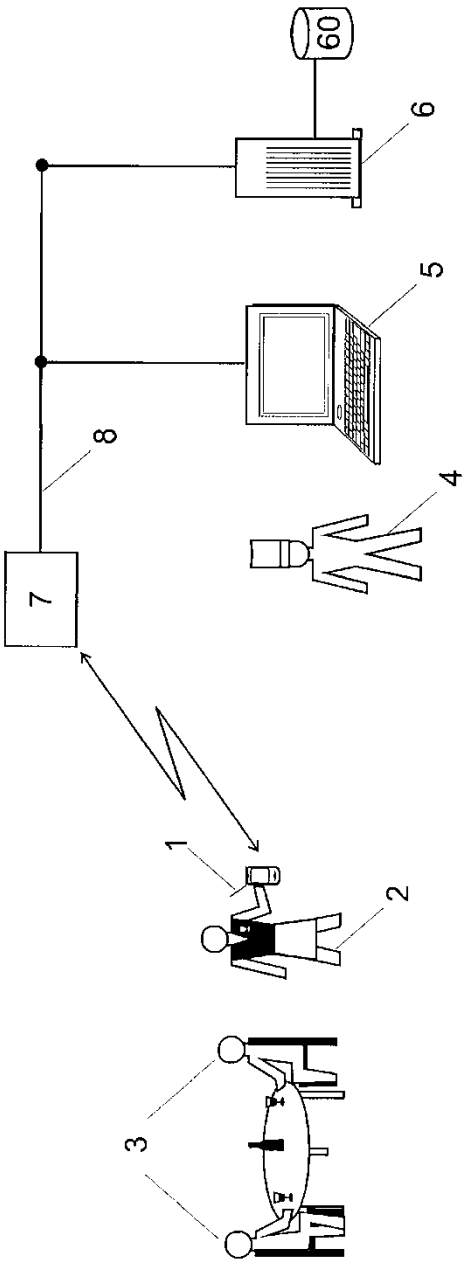

Fig. 1

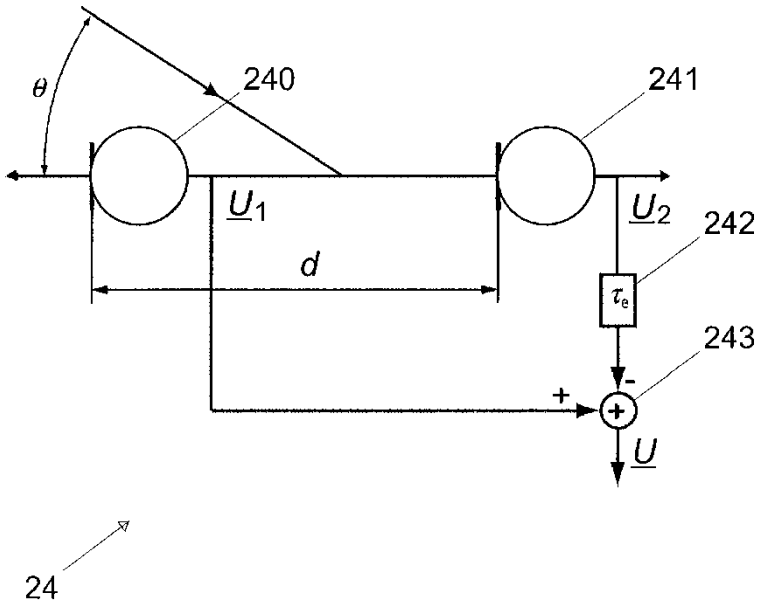

Fig. 2

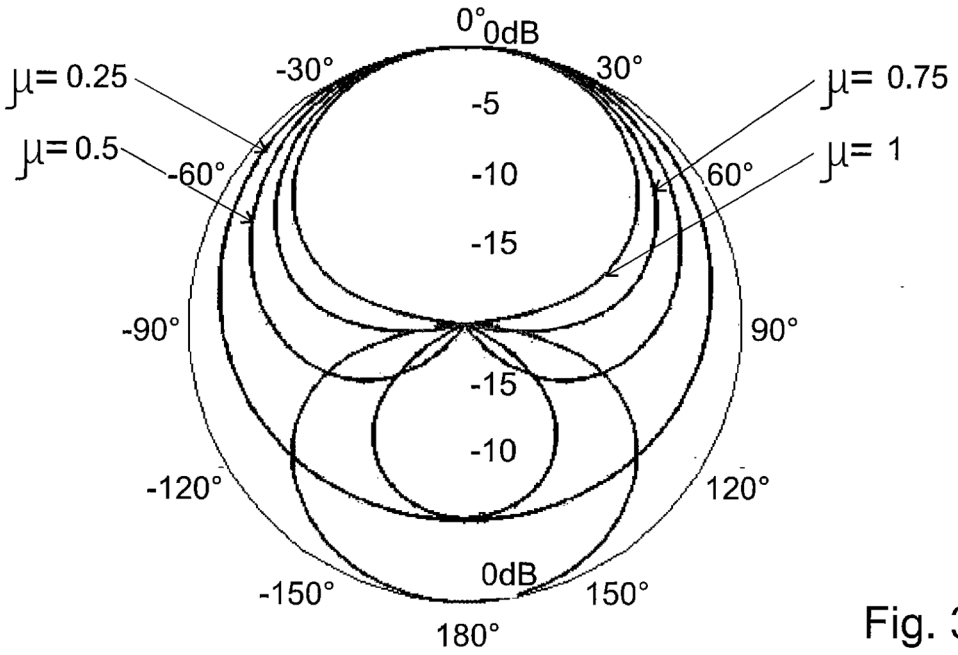
0°
0dB
-30°
30°
μ= 0.25
μ= 0.75
-5
μ= 0.5
μ= 1
-60°
60°
-10
-15
-90°
90°
-15
-10
-120°
120°
0dB
-150°
150°
180°

Fig. 3

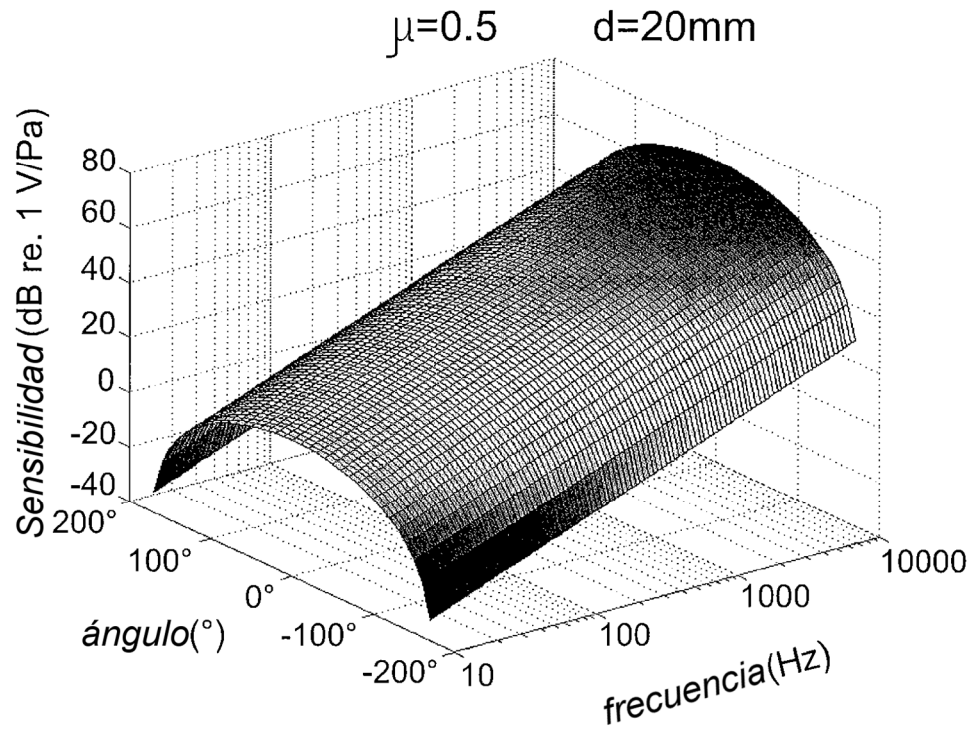
μ=0.5
d=20mm
Sensibilidad (dB re. 1 V/Pa)
80
60
40
20
0
-20
-40
200°
100°
0°
-100°
-200°
ángulo(°)
10
100
1000
10000
frecuencia(Hz)

Fig. 4

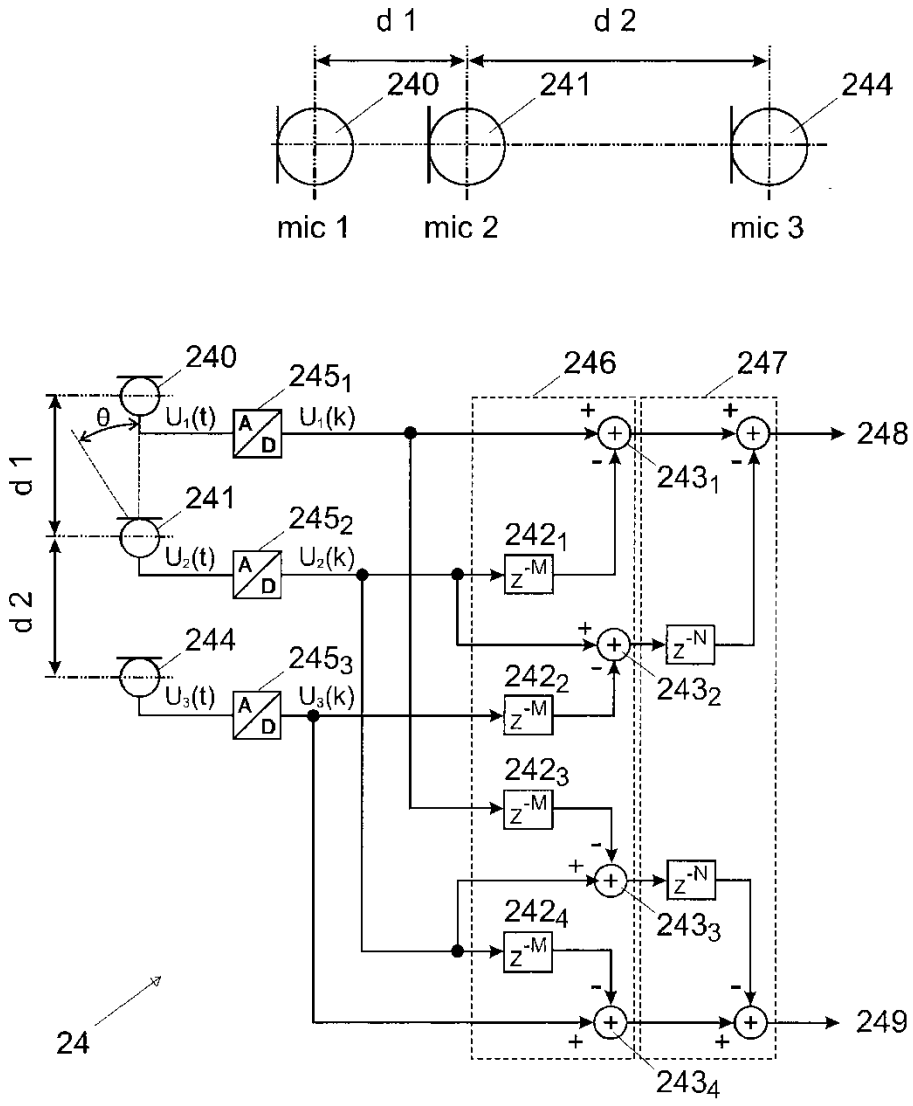
d 1
d 2
240
241
244
mic 1
mic 2
mic 3
240
245₁
246
247
θ
U₁(t)
A/D
U₁(k)
248
243₁
241
245₂
242₁
U₂(t)
U₂(k)
z^-M
z^-N
243₂
244
245₃
242₂
U₃(t)
U₃(k)
242₃
243₃
242₄
249
243₄
d 1
d 2
24

Fig. 5

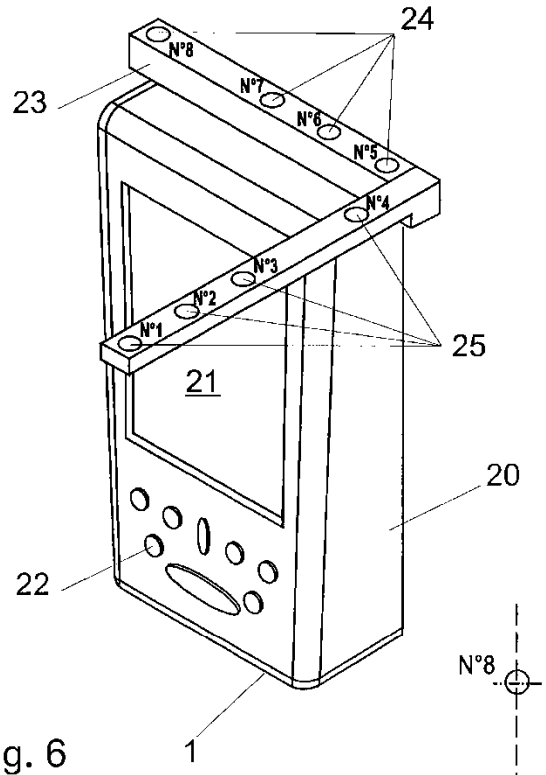

Fig. 6

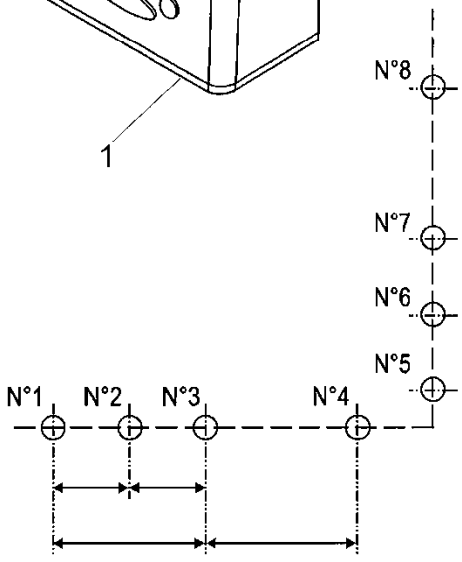

Fig. 7

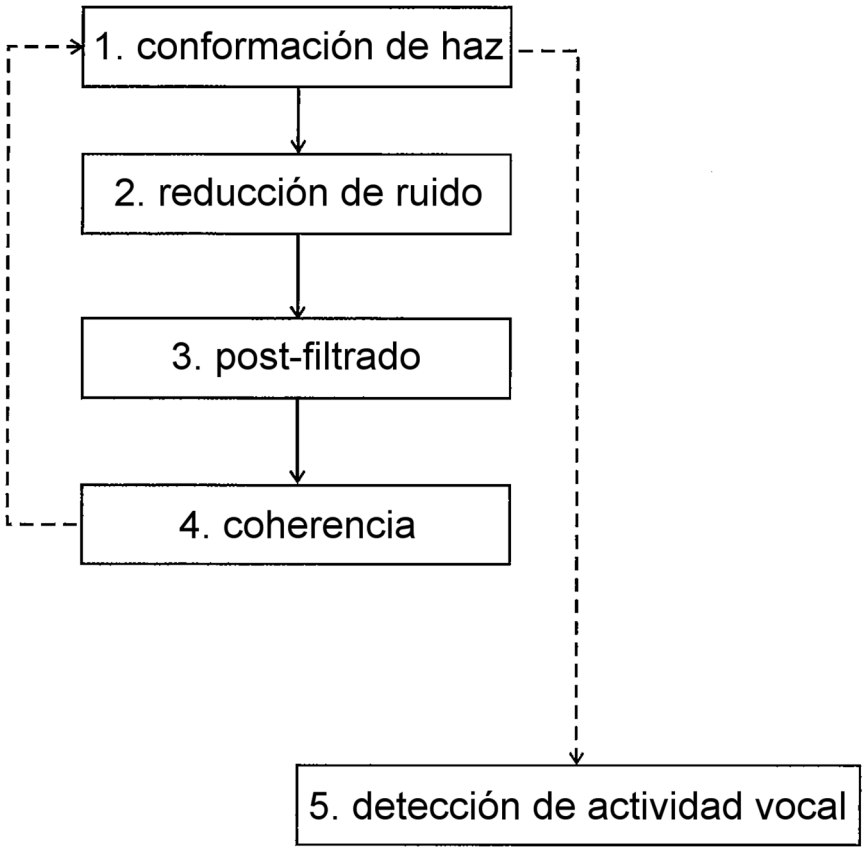
1. conformación de haz
2. reducción de ruido
3. post-filtrado
4. coherencia
5. detección de actividad vocal

Fig. 8