

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 728 708**

51 Int. Cl.:

G10L 13/10

(2013.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **19.08.2014** **PCT/GB2014/052544**

87 Fecha y número de publicación internacional: **26.02.2015** **WO15025155**

96 Fecha de presentación y número de la solicitud europea: **19.08.2014** **E 14756113 (8)**

97 Fecha y número de publicación de la concesión europea: **13.03.2019** **EP 3036737**

54 Título: **Sistema y procedimiento de diálogo audiovisual**

30 Prioridad:

23.08.2013 GB 201315142

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

28.10.2019

73 Titular/es:

**UCL BUSINESS PLC. (100.0%)
The Network Building 97 Tottenham Court Road
London W1T 4TP, GB**

72 Inventor/es:

**LEFF, JULIAN;
WILLIAMS, GEOFFREY y
HUCKVALE, MARK**

74 Agente/Representante:

DURAN-CORRETJER, S.L.P

ES 2 728 708 T3

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

Sistema y procedimiento de diálogo audiovisual

5 Sector técnico

La presente invención se refiere a un sistema y a un procedimiento de diálogo audiovisual. En particular, pero no exclusivamente, aspectos de la presente invención se refieren a un sistema de diálogo audiovisual y a un procedimiento que incluyen la conversión de la voz en tiempo real, la animación facial y la personalización de la voz.

10 Antecedentes de la invención y técnica anterior

Los sistemas de diálogo hablado son utilizados de forma regular en una amplia gama de sectores técnicos, por ejemplo, en dispositivos de telefonía móvil para procesar el habla del usuario y transmitirla a un dispositivo de recepción de comunicaciones. Los sistemas de diálogo hablado emplean habitualmente una combinación de síntesis del habla y técnicas de reconocimiento de la voz. Las técnicas de síntesis del habla son utilizadas a menudo en producciones de diversión tales como videojuegos y presentaciones animadas, pero asimismo puede ser una herramienta útil en la tecnología de ayuda a personas con una diversidad de incapacidades, por ejemplo, deficiencias visuales y de habla, y dislexia.

Una técnica habitual de síntesis del habla es un sistema de texto-a-habla (TTS), en el que el texto sin formato es analizado, convertido en expresiones fonéticas y a continuación convertido en habla, mediante la generación de ondas de forma. Sin embargo, el sistema TTS puede tener problemas para identificar la pronunciación correcta de ciertos aspectos del texto, por ejemplo, números, abreviaciones y deletreos que tienen pronunciaciones diferentes dependiendo del contexto de la palabra. Como resultado, el TTS puede ser un proceso largo y complejo y, por consiguiente, a menudo es necesario conocer el texto mucho antes de la síntesis del habla. En vista de esto, los sistemas TTS no son adecuados para la conversión de la voz en tiempo real.

El tratamiento de las alucinaciones auditivas ha sido divulgado por un cierto número de medios de publicación. Artículos de noticias publicados por la BBC (<http://www.bbc.co.uk/news/health-22691718>) y el periódico "The Guardian" (<http://www.guardian.co.uk/society/2013/may/29/schizophrenia-mental-health>) comentan ambos la posibilidad de la utilización de técnicas de realidad virtual (VR) para producir un "avatar" que representa la "voz" oída por pacientes esquizofrénicos con el fin de ayudarles a controlar sus alucinaciones. Otro artículo de noticias de la BBC (<http://www.bbc.co.uk/news/science-environment-23144208>) informa de cómo la tecnología utilizada en un ensayo clínico es capaz de ajustar la voz de un avatar virtual facilitado por el profesional terapéutico, para que se parezca a la voz de la alucinación auditiva del paciente.

Otras técnicas anteriores incluyen: "Computer-assisted therapy for medication resistant auditory hallucinations: proof-of-concept study" de Leff y otros, (Terapias asistidas por ordenador para alucinaciones auditivas resistentes a la medicación: estudio de demostración conceptual), British Journal of Psychiatry, vol. 202 nº 6, 21 de febrero de 2013, páginas 428-433, así como "Continuous Probabilistic Transform for Voice Conversion" (Transformación probabilística continua para la conversión de la voz) de Stylianou y otros, Actas de la IEEE sobre el habla y el procesamiento del audio, Vol. 6, nº 2, 1 de marzo de 1998.

45 Resumen de la invención

La presente invención da a conocer un sistema de diálogo audiovisual que permite a un usuario crear un "avatar" que puede ser personalizado para parecer y sonar de una forma determinada. El avatar puede ser creado para que se parezca, por ejemplo, a una persona, animal o criatura mítica, y puede ser generado para tener una voz variable que puede ser femenina o masculina. El sistema utiliza a continuación una conversión de la voz en tiempo real con el objeto de transformar una entrada de audio, por ejemplo, una palabra hablada, en una voz de interés que es seleccionada y personalizada por el usuario. El sistema está dispuesto para animar facialmente el avatar utilizando un algoritmo de sincronización de los labios en tiempo real, de tal modo que el avatar generado y la voz de interés estén sincronizados. Es decir, el avatar parece que dice las palabras de una entrada de audio con la voz de interés. La presente invención puede ser utilizada en múltiples aplicaciones, por ejemplo, para software de juegos y para películas de animación.

Una aplicación particular de la presente invención es el tratamiento de la esquizofrenia, en la que los pacientes que padecen audición de voces pueden crear una representación visual y audible de su alucinación auditiva con la forma del avatar. El paciente puede personalizar la apariencia visual del avatar y personalizar la voz del avatar, de modo que se parezca estrechamente a su alucinación auditiva. Una vez que el paciente ha diseñado su avatar, un profesional terapéutico puede hablar al paciente a través del avatar de modo que el paciente puede enfrentarse a su alucinación y entrar en conversación con ella, permitiendo de este modo conseguir el control de la enfermedad.

A la vista de lo anterior, un aspecto de la presente invención da a conocer un sistema de diálogo audiovisual, que comprende: un dispositivo de entrada de audio; un dispositivo de salida de audio; un dispositivo de salida visual y un

procesador, estando el procesador dispuesto para: recibir una señal de entrada de audio que representa una fuente de voz procedente del dispositivo de entrada de audio; realizar una conversión de la voz, sustancialmente en tiempo real, de la señal de entrada de audio para producir una señal de salida de audio que representa una voz de interés, en la que la señal de salida de audio es proporcionada al dispositivo de salida de audio, y en la que el proceso de conversión de la voz en tiempo real incluye: i) descomponer la señal de entrada de audio en un conjunto de características de filtrado variables con el tiempo y una señal residual de excitación; ii) transformar espectralmente las características de filtrado variables con el tiempo, y/o modificar el tono de la señal residual de excitación; y iii) sintetizar la señal de salida de audio dependiendo de las características transformadas de filtrado variables con el tiempo y/o de la señal residual de excitación del tono modificada; generar un avatar, en el que el avatar es presentado visualmente en el dispositivo visual de salida; y animar facialmente el avatar generado, en el que la animación está sincronizada con la señal de salida de audio; en el que el procesador está dispuesto además para personalizar la conversión de la voz en tiempo real, comprendiendo la personalización: 1) seleccionar una de una pluralidad de voces de interés predefinidas, en que las voces de interés predefinidas están representadas por medio de un conjunto de las respectivas transformaciones lineales que incluyen un conjunto de características de filtrado variables con el tiempo y un factor de escalado del tono; y 2) ajustar las características de filtrado variables con el tiempo y/o el factor de escalado del tono de la voz de interés predefinida seleccionada para proporcionar parámetros personalizados a la voz de interés, en la que las características de filtrado variables con el tiempo y/o el factor de escalado del tono de la voz de interés predefinida seleccionada, son ajustados utilizando una pluralidad de controles deslizantes visualizados en la interfaz de usuario que, cuando son activados por el usuario, establecen una magnitud de cambio mediante la cual se ajustan las características de filtrado variables con el tiempo y/o el factor de escalado del tono.

En una realización, la conversión de la voz sustancialmente en tiempo real incluye la descomposición de la señal de entrada de audio en un conjunto de características de filtrado variables con el tiempo y en una señal residual de excitación, transformando espectralmente las características de filtrado variables con el tiempo, y/o modificando el tono de la señal residual de excitación, y sintetizando la señal de salida de audio dependiendo de las características de filtrado transformadas variables con el tiempo y/o de la señal residual de excitación del tono modificada.

Preferentemente, la descomposición de la señal de entrada de audio es realizada por medio de un proceso de codificación de predicción lineal. Se debe tener en cuenta que las características de filtrado variables con el tiempo se refieren a los coeficientes de predicción del procesamiento de predicción lineal del habla.

En otra realización, las características de filtrado variables con el tiempo son estimadas en secciones cortas con ventanas de la señal de audio de entrada. Preferentemente, las secciones cortas con ventanas de la señal de audio de entrada son de una duración de 20 a 40 ms y están superpuestas de 5 a 15 ms. Adicionalmente, el error de predicción de la codificación lineal de predicción puede ser calculado por separado para las secciones cortas de ventanas, siendo posteriormente superpuestos los errores de predicción y sumados para producir la señal residual de excitación.

En una realización preferente, las características de filtrado variables con el tiempo son transformadas de Fourier en un punto múltiple, por ejemplo, 256 puntos de amplitud de respuesta antes de la transformación espectral de las características de filtrado variables con el tiempo. Preferentemente, la amplitud de respuesta transformada espectralmente es, a continuación, una transformación inversa de Fourier transformada de nuevo en las características de filtrado variables con el tiempo.

En otra realización, la modificación el tono incluye un nuevo muestreo de la señal residual de excitación y el cambio de la frecuencia fundamental de la señal residual de excitación muestreada de nuevo por medio de un factor constante de escalado del tono. Además, el cambio en la duración de la señal residual de excitación producido por el nuevo muestreo puede ser corregido mediante la compresión o la extensión de la señal residual de excitación en el tiempo para restaurar la duración original.

Según una realización, la transformación espectral de las características de filtrado variables con el tiempo es llevada a cabo por medio de un conjunto de transformaciones lineales. Los conjuntos de transformaciones lineales pueden ser generados a continuación entre la señal de entrada de audio y una pluralidad de voces de interés predefinidas. En una realización, se utilizan más de 50 voces de interés predefinidas para generar los conjuntos de transformaciones lineales. Preferentemente, se utilizan por lo menos 82 voces de interés, incluyendo al menos 40 locutores masculinos de interés y al menos 42 locutores femeninos de interés. En particular, en una realización, puede ser generado un conjunto de 8 transformaciones lineales entre la señal de entrada de audio y una voz de interés predefinida. En otras realizaciones, se pueden utilizar 16, 24, u otras cantidades.

Según otra realización, se pueden pronunciar una serie de frases tanto por medio de las pluralidades de voces de interés predefinidas como por la señal de entrada de audio y, preferentemente, en una realización, se pronuncian por lo menos 20 frases. La pluralidad de frases pronunciadas por la señal de audio de entrada y la pluralidad de frases pronunciadas por la pluralidad de voces de interés predefinidas están alineadas temporalmente. La pluralidad de frases puede estar alineada temporalmente utilizando los coeficientes cepstrales de frecuencia Mel en combinación con un algoritmo de programación dinámica. En una realización, en el caso de una pluralidad de

secciones de señal dentro de los pares de frases alineadas temporalmente, se calculan los coeficientes de predicción y el espectro de codificación de predicción lineal, y se halla la asignación de la frecuencia óptima utilizando un algoritmo de programación dinámica.

5 Preferentemente, se utilizan las asignaciones de frecuencias para la pluralidad de secciones de señal en el caso de la señal de entrada de audio para hallar el conjunto de transformaciones lineales que minimice el error de transformación de la media cuadrática. En una realización concreta, se puede hallar un conjunto de 8 transformaciones lineales. Otros números de transformaciones pueden ser utilizados en otras realizaciones.

10 La voz de interés predefinida seleccionada puede estar asociada con un conjunto de transformaciones lineales. Preferentemente, los vectores de transformación del conjunto de transformaciones lineales se reducen a un vector de transformación media y a una pluralidad de vectores de cambio ortogonales, en los que se puede utilizar un control deslizante para ajustar la magnitud por medio de la cual se añade un vector de cambio al vector medio, de tal modo que las características de filtrado variables con el tiempo son ajustadas.

15 Según otra realización, el procesador está dispuesto además para personalizar facialmente el avatar generado, en el que la personalización facial del avatar generado puede incluir una disposición visual de distintas caras para su selección. Preferentemente, la disposición visual de las distintas caras incluye, al menos 250 caras distintas. Además, la disposición visual de las distintas caras puede variar en género, edad, etnicidad y tipo de cabello y se pueden disponer una serie de accesorios y de tipos de cabello adicionales para su selección.

Otros aspectos de la presente invención son evidentes a partir de las reivindicaciones adjuntas,

Breve descripción de los dibujos

25 La presente invención será descrita a continuación únicamente a modo de ejemplo y haciendo referencia a los dibujos adjuntos, en los que:

30 la figura 1 muestra un conjunto de avatares escogidos por los pacientes;

la figura 2 muestra la página frontal del selector de caras en línea;

la figura 3 muestra la interfaz de usuario de registro de la voz;

35 la figura 4 muestra la ejecución del sistema en una realización de la presente invención;

la figura 5a es un análisis que muestra esquemáticamente la señal hablada;

40 la figura 5b es una reconstrucción de la señal hablada mostrada de forma esquemática;

la figura 6 es un diagrama de bloques que muestra un aparato según una realización de la presente invención;

la figura 7 es un diagrama de flujo que muestra un procedimiento según una realización de la presente invención;

45 la figura 8 es un diagrama de flujo que muestra un procedimiento de conversión de la voz según una realización de la presente invención;

la figura 9 es un diagrama de flujo que muestra el procedimiento de distorsión espectral según una realización de la presente invención;

50 la figura 10 es un diagrama de flujo que muestra el procedimiento de cambio del tono residual según una realización de la presente invención;

55 la figura 11 es un diagrama de flujo que muestra el procedimiento para generar la función de distorsión espectral según una realización de la presente invención.

Descripción detallada de las realizaciones preferentes

Resumen

60 La presente invención propone un sistema altamente novedoso de tecnología del habla para la ejecución de terapias basadas en CBT para alucinaciones auditivas, denominado "Terapia de avatares". La base del sistema es una serie de componentes para la creación y el funcionamiento de un avatar individualizado que "habla" en la voz que oye el paciente, y virtualmente se parece a la cara que el paciente percibe. En los casos en que el paciente no percibe claramente una cara, a él o a ella se les solicita que escojan una cara con la que se sientan cómodos al hablar con ella. Las características del avatar se escogen mediante consulta previa con cada paciente utilizando una serie de

herramientas informatizadas. En la figura 1, se muestra un conjunto de imágenes de avatares escogidos por los pacientes.

Durante las sesiones de terapia, el profesional terapéutico y el paciente están sentados frente a la pantalla de un ordenador en habitaciones separadas y se comunican mediante una conexión dual de audio. Las expresiones del avatar son pronunciadas por el profesional terapéutico y las respuestas del paciente al avatar son realimentadas (solamente el audio) de modo que el paciente puede interactuar y cuestionar el avatar. La voz del avatar es producida mediante la modificación del habla del profesional terapéutico en tiempo real, de modo que las palabras del profesional terapéutico son pronunciadas por el avatar en la voz simulada en el ordenador del cliente con sincronización labial. En el mismo canal, el profesional terapéutico puede comunicar asimismo instrucciones o consejos y dar ánimos al paciente en su propia voz, tal como sería el caso en una sesión de terapia normal. Los resultados a pequeña escala de un estudio RCT basado en NIHR están divulgados en [9] y comentados además en [10].

Introducción

El fenómeno de las alucinaciones auditivas (“oír voces”) es un problema permanente en el tratamiento de enfermedades mentales graves tales como la esquizofrenia. Aproximadamente un 30% de las personas con este diagnóstico continúan experimentando alucinaciones y delirios a pesar del tratamiento con medicación anti-psicótica [1]. Oír voces no solamente es angustioso para los que lo sufren, sino también tiene un serio impacto en sus cuidadores y en el resto de personas con las que entran en contacto. Las alucinaciones auditivas se manifiestan de un cierto número de modos, incluyendo voces que dicen en voz alta lo que el paciente está pensando, voces que expresan un comentario sobre las acciones del paciente o sobre actos externos imaginados, dos o más personas conversando, refiriéndose a menudo al paciente en tercera persona y órdenes que obligan al paciente a realizar ciertas acciones (a menudo violentas). Las voces persistentes limitan severamente la capacidad del paciente para concentrarse en las tareas, y por ello dificultan los intentos de rehabilitación. Los costes directos del tratamiento en el Reino Unido se estiman en 2 billones de libras anualmente, mientras que los costes indirectos, incluyendo la pérdida de empleo de los pacientes y cuidadores ascienden a otros 2 billones de libras [2].

Aproximadamente en los últimos 15 años, se han llevado a cabo en Gran Bretaña un cierto número de pruebas aleatorias controladas (RCT) para evaluar el valor de la terapia del comportamiento cognitivo (CBT) en el caso de síntomas de psicosis persistente resistente a la medicación [3, 4, 5, 6]. Mientras que las pruebas han mostrado algún efecto en la reducción de las alucinaciones auditivas, han sido criticadas en base a su diseño experimental. Una RCT más reciente del CBT, aunque no afecta a la frecuencia o a la intensidad de las alucinaciones auditivas, tuvo éxito en la reducción de la potencia de la voz dominante tal como es percibida por los pacientes, y en su angustia. [7].

Cuando se les pregunta sobre el peor aspecto de oír voces persecutorias, muchos pacientes explican “el sentimiento de impotencia” a que inducen. Los que son capaces de establecer un diálogo con su voz sienten que la controlan mucho más y en consecuencia su sufrimiento se reduce. Muchas personas que oyen voces visualizan también una cara asociada a la voz. Esta cara puede ser la de alguna persona conocida de ellos, la de una personalidad muy conocida o una figura imaginaria, o quizás representando un ángel, un demonio u otra figura religiosa o mítica. Un medio mediante el cual un paciente podría ser ayudado a tomar el control de su voz es mediante la creación de un avatar virtual que represente a la persona que ellos creen que les habla, y entonces permite que el avatar pase a estar progresivamente bajo el control del paciente. Por lo tanto, la incorporación de la voz dentro del contexto de un avatar es un paso natural a realizar desde un punto de vista clínico.

Las técnicas de realidad virtual (VR) han sido exploradas anteriormente para el modelado de episodios psicóticos implicados en la esquizofrenia. Banks y otros [8] informan de un entorno de VR sofisticado para simular alucinaciones auditivas y visuales y motivar su utilización en la educación médica, pero no han ensayado su sistema en un entorno clínico, ni han intentado individualizar características de las voces.

Consideraciones sobre el sistema de diseño

Generación de las expresiones del avatar

Un parámetro importante en el diseño de un sistema de diálogo con una voz sintética es el procedimiento de generación de expresiones sintéticas, bien sintetizando directamente el habla con las características deseadas, o bien transformando el habla natural para producir una cualidad diferente de la del locutor original. Dado que el contenido de las expresiones del profesional terapéutico para ser pronunciadas por el avatar no pueden ser conocidas anticipadamente, el sistema estándar de texto-a-habla (TTS) no es utilizable en la presente invención. Teclear el texto durante las sesiones de terapia podría introducir frecuentemente largos retrasos inaceptables, interrumpiendo la fluencia del diálogo. Por consiguiente, para permitir una interacción tan próxima al tiempo real como sea posible, se utiliza un sistema basado en la conversión de la voz en vez de la síntesis del habla.

Los parámetros de diseño del estudio experimental requieren que la voz de interés y la cara sean generados en una

sesión inicial de registro con cada paciente, y que no duren más de una hora. Teóricamente, esto sería seguido por la primera de las series de sesiones de terapia, lo que significa que el avatar debe ser obtenido en su forma definitiva durante la sesión de registro. Se reclutaron temas de manera continua durante el estudio y por lo tanto las voces de interés y las voces no estaban disponibles previamente al desarrollo tecnológico.

En aplicaciones convencionales de conversión de la voz en las que el conjunto de locutores de interés es conocido anticipadamente, se aplica un proceso de aprendizaje para producir la asignación, o transformación, desde la fuente a la voz de interés. Esto requiere la disponibilidad para el aprendizaje de datos del habla tanto de la fuente como de los locutores de interés. En el desarrollo de la presente invención el principal problema técnico fue el hecho de que las voces de interés no son conocidas hasta que los pacientes los han registrado y, por supuesto, no se pueden obtener muestras reales en ningún caso. Esto significa que las asignaciones no pueden ser aprendidas individualmente y que, por el contrario, se debe desarrollar un conjunto predeterminado de voces transformadas utilizando la voz del profesional de la terapéutica como fuente. Aunque esto limita el campo de las voces que pueden ser producidas con buena calidad, todavía se puede producir una diversidad de voces diferentes, siempre que se disponga de algunos medios adicionales de sintonización fina de las transformaciones.

Componentes del sistema para el registro y la terapia

El sistema comprende una combinación de procedimientos en línea y fuera de línea, e incluye un sistema de conversión de la voz en tiempo real, unos sistemas de personalización de la voz y la cara (registro), un sistema personalizable de animación facial con sincronización de los labios en tiempo real y un canal de audio de dos vías con conmutación entre las voces del profesional de la terapéutica y del avatar.

Fuera de línea, o en los procedimientos de registro, el paciente escoge la voz y la cara apropiadas utilizando un conjunto de herramientas basadas en la informática, asistido por un operador formado y supervisado por el profesional terapéutico. Se debe tener en cuenta que algunas de estas herramientas son productos disponibles comercialmente o juegos de herramientas de software que han sido personalizados para adecuarse a las necesidades de la presente invención.

Sistema de conversión de la voz en tiempo real

Procesamiento de la predicción lineal del habla

Las señales habladas son producidas por medio de señales de excitación producidas en la garganta que son modificadas dependiendo de la forma del tracto vocal. Con el objeto de analizar y sintetizar el habla, las señales de habla deben ser descompuestas en sus componentes separados y deben ser consideradas independientemente. Un modo de conseguir esto es el procesamiento de predicción lineal del habla [19, 20] en el que la señal de habla es modelada como una suma ponderada de p muestras previas de habla, más una forma de la onda residual de excitación que puede ser denominada asimismo como un error de predicción. La señal del habla puede ser expresada como:

$$x(n) = \varepsilon(n) + \sum_{k=1}^p a_k y(n-k) \quad (1)$$

Cuando $x(n)$ es la señal de habla, $\varepsilon(n)$ es el error de predicción, a_k son los coeficientes de predicción y es la suma ponderada de las p muestras previas del habla.

En consecuencia, en el caso de una señal de habla determinada, se aplica una función de transferencia que descompone la señal de habla en los dos componentes, tal como se muestra por medio de la figura 5a, en la que la función de transferencia está expresada como:

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k} \quad (2)$$

A la inversa, para un error de predicción dado se puede aplicar la inversa de la función de transferencia que permite que la señal de habla sea reconstruida, tal como se muestra por medio de la figura 5b. Es decir, es posible modelar la señal de excitación que conduce a la señal de habla.

La señal de habla es muestreada en secciones cortas de ventanas, y con el fin de descomponer o reconstruir la señal de habla, se deben calcular los coeficientes de predicción a_k para la función de transferencia, de tal modo que minimicen la media cuadrática del error de predicción $\varepsilon(n)$. El error de predicción del modelo de predicción lineal es esencialmente la diferencia entre la señal de habla real y la señal de habla predicha, y el valor de su media cuadrática es indicativo de la energía en la onda de forma de excitación residual. Dos técnicas corrientes de minimización del error de predicción de la media cuadrática y del cálculo de los coeficientes de predicción incluyen el

procedimiento de covariancia y el procedimiento de autocorrelación. Ambas técnicas tienen como resultado una matriz de ecuaciones que pueden ser resueltas por medio de un algoritmo universal. Una vez que estos coeficientes de predicción han sido calculados es posible descomponer la señal de habla, manipular los componentes de la señal de habla y a continuación reconstruir la señal de habla.

Adicionalmente, la respuesta de la frecuencia del modelo de predicción lineal corresponde a la envoltura espectral de la señal de habla y está libre de armónicos de tono. Es la envoltura espectral la que proporciona las características del tracto vocal. La forma de la envoltura espectral depende del número de parámetros que son utilizados en el modelo de predicción lineal, esto es, el número de p muestras previas utilizadas para modelar la señal de habla. A medida que p aumenta, más armónicos de tono son capturados. Por consiguiente, si se escoge que p sea muy elevado, los picos espectrales capturan estos armónicos del tono. Sin embargo, si se escoge un p que sea demasiado bajo, las zonas clave de resonancia estarán ausentes del espectro. Como norma general, p debe ser escogido como el doble del número de resonancias del tracto vocal (conformadores) en el ancho de banda de la señal. Asimismo, se debe tener en cuenta que el número de parámetros (p) de predicción lineal no tiene efecto apreciable en el error de la media cuadrática del modelo de predicción lineal.

Tecnología

La tecnología de conversión de la voz se basa en líneas generales en el planteamiento descrito por Stylianou [11]. En este planteamiento, tal como se muestra en la figura 8, el habla se descompone en dos elementos: un filtro que varía con el tiempo y una onda de forma de la excitación residual (etapa 81). Las transformaciones espectrales del filtro combinadas con la modificación del tono (etapa 82) del residuo permiten la síntesis (etapa 83) de versiones de una pronunciación en las que las características del locutor parecen haber cambiado.

La descomposición del habla es llevada a cabo por medio de la predicción lineal, tal como se ha descrito anteriormente, siendo estimado el filtro de predicción en secciones cortas de ventanas de la señal de una duración de 30 ms y una superposición de 10 ms. En otras realizaciones, se pueden utilizar otras longitudes de ventana y de superposición. A continuación se calcula el error de predicción del filtro por separado para cada ventana y luego son superpuestos y sumados para crear la forma de onda residual de excitación. Para hacer que los coeficientes de predicción sean más susceptibles a la asignación espectral, los coeficientes de predicción son transformadas de Fourier en un multipunto, por ejemplo, 256 puntos de amplitud de respuesta. Esta transformación es simplemente una comodidad matemática, permitiendo que la respuesta del filtro sea ajustada por medio de una función de distorsión espectral para la transformación de las características de la voz. En el caso de la síntesis, el espectro distorsionado es convertido de nuevo en coeficientes de predicción. Con una distorsión espectral uniforme, la conversión desde y hacia la respuesta de amplitud no introduce ninguna distorsión apreciable de la señal.

La manipulación espectral de la respuesta del filtro es llevada a cabo por medio de un conjunto de transformaciones lineales (etapa 82). En este trabajo, un conjunto de 8 transformaciones es utilizado para cada voz de interés, cuyas formas exactas se hallan durante el procedimiento de formación y de personalización descrito más adelante. Para seleccionar cómo se utilizan las transformaciones para un alargamiento dado de la señal de habla, la envoltura de la señal espectral para el locutor fuente se modela utilizando un modelo gaussiano de mezclado (Gaussian Mixture Model, (GMM)) de 8 mezclas. El comportamiento puede ser mejorado con 64 mezclas [11, 12] pero como el sistema de conversión de la voz de la presente invención está diseñado para un funcionamiento con flexibilidad y en tiempo real, sin ningún concepto de "precisión" que recree una voz específica, 8 mezclas constituyen el mejor compromiso entre la calidad requerida y una carga de cálculo mínima. Cada mezcla es asociada a continuación con una transformación lineal y la transformación final aplicada al filtro se halla a partir de la suma de transformaciones ponderadas según la probabilidad de la mezcla para la correspondiente envoltura de la fuente de la señal de habla. Ver [11] para una descripción matemática.

En el caso de un escalado del tono, la señal residual del locutor fuente, es muestreada de nuevo en primer lugar para cambiar la frecuencia fundamental por medio de un factor constante. Para corregir el cambio en la duración producido por el nuevo muestreo, el residuo es comprimido o alargado a continuación en el tiempo para restaurar la duración original utilizando una Waveform-Similarity Overlap-Add (WSOLA) (Adición-superposición por similitud de formas de onda) (ver [13]).

Las figuras 9 a 11 muestran, a modo de ejemplo, procedimientos para la distorsión espectral, el cambio residual de tono y la generación de la función de distorsión espectral que deben ser utilizados en la conversión de la voz. La conversión de la voz empieza con el muestreo de la señal fuente del habla en secciones de la señal de 30 ms, (9-11a), superpuestas en 10 ms. Estas secciones de la señal son dispuestas a continuación en ventanas utilizando la función de ventana de Hamming (9-11b) con el fin de determinar todos los valores fuera de $0 \leq n < N$, a cero. A partir de esta señal de ventana, se puede realizar el procedimiento de distorsión espectral mostrado en la figura 9, y el procedimiento de cambio de tono mostrado en la figura 10. Los coeficientes del filtro de predicción de la señal de la ventana se calculan por medio del procedimiento de autocorrelación para hallar en primer lugar un conjunto de coeficientes de autocorrelación (9-10c), y del procedimiento recurrente de Levinson-Durbin para encontrar a continuación los coeficientes de predicción (9-10d). Una vez han sido hallados, los coeficientes de predicción son transformados mediante Fourier para obtener la amplitud de respuesta (9e), y dicha amplitud de respuesta puede

ser distorsionada a continuación utilizando una función espectral de distorsión (9f) calculada para dicha señal de ventana.

Para obtener la función espectral de distorsión, tal como se muestra mediante la figura 11, se calculan los coeficientes cepstrales escalados Mel (MFCC) (11c) para la señal de ventana. Adicionalmente, el espectro envuelto para la señal de ventana es modelado mediante un modelo gaussiano de mezclado (Gaussian Mixture Model, (GMM)) de 8 mezclas y se hallan las probabilidades de que los datos MFCC pudieran haber sido generados por medio de cada mezcla de los GMM (11d). Cada mezcla está asociada con una transformación lineal (correspondiente a las voces de interés) y la suma de estas transformaciones lineales ponderadas mediante las probabilidades anteriores proporciona la función espectral de distorsión a aplicar a la respuesta de la amplitud (11e).

Una vez que la función de distorsión espectral ha sido aplicada a la respuesta de la amplitud, la respuesta de la amplitud resultante distorsionada es una transformada inversa de Fourier para obtener el correspondiente conjunto de coeficientes de autocorrelación (9g). A continuación, utilizando el procedimiento recurrente de Levinson-Durbin, los coeficientes de correlación son utilizados para obtener un conjunto de coeficientes de predicción distorsionados (9h).

La segunda etapa de conversión de la voz es la modificación del tono de la señal de habla. Los coeficientes de predicción de cada señal de ventana, tal como se ha descrito anteriormente, son utilizados para calcular el error de predicción de las señales de ventana, y a continuación los errores de predicción son superpuestos y sumados para obtener la señal residual de la señal de habla (10e). Esta señal residual es muestreada de nuevo mediante un factor de 4, (10f), de tal modo que aumenta la velocidad de muestreo y la señal muestreada, es muestreada otra vez para obtener el cambio de tono de la señal residual (10g). La señal residual con el tono cambiado es superpuesta y añadida a continuación a un acumulador histórico de señales de ventana de 3 x 30 ms, (10h), superpuestas en 10 ms. Con el objeto de restaurar la señal cambiada de tono a la duración original de la señal residual, se utiliza la adición-superposición por similitud de formas de onda para modificar la escala de tiempo sin afectar al tono, (10i). Finalmente, el residuo cambiado de tono es muestreado, reduciéndolo a continuación mediante un factor de 4, (10j).

Los coeficientes de predicción distorsionados y la señal residual con cambio de tono pueden ser utilizados para sintetizar la señal de salida (9i) que se refiere a la voz de interés.

Formación

La entrada al proceso de personalización de la voz la constituye un gran número de transformaciones de la voz formada, generadas entre el único locutor fuente (el profesional terapéutico) y un cierto número de locutores formados. En el caso de la presente invención, se tomaron 55 locutores del cuerpo de acentos del inglés británico [14], y se tomaron 27 locutores del cuerpo UCL de locutores de variabilidad [15]. En total, fueron utilizados 40 locutores masculinos y 42 locutores femeninos de interés y en cada caso, se obtuvo una selección de 20 frases pronunciadas por cada locutor y un conjunto coincidente de 20 frases pronunciadas por el locutor fuente.

Los materiales hablados por el locutor fuente fueron utilizados para construir una mezcla de 8 GMM en base a las características MFCC de la señal [16]. Esta GMM fue utilizada a continuación para formar la totalidad de los 82 conjuntos de transformaciones lineales utilizando el procedimiento de Stylianou. Esto, en primer lugar, implicaba, en el caso del locutor fuente y de cada locutor de interés, una alineación temporal entre cada par de frases coincidentes. Los vectores MFCC para esta alineación fueron utilizados en combinación con un algoritmo de programación dinámica. A continuación, para cada par coincidente de secciones de señal, se calcularon los coeficientes de predicción y el espectro LPC y se halló la frecuencia óptima de asignación para dicho par de estructuras utilizando un algoritmo de programación dinámica. Obtenidas todas las asignaciones para todas las estructuras emparejadas en todas las frases, junto con el conjunto de probabilidades de mezcla GMM para cada estructura fuente, es posible hallar el conjunto de la media de 8 transformaciones que minimizan la media cuadrática del error de transformación. Adicionalmente, se midió la frecuencia media fundamental de cada locutor y del locutor fuente. De este modo, para cada locutor de interés, el procedimiento de formación genera un conjunto de 8 transformaciones lineales de frecuencia y un factor de escalado del tono, que tomados conjuntamente constituyen el mejor intento de obtener la morfología de las características de la voz del locutor fuente en el locutor de interés (etapa 83).

Procedimientos de registro

Selección y personalización de las caras

El procedimiento de selección de caras se basa en el software FaceGen© Modeller (modelador para la generación de caras) desarrollado por la firma Singular Inversions, tal como se muestra en la figura 2. Este programa permite la creación de una gama virtualmente infinita de caras en tres dimensiones y guardarlas en diversos formatos de archivo. Como punto inicial de la selección se creó una disposición virtual o "paleta" de unas 250 caras distintas abarcando un amplio margen de características. Las caras se escogieron para cubrir una amplia variedad de tipos de cara que varían en género, edad y grupo étnico así como de diversos tipos corrientes de cabello.

Está disponible un limitado número de tipos de cabello y de accesorios. Cada uno de estos archivos fue guardado tanto como un archivo de modelos FaceGen como un archivo de imágenes JPEG. Los archivos de imágenes fueron compilados en páginas web que se conectaron entre sí para formar un sitio web para seleccionar la paleta inicial de caras.

Cada una de las imágenes del sitio de selección de imágenes tiene asociado un archivo de modelos FaceGen, almacenado en una estructura de directorio en paralelo. Una vez que el cliente ha escogido una cara inicial adecuada, el correspondiente archivo de modelos de caras es localizado fácilmente para una futura mejora en el FaceGen de acuerdo con los deseos del cliente. La imagen de la cara resultante es exportada como un conjunto de objetivos de animación en el formato de gráficos .OBJ 3D. Estos consisten esencialmente en un conjunto de archivos de imágenes y texturas para el conjunto preseleccionado de "visemas" que son utilizados en el sistema de animación facial descrito más adelante.

Registro de la voz

En el estudio piloto, la personalización de la voz fue realizada en dos etapas. No obstante, el proceso de la presente invención está integrado en una única aplicación. En primer lugar, una muestra de unas transformaciones de 20 voces para el género apropiado es escogida del conjunto de formación y se le pide al cliente que escoja cuál de estas voces tiene las características de locución más próximas a la voz requerida. En la pantalla de selección cada burbuja de habla representa una voz diferente y cuando es pinchada interpreta un archivo audio con sincronización de los labios en la voz que representa, tal como se muestra en la figura 3.

En una segunda etapa, el cliente manipula el conjunto de la transformación escogida utilizando un reducido conjunto de controles deslizantes que alteran el factor de escalado del tono y los coeficientes de transformación que pueden ser integrados en la pantalla de la figura 3. Para permitir la manipulación de las transformaciones utilizando un reducido número de controles deslizantes, se realiza un análisis de los componentes principales en las 82 transformaciones estimadas durante la formación. Esto reduce el vector de transformación escogido a un vector de transformación medio más un cierto número de vectores ortogonales de cambio. Cada control deslizante manipula la cantidad en la que cada vector de cambio es añadido al vector medio, y un control deslizante distinto manipula el factor de escalado del tono. La interfaz de usuario está diseñada para permitir que el cliente experimente con múltiples variantes de la voz escogida y conserve copias de alternativas prometedoras, de tal manera que al final del proceso se pueda guardar la mejor transformación posible.

Temas de rendimiento

Durante la fase de desarrollo de la presente invención, se ensayaron los procedimientos de selección de voz y de cara con tres oyentes de la voz que no tomaron parte en el estudio posterior. Estos oyentes informaron de una precisión de similitud de aproximadamente el 60 - 70% con respecto a la cara objetivo. Una cifra similar se obtuvo con los pacientes que tomaron parte en el estudio. Sin embargo, en unos pocos casos, el efecto fue más espectacular. Un pequeño número de pacientes, cuando se les mostró el avatar por primera vez, encontraron que el realismo era tan grande que fueron incapaces de verlo durante mucho tiempo.

Dadas las limitaciones técnicas y el requisito de que todas las voces puedan ser derivadas de una fuente única (un locutor masculino de unos 70 años y con voz de bajo), resultó difícil en algunas ocasiones hacer coincidir la voz exactamente. De una manera no sorprendente, la mayor dificultad surgió al hacer encajar voces femeninas. En general, los pacientes informaron de que la precisión de la coincidencia estaba dentro de un margen de 60 - 90%. La duración del proceso depende de factores tales como la capacidad de comunicación del cliente, el nivel de detalle deseado y el grado de personalización requerido. Habitualmente se tarda unos 20 minutos, raramente menos de 15 o más de 30 minutos.

Sistema de ejecución

La ejecución en tiempo real de la voz convertida se realiza utilizando un software personalizado que funciona en dos ordenadores personales conectados en red: un proceso "servidor" que funciona en el lado médico y un proceso "cliente" que funciona en el lado del paciente. En la figura 4 se muestra una ilustración esquemática de la disposición experimental.

El proceso servidor captura el audio del micrófono del profesional terapéutico y realiza la totalidad del procesamiento de la señal y de la conversión del habla. El habla emitida es transferida a continuación mediante la red al ordenador del cliente. El proceso del cliente interpreta de nuevo el habla convertida, y captura el habla del paciente, cargándola en el ordenador del profesional terapéutico. El profesional terapéutico utiliza dos botones de "pulsar para hablar" que controlan si se emite al cliente su voz natural o la voz convertida. El profesional terapéutico escucha con los auriculares, de modo que las respuestas del cliente no son realimentadas a través del micrófono del profesional terapéutico. De manera similar, el micrófono del cliente queda silenciado mientras se está enviando el audio desde el servidor, y también para evitar la realimentación a través del sistema. Esto tiene un efecto limitado pero aceptable

en el diálogo.

Es decisivo crear la separación entre el profesional terapéutico y el avatar para asegurar que solamente las expresiones del avatar pasan a través de la cabeza parlante. Dado que ambas señales son conducidas por el mismo canal, esto se consigue por medio de la captura del habla en estéreo en el lado del servidor (profesional terapéutico) y pasando el audio sin procesar y el convertido a través de los canales izquierdo y derecho respectivamente. Solamente el canal derecho pasa al dispositivo de reconocimiento y al motor de sincronización de los labios en la animación del cliente, con lo que se consigue el efecto deseado.

Finalmente, el lado del cliente está equipado con un pulsador contra el pánico que desconecta el avatar inmediatamente cuando es oprimido, presenta una imagen de un paisaje y hace sonar música tranquilizante. En el estudio, pocos clientes encontraron necesario utilizar este dispositivo pero su sola presencia demostró ser de ayuda para aliviar la ansiedad de algunos clientes.

Comentarios adicionales

Se han conseguido los puntos fuertes del sistema de audio, que son la personalización de la voz y la ejecución en tiempo real de una voz personalizada. Cuando la voz de interés es masculina y la transformación no es demasiado grande, la calidad de la transformación es buena y relativamente sin distorsión. El principal punto débil es que la gama de voces de buena calidad disponibles es algo limitada, y cuando las transformaciones requeridas son demasiado grandes, se produce asimismo una pérdida de calidad. Las voces femeninas generadas a partir de la voz de un profesional terapéutico masculino pueden sonar algo distorsionadas y poco convincentes para algunos pacientes.

Además, debido a que todo el procesamiento del audio es realizado utilizando tarjetas de sonido y controladores de sonido estándar, el sistema de procesamiento introduce unos 50 ms de retraso en el diálogo. Adicionalmente, los retrasos variables introducidos por la conexión de la red pueden ocasionar problemas más graves que interfieren en la conducción de la sesión. Por este motivo, el sistema debe ser implementado de nuevo para trabajar con una conexión de audio directa, con el procesamiento del audio reestructurado para utilizar la plataforma de audio "Windows core" con el resultado de unas latencias mucho menores.

Se debe tener en cuenta que el término "conversión de la voz en tiempo real" debe ser interpretado como que significa que la conversión de la voz es llevada a cabo en vivo, de tal modo que se puede intercambiar un diálogo entre dos individuos, por ejemplo, un profesional terapéutico y un paciente para el tratamiento de alucinaciones auditivas. Debe quedar claro al experto que cualquier retraso añadido por el sistema de procesamiento es aceptable siempre que no tenga un efecto apreciable en el flujo de la conversación entre los dos individuos. Tal como se ha mencionado anteriormente, un retraso aceptable puede ser un retraso de unos 50 ms o, a modo de ejemplo, un retraso comprendido dentro de un margen de 20 ms a 250 ms.

Plataforma de animación facial - Implementación

La plataforma de animación está basada en la sincronización de labios en tiempo real SDK de la firma Annosoft LLC [17]. Habitualmente utilizada para la animación en 3D de personajes que hablan en videojuegos o en películas, este conjunto de herramientas permite que personajes en 3D diseñados en otras aplicaciones puedan ser importados en una diversidad de formatos y convertidos con un dispositivo de reconocimiento fonético del habla en tiempo real que produce movimientos de los labios. El sistema puede funcionar tanto con entradas directas de audio como con conversaciones grabadas previamente. La información sobre los personajes y los objetivos de la animación es cargada desde un simple archivo de configuración de texto conectado a un conjunto de archivos gráficos, y puede ser compilada antes en formato binario para una carga y una visualización más rápidas. En la presente invención se puede utilizar un conjunto de 12 "visemas" (ver, por ejemplo, [18]), incluyendo uno para el silencio. Se ha hallado que este es el mejor compromiso entre la precisión de los movimientos de la boca y la uniformidad y la velocidad de la respuesta. El archivo de configuración para cada personaje simplemente especifica qué objetivo de animación del modelo 3D debe ser utilizado para cada "visema" y qué transformación de audio debe ser utilizada. La creación de los archivos de configuración está incorporada en los programas de registro basados en una lista predeterminada de asignación de fonemas y de "visemas".

La estimación de las formas de los labios mejora si el dispositivo de reconocimiento tiene acceso a fragmentos más largos del discurso, de lo que resulta un compromiso entre la precisión de la sincronización de los labios y la latencia de la respuesta. El mejor compromiso obtenido ha sido un valor de la latencia de 150 ms.

Una realización de la presente invención que implementa las técnicas antes descritas será descrita a continuación. De este modo, las técnicas no serán descritas de nuevo y los detalles de las mismas se incorporan como referencia.

La figura 6 muestra un ejemplo de un sistema informático general 1 que puede formar la plataforma para las realizaciones de la invención. El sistema informático 1 comprende una unidad central de procesamiento (CPU) 5, una memoria de trabajo 6 y una interfaz de entrada 2 dispuesta para recibir entradas de control de un usuario a

través de un dispositivo de entrada 22, tal como un teclado, ratón, u otro dispositivo de control, y un hardware 3 de salida dispuesto para proporcionar la emisión de información a un usuario. La interfaz de entrada 2 está dispuesta además para recibir entradas de audio del usuario a través de un dispositivo 21 de entrada de audio, tal como un micrófono u otro dispositivo capaz de capturar sonidos. El equipo de salida 3 incluye una unidad de presentación visual 32 tal como la pantalla de un ordenador, un televisor, u otro medio de presentación visual, y un dispositivo 31 de salida de audio tal como un altavoz u otro dispositivo capaz de producir sonido.

El sistema informático 1 está dotado además de un medio de almacenamiento 4 que puede ser leído por un ordenador, tal como un disco duro (HDD), una memoria USB, una unidad de estado sólido, o cualquier otra forma de almacenamiento de datos de uso general, en la que los datos almacenados 43, 46 y diversos programas de control están dispuestos para controlar el sistema informático 1 para que funcione de acuerdo con las realizaciones de la presente invención. Por ejemplo, el programa 40 de control está dispuesto y preparado para proporcionar un control global del sistema para llevar a cabo las realizaciones de la presente invención. Este programa 40 de control, puede, por ejemplo, recibir entradas del usuario y lanzar otros programas para realizar tareas específicas. Los demás programas bajo el control del programa 40 de control incluyen un programa 42 de generación de un avatar y un programa 44 de personalización facial 44 que están preparados para generar un avatar de acuerdo con las especificaciones de las entradas del usuario, tal como se ha descrito anteriormente en la sección de selección y personalización de caras. Asimismo está dispuesta una disposición visual de distintas caras 46, de las cuales el usuario puede escoger una y personalizarla. Está dispuesto un conjunto de voces de interés 43 predefinidas de las cuales el usuario puede escoger y a continuación seleccionar una, utilizando un programa 41 de personalización de la voz tal como ha sido descrito en la sección de registro de voces. Un programa 45 de conversión de voces en tiempo real según la técnica de conversión de la voz descrita anteriormente está asimismo dispuesto, de modo que las entradas de audio recibidas a través del dispositivo 21 de entrada de audio son transformadas en la voz de interés, tal como ha sido personalizado por el usuario. Finalmente, está dispuesto y preparado un programa 47 de animación facial para animar los movimientos faciales del avatar generado y sincronizar los movimientos de los labios del avatar de acuerdo con la salida del programa 45 de conversión de la voz en tiempo real. El procedimiento de animación facial ha sido descrito anteriormente.

Se debe comprender que, por supuesto, otros varios componentes y sistemas son conocidos de un experto en la técnica para permitir que el sistema informático 1 funcione.

Una aplicación de las realizaciones de la presente invención, tal como se ha descrito en detalle previamente, es el tratamiento de las alucinaciones auditivas. El sistema informático, tal como se muestra en la figura 6, es capaz de llevar a cabo el sistema de diálogo audiovisual descrito en la presente invención e ilustrado mediante las etapas de la figura 7. Un paciente afectado de alucinaciones auditivas puede proporcionar entradas de control a través del dispositivo de entrada 22 para instruir al sistema informático 1 para que genere un avatar (etapa 71). El sistema informático 1 puede comprender un medio de almacenamiento 4 que puede ser leído, incluyendo un programa 40 de control dispuesto para proporcionar el control global del sistema. El programa 40 de control recibe las entradas del paciente y lanza un programa 42 de generación de un avatar que está preparado para generar un avatar (etapa 71). A continuación, el programa 40 de control puede lanzar un programa 44 de personalización de caras que permite que el paciente personalice el avatar generado utilizando una disposición visual de distintas caras 46 almacenadas en el medio 4 que puede ser leído por el ordenador. Una vez que el paciente ha completado la personalización facial, el avatar completado es presentado en una unidad 32 de presentación visual (etapa 71). El programa 40 de control puede lanzar asimismo un programa 41 de personalización de la voz que está dispuesto para permitir que el paciente escoja la voz de interés del avatar generado, de modo que se parezca lo más posible a su alucinación auditiva (etapa 70). El programa 41 de personalización de la voz utiliza un conjunto 43 de voces de interés predefinidas que incluye una gama de voces masculinas y femeninas de las cuales el paciente puede seleccionar una. Entonces, el paciente es capaz de personalizar los parámetros de la voz de interés seleccionada, tal como se muestra en la figura 3, utilizando una selección de controles deslizantes que aparecen en el dispositivo de entrada 22 hasta conseguir un cercano parecido con la alucinación auditiva.

Una vez completados los procesos de personalización facial y de la voz, el profesional terapéutico y el paciente pueden entrar en conversación a través del avatar generado. El profesional terapéutico proporciona una entrada de audio (etapa 72) a través de un dispositivo 21 de entrada de audio y el programa 40 de control lanza un programa 45 de conversión de la voz en tiempo real para transformar la entrada de audio, es decir, la voz del profesional terapéutico en la voz de interés, utilizando los parámetros personalizados escogidos por el paciente (etapa 73), en que el programa de conversión de la voz emplea la técnica de conversión de la voz antes descrita. La salida del programa 45 de conversión de la voz en tiempo real es proporcionada a continuación al paciente a través de un dispositivo 31 de salida de audio. De forma simultánea, el programa 40 de control lanza un programa 47 de animación facial que está dispuesto para animar los movimientos faciales del avatar generado, concretamente, sincronizando los labios del avatar en tiempo real con la salida del programa 45 de conversión de la voz en tiempo real (etapa 74). Como resultado, cuando el profesional terapéutico habla en el dispositivo 21 de entrada de audio, el avatar parece decir las palabras pronunciadas por el profesional terapéutico en la voz de la alucinación auditiva del paciente.

Conclusiones

- 5 Parece que no existe ningún otro estudio para el tratamiento de las alucinaciones auditivas utilizando avatares generados mediante ordenador. La tecnología utilizada en el estudio piloto original fue desarrollada en un breve periodo de tiempo y actualmente se está desarrollando más como parte de un ensayo clínico de 3 años que implica un grupo mayor de pacientes (140), incluyendo su extensión a otros idiomas. Todos los componentes básicos son bien conocidos en el sector del habla, pero su combinación y adaptación para adecuarse a la aplicación concreta, es novedosa. En la práctica es relativamente sencillo y de coste reducido para ser implementado en un entorno clínico, requiriendo solamente dos ordenadores de sobremesa estándar conectados mediante un cable de audio. Los
- 10 pacientes captan rápidamente la idea del sistema y son motivados por el mismo. Desde el punto de vista psiquiátrico, la introducción de un avatar permite a los investigadores estudiar la relación entre los pacientes y sus voces, de primera mano y por primera vez.
- 15 Los resultados clínicos han demostrado hasta el momento ser prometedores, con resultados espectaculares en un reducido número de pacientes [9, 10]. Las limitaciones presupuestarias permitieron inicialmente solamente un único profesional terapéutico (masculino), sin embargo, el estudio actual incluye tanto hombres como mujeres, ampliando de este modo el conjunto de voces disponibles. El propósito principal del nuevo estudio es repetir la terapia del avatar en un nuevo entorno con diferentes profesionales terapéuticos para establecer si se pueden obtener beneficios terapéuticos independientemente por cualquier profesional terapéutico formado. Otras cuestiones de la
- 20 investigación incluyen el ensayo del efecto de la exactitud en la coincidencia de la voz y de la cara, el nivel de realismo requerido para obtener el mejor efecto terapéutico, si la creación del avatar es por sí misma terapéutica, y qué subgrupos de pacientes es más probable que se beneficien de la terapia.

REIVINDICACIONES

1. Sistema audiovisual de diálogo (1) que comprende:

5 un dispositivo de entrada de audio (21);
un dispositivo de salida de audio (31);
un dispositivo de salida visual (32); y
un procesador (5), estando el procesador dispuesto para:

10 recibir (72) una señal de entrada de audio que representa una voz fuente desde el dispositivo de entrada de audio;
realizar (73) una conversión sustancialmente en tiempo real de la conversión de la voz en la señal de entrada de
audio para producir una señal de salida de audio que representa una voz de interés, en la que la señal de salida de
audio es proporcionada al dispositivo de salida de audio, y en la que el proceso de conversión de la voz en tiempo
real incluye:

15 i) descomponer la señal de entrada de audio en un conjunto de características de filtrado variables con el
tiempo y una señal residual de excitación;
ii) transformar espectralmente las características de filtrado variables con el tiempo, y/o modificar el tono de
la señal residual de excitación; y
20 iii) sintetizar la señal de salida de audio dependiendo de las características de filtrado variables con el
tiempo transformadas y/o de la señal residual de excitación con el tono modificado;

generar (71) un avatar, en el que el avatar es presentado visualmente en el dispositivo de salida visual; y
25 animar facialmente el avatar generado (74), en el que la animación está sincronizada con la señal de salida de
audio;
en el que el procesador está dispuesto además para personalizar la conversión de la voz en tiempo real,
comprendiendo la personalización:

30 1) seleccionar una de una pluralidad de voces de interés predefinidas, en que las voces de interés
predefinidas están representadas por un conjunto de las respectivas transformaciones lineales que incluyen
un conjunto de características de filtrado variables con el tiempo y un factor de escalado del tono; y
2) ajustar (82) las características de filtrado variables con el tiempo y/o el factor de escalado del tono de la
voz de interés predefinida, seleccionada para proporcionar los parámetros personalizados a la voz de
interés,

35 en el que las características de filtrado variables con el tiempo y/o el factor de escalado del tono de la voz de interés
predefinida seleccionada, son ajustados utilizando una pluralidad de controles deslizantes (A, B, C, D) dispuestos en
la interfaz del usuario, los cuales cuando son activados por el usuario establecen una magnitud de cambio mediante
la cual se ajustan las características de filtrado variables con el tiempo y/o el factor de escalado del tono.

40 2. Sistema, según la reivindicación 1, en el que las características de filtrado variables con el tiempo son calculadas
en secciones cortas de ventana de la señal de entrada de audio, en el que las secciones cortas de ventana de la
señal de entrada de audio tienen una duración de 20 a 40 ms y están superpuestas de 5 a 15 ms.

45 3. Sistema, según cualquier reivindicación anterior, en el que las características de filtrado variables con el tiempo
son transformadas de Fourier con una amplitud de respuesta de múltiples puntos antes de la transformación
espectral de las características de filtrado variables con el tiempo, y en el que la amplitud de respuesta transformada
espectralmente es una transformada inversa de Fourier de nuevo con las características de filtrado variables con el
tiempo.

50 4. Sistema, según cualquiera de las reivindicaciones anteriores, en el que un conjunto de, por lo menos, 8
transformadas lineales es generado entre la señal de entrada de audio y una voz de interés predefinida, y en el que
más de 50 voces de interés predefinidas son utilizadas para generar los conjuntos de transformaciones lineales.

55 5. Sistema, según cualquiera de las reivindicaciones anteriores, en el que una pluralidad de frases son pronunciadas
tanto por la pluralidad de voces de interés predefinidas como por la señal de entrada de audio, y en el que por lo
menos son pronunciadas 20 frases.

60 6. Sistema, según la reivindicación 5, en el que la pluralidad de frases pronunciadas por la señal de entrada de audio
y la pluralidad de frases pronunciadas por la pluralidad de voces de interés predefinidas están alineadas
temporalmente, en el que la pluralidad de frases están alineadas temporalmente utilizando los coeficientes
cepstrales de frecuencia Mel de la señal de entrada de audio en combinación con un algoritmo de programación
dinámica.

65 7. Sistema, según la reivindicación 6, en el que en el caso de una pluralidad de secciones de señal dentro de la
pluralidad de frases temporalmente alineadas:

- a) se calculan los coeficientes de predicción y el espectro de codificación de la predicción lineal; y
- b) se halla la asignación de la frecuencia óptima utilizando un algoritmo de programación dinámica.

- 5 8. Sistema, según cualquiera de las reivindicaciones anteriores, en el que los vectores de transformación del conjunto de transformaciones lineales son reducidos a un vector de transformación medio y a una pluralidad de vectores de cambio ortogonales.
- 10 9. Sistema, según la reivindicación 8, en el que se utiliza un control deslizante para ajustar la magnitud por medio del cual se añade un vector de cambio al vector de transformación medio, de tal modo que se ajustan las características de filtrado variables con el tiempo.
- 15 10. Sistema, según cualquiera de las reivindicaciones anteriores, en el que el procesador está además dispuesto para personalizar facialmente el avatar generado, en el que la personalización facial del avatar generado incluye proporcionar una disposición visual con caras distintas para la selección.
- 20 11. Sistema, según cualquiera de las reivindicaciones anteriores, en el que el dispositivo de entrada de audio y el dispositivo de salida de audio pueden ser conectados para formar un canal de audio de dos vías.
- 25 12. Procedimiento de diálogo audiovisual que comprende:
 - recibir (72) una señal de entrada de audio que representa una voz fuente, desde un dispositivo de entrada de audio;
 - realizar (73) la conversión de la voz sustancialmente en tiempo real en la señal de entrada de audio para producir una señal de salida de audio que representa una voz de interés, en la que la salida de la señal de audio es proporcionada a un dispositivo de salida de audio, y en el que la conversión de la voz sustancialmente en tiempo real incluye:
 - 30 i) descomponer (81) la señal de entrada de audio en un conjunto de características de filtrado variables en el tiempo y una señal residual de excitación;
 - ii) transformar espectralmente (82) las características de filtrado variables en el tiempo, y/o modificar el tono de la señal residual de excitación; y
 - iii) sintetizar (83) la señal de salida de audio dependiendo de las características de filtrado variables en el tiempo y/o de la señal residual de excitación con el tono modificado;
 - 35 generar un avatar (71), en el que el avatar es mostrado visualmente en un dispositivo de salida visual; y animar facialmente (74) el avatar generado, en el que la animación está sincronizada con la señal de audio de salida;
 - 40 el procedimiento comprende además la personalización de la conversión de la voz en tiempo real, comprendiendo la personalización:
 - 1) seleccionar una de una pluralidad de voces de interés predefinidas, en que las voces de interés predefinidas están representadas por medio de un conjunto de las respectivas transformaciones lineales que incluyen un conjunto de características de filtrado variables en el tiempo y un factor de escalado del tono; y
 - 45 2) ajustar (82) las características de filtrado variables en el tiempo y/o el factor de escalado del tono de la voz de interés predefinida seleccionada para proporcionar parámetros personalizados a la voz de interés,
 - 50 en el que las características de filtrado variables en el tiempo y/o el factor de escalado del tono de la voz de interés predefinida seleccionada son ajustados utilizando una pluralidad de controles deslizantes (A, B, C, D) presentados en una interfaz de usuario que, cuando es activada por el usuario, establece una magnitud de cambio por medio de la cual se ajustan las características de filtrado variables en el tiempo y el factor de escalado del tono de la voz.



FIG. 1

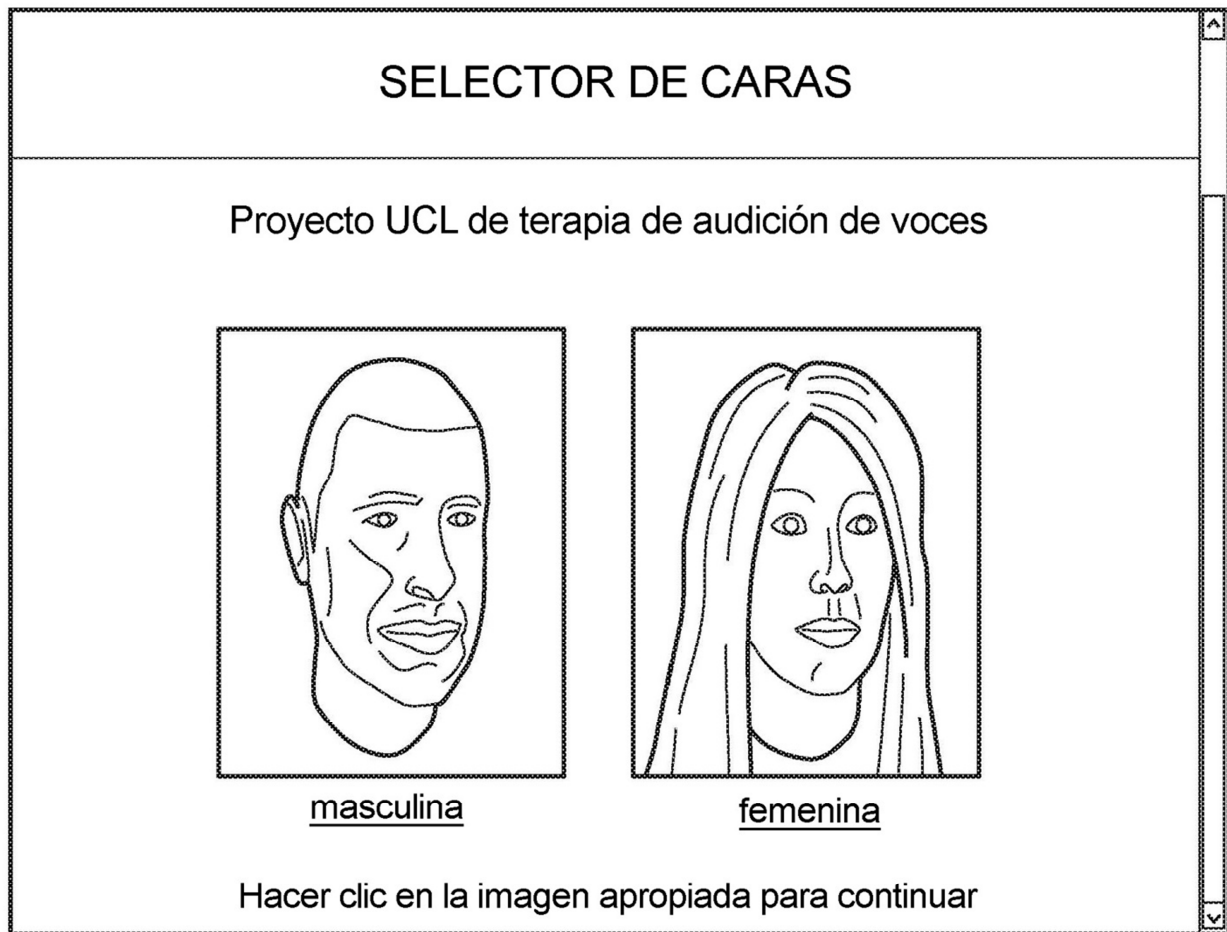


FIG. 2

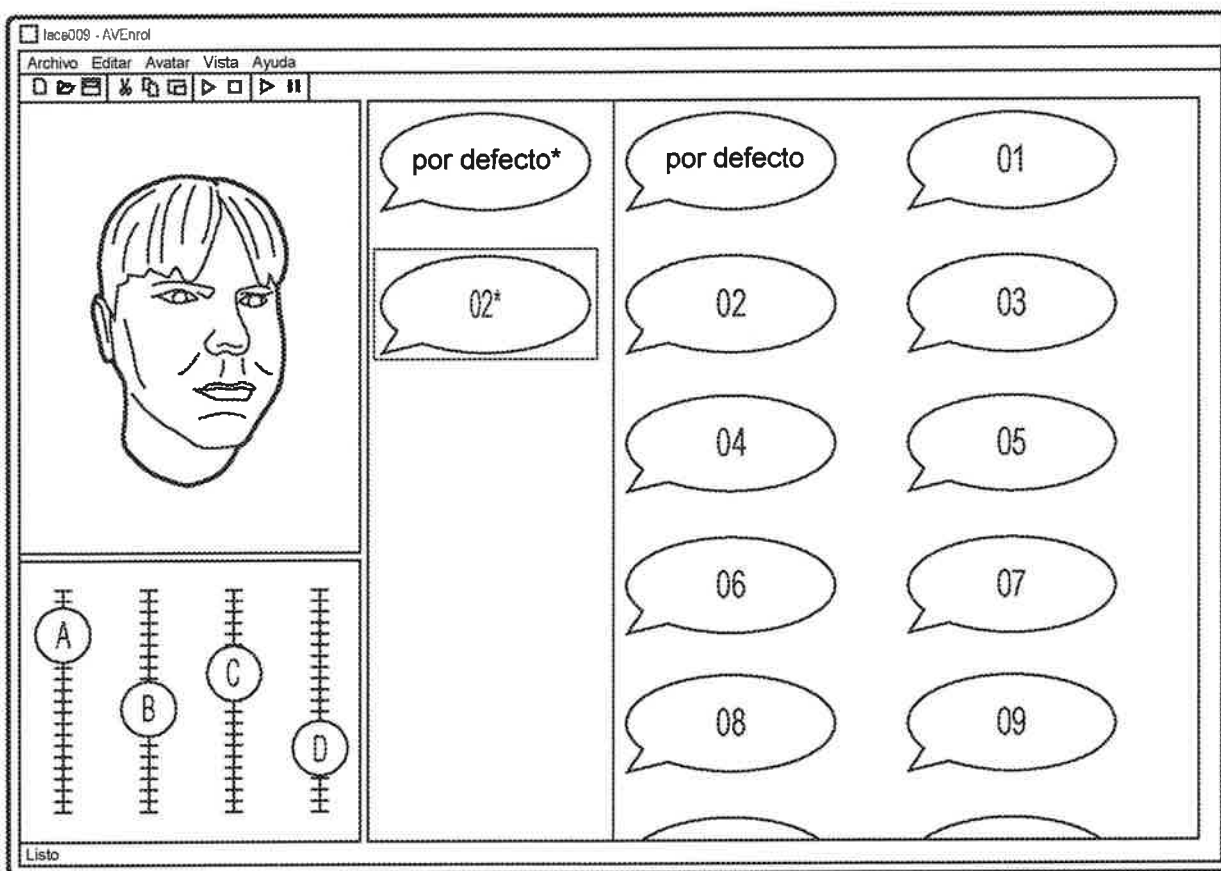


FIG. 3

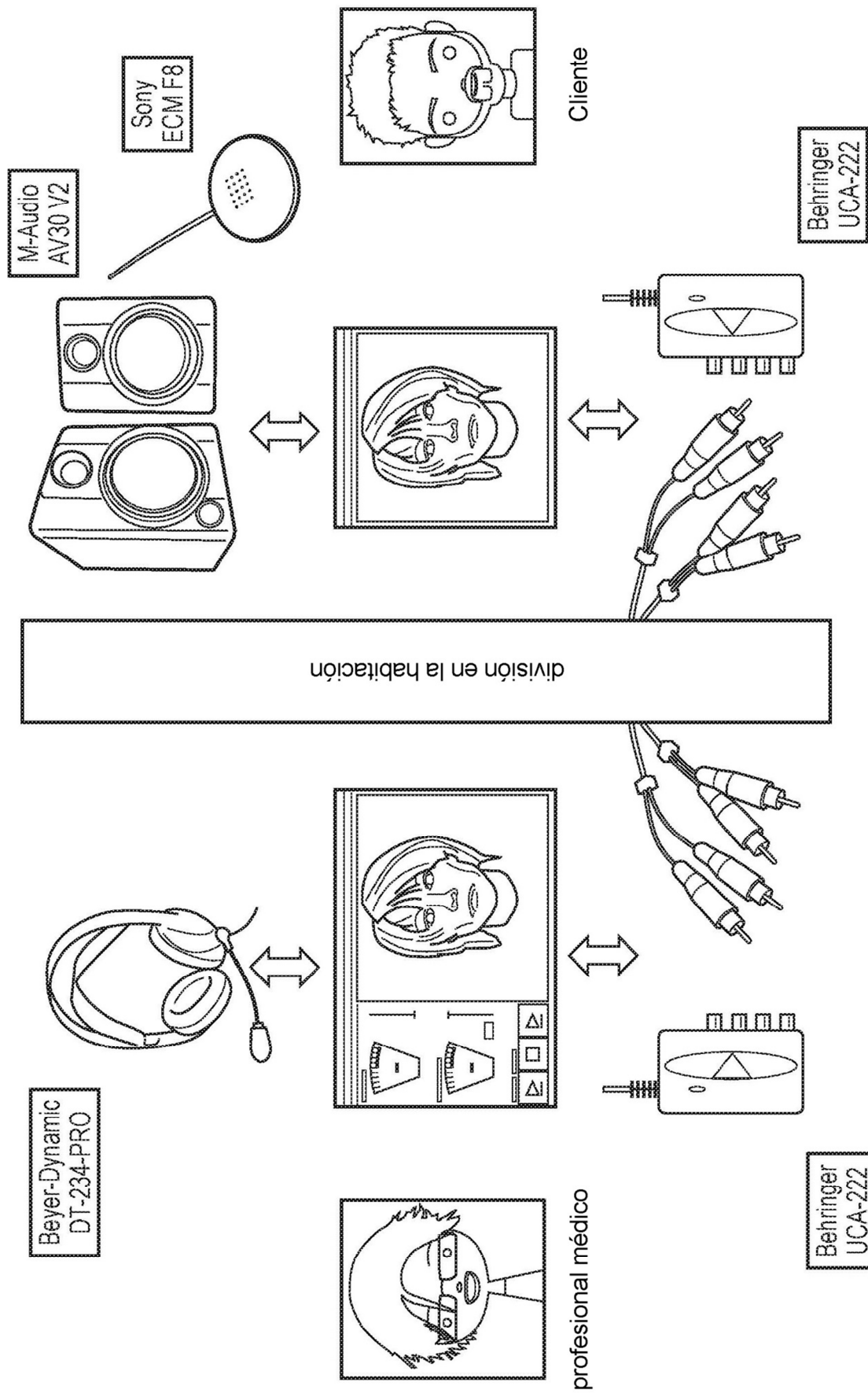


FIG. 4

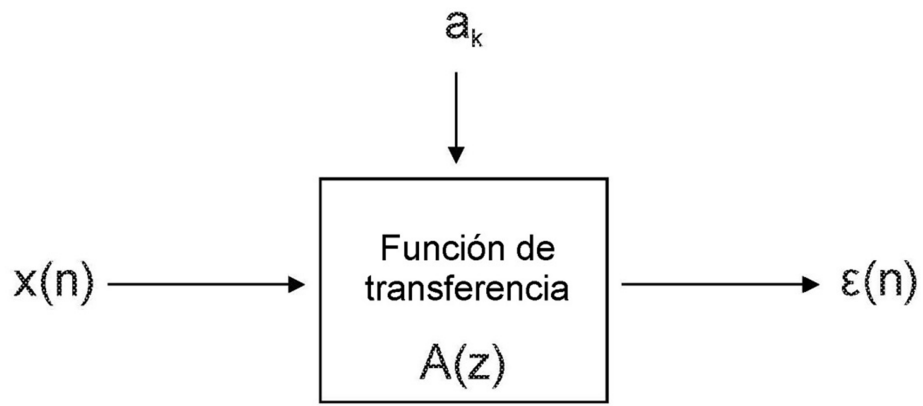


FIG. 5a

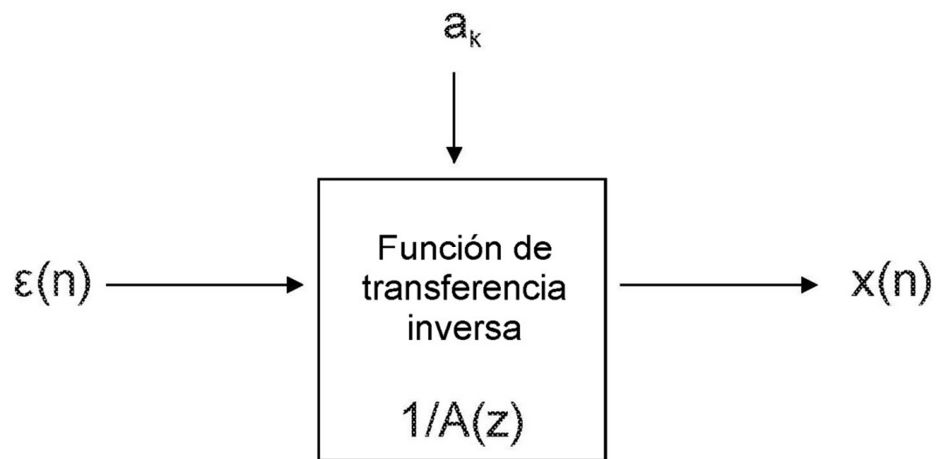


FIG. 5b

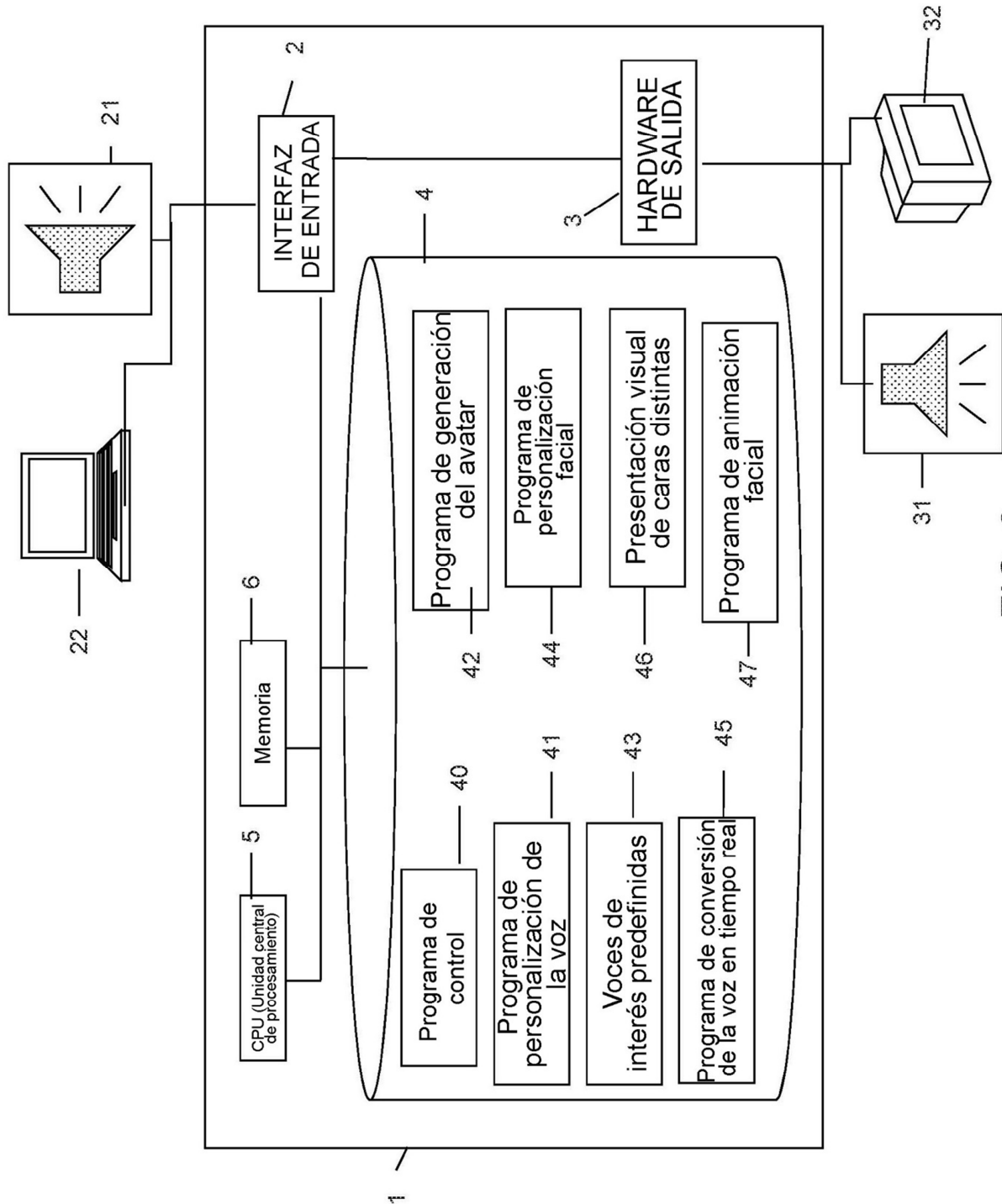


FIG. 6

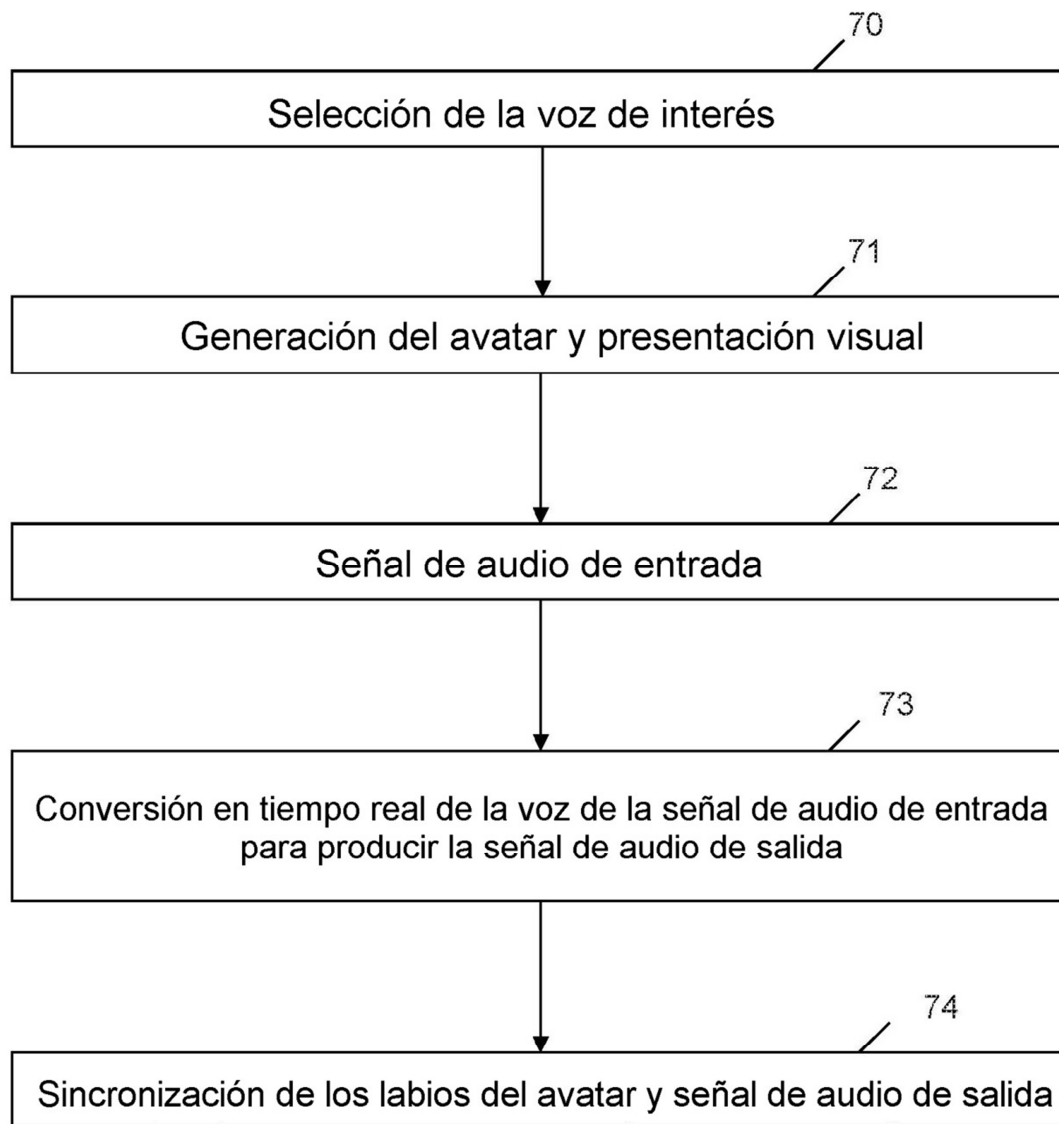


FIG. 7

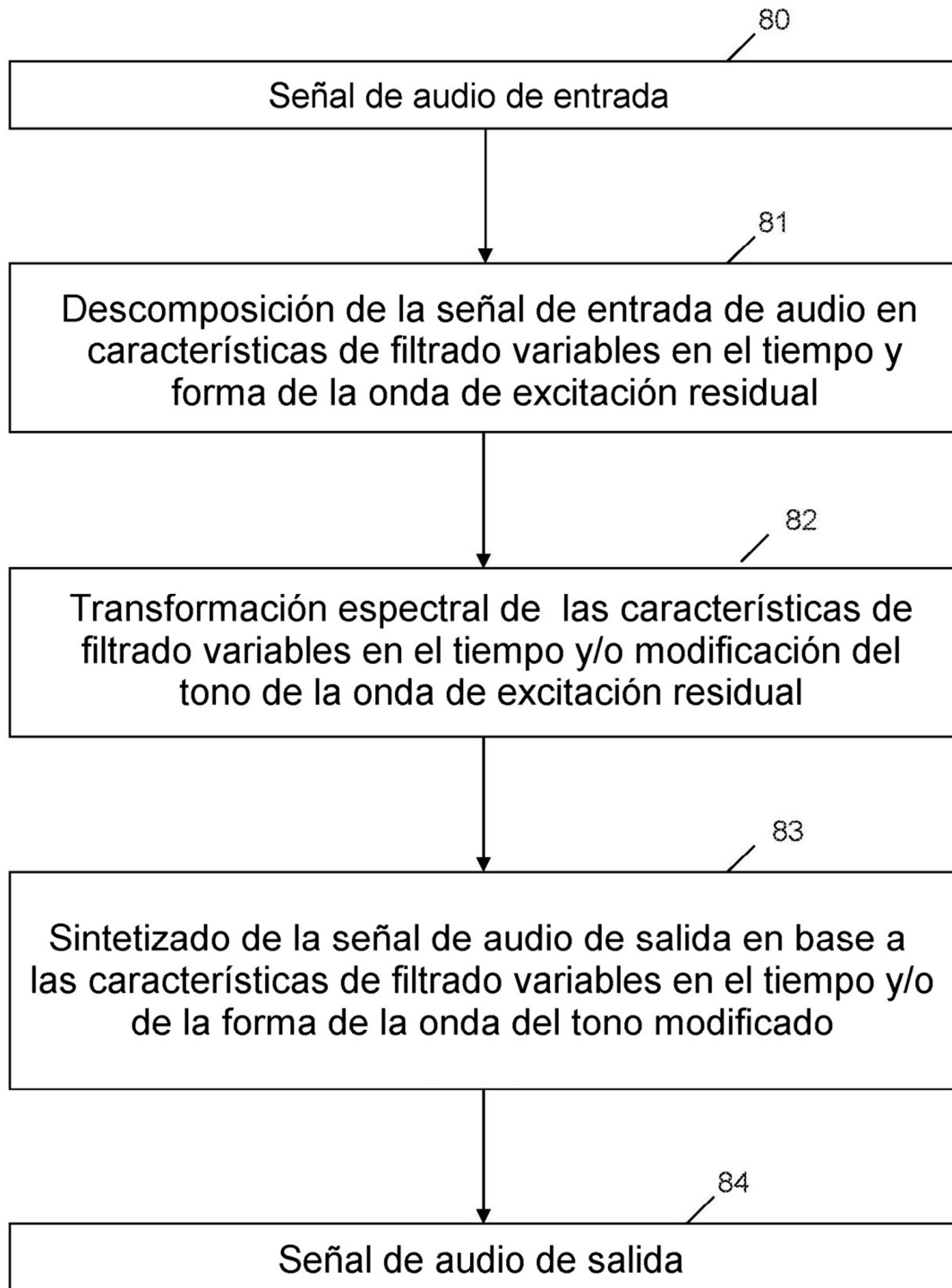


FIG. 8

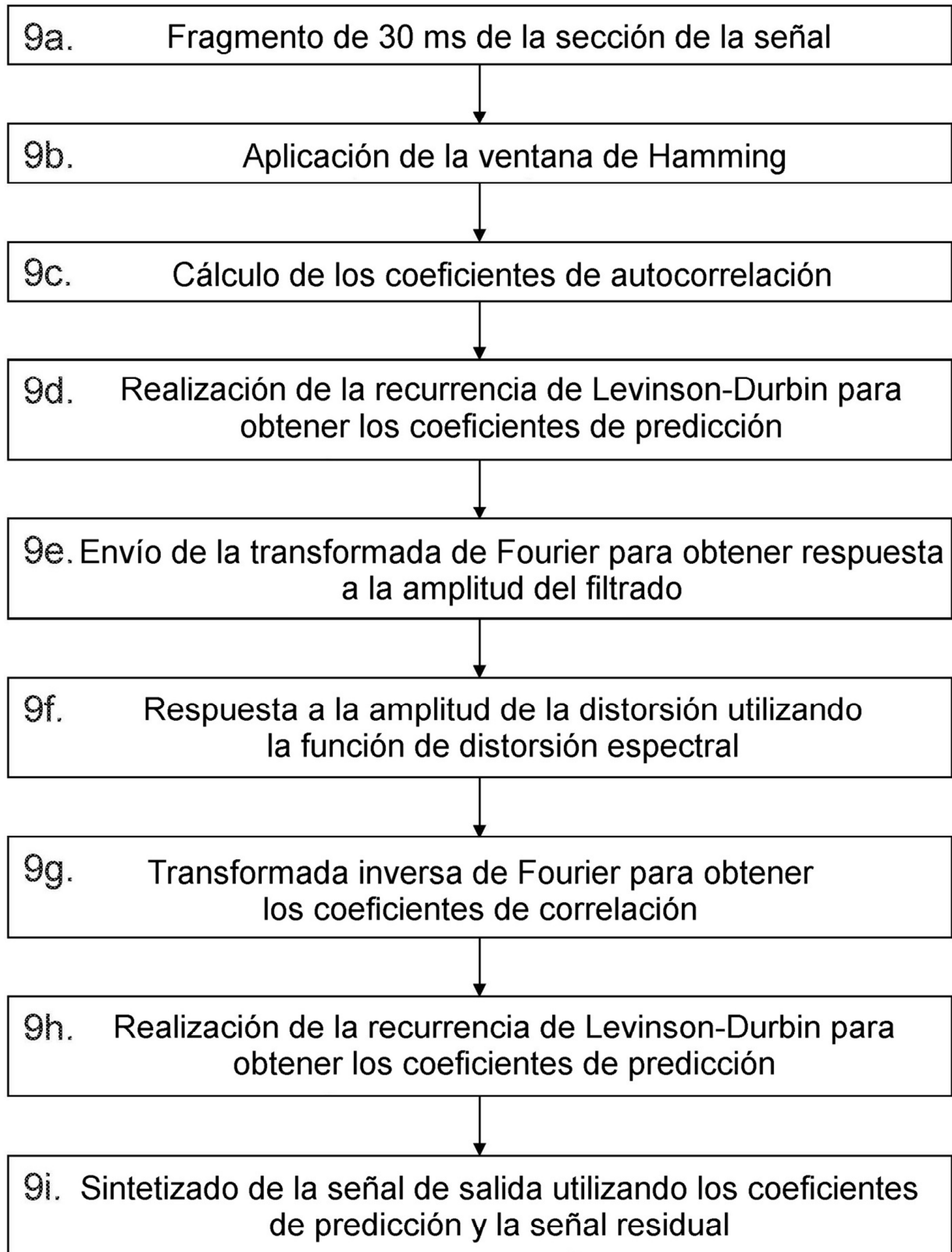


FIG. 9

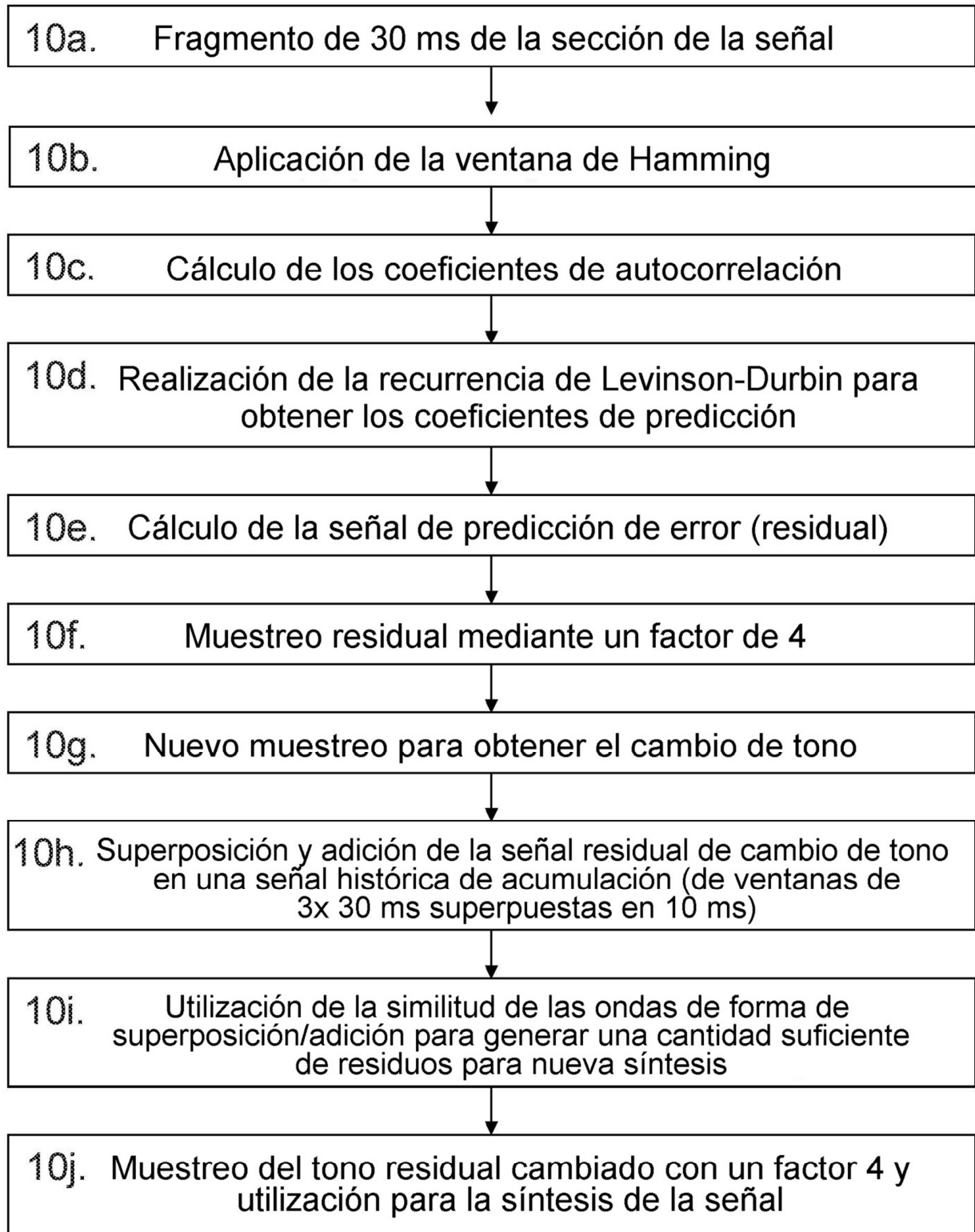


FIG. 10

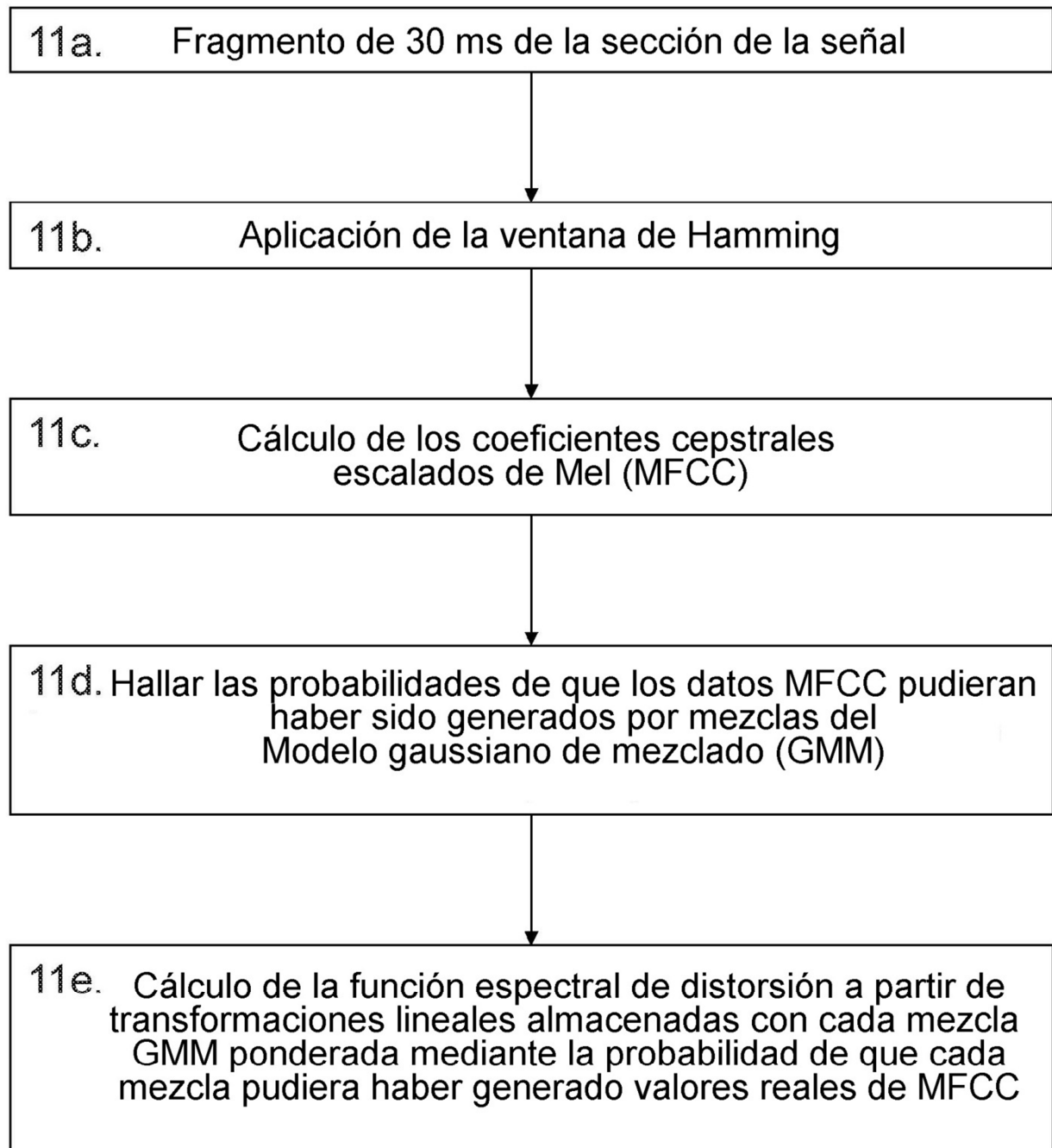


FIG. 11