(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 878 137**

(51) Int. CI.:

G10L 15/06 (2013.01)
G10L 15/08 (2006.01)
G10L 15/16 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **26.06.2018 PCT/CN2018/092899**

(87) Fecha y número de publicación internacional: **03.01.2019 WO19001428**

(96) Fecha de presentación y número de la solicitud europea: **26.06.2018 E 18823086 (6)**

(97) Fecha y número de publicación de la concesión europea: **09.06.2021 EP 3579227**

(54) Título: **Método y dispositivo de activación de voz y dispositivo electrónico**

(30) Prioridad:

29.06.2017 CN 201710514348

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
18.11.2021

(73) Titular/es:

ADVANCED NEW TECHNOLOGIES CO., LTD. (100.0%)
Cayman Corporate Centre, 27 Hospital Road George Town, Grand Cayman KY1-9008, KY

(72) Inventor/es:

WANG, ZHIMING;
ZHOU, JUN y
LI, XIAOLONG

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Método y dispositivo de activación de voz y dispositivo electrónico

Campo técnico

Esta especificación se refiere al campo de las tecnologías de software informático y, en particular, a un método, aparato y dispositivo electrónico de activación de voz.

Técnica antecedente

Con el rápido desarrollo de los dispositivos móviles, las tecnologías relacionadas con la voz también se están volviendo cada vez más comunes. Por ejemplo, el reconocimiento de voz se utiliza en asistentes de conversación cada vez más populares como Siri de Apple, Cortana de Microsoft, y Alexa de Amazon para mejorar la experiencia del usuario y el nivel natural de interacción humano-ordenador.

Una tecnología de interacción de voz importante es la Detección de Palabras Clave (KWS), la cual también puede denominarse en general como activación de voz. Con base en la técnica anterior, existe la necesidad de una solución de activación de voz que no dependa únicamente de datos de voz específicos de palabras clave.

ZHANG et al, Deep Recurrent Convolutional Neural Network: Improving Performance for Speech Recognition, CORR (ARXIV), vol. 1611.07174v2, 27 de Diciembre de 2016, páginas 1-11, XP055524882, se refiere a la arquitectura para el modelado de secuencias y al reconocimiento de voz utilizando redes convolucionales recurrentes profundas con aprendizaje residual profundo.

HORI et al, Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder y RNN-LM, ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 8 de Junio de 2017, XP080768573, se refiere a un patrón de Reconocimiento Automático de Voz (ASR) de extremo a extremo en el cual una red CTC se ubica en la parte superior de un decodificador y se entrena conjuntamente con el decodificador con base en la atención, y donde el proceso de búsqueda de haz combina predicciones CTC, las predicciones del decodificador con base en la atención, y un patrón de lenguaje LSTM entrenado separadamente.

TIAN et al, Frame Stacking and Retaining for Recurrent Neural Network Acoustic Model, ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY, 17 de Mayo de 2017, XP080948417 se refiere a un método de retención de marco aplicado a la decodificación.

WANG et al, Small-Footprint Keyword Spotting Using Deep Neural Network y Connectionist Temporal Classifier, Al Deparment, Ant Financial Group, Hangzhou, China, 11 de Septiembre de 2017 [obtenido el 25 de Marzo de 2019], obtenido de https://arxiv.org/pdf/1709.03665.pdf, se refiere a un sistema de Detección de Palabras Clave (KWS) que utiliza una Red Neuronal Profunda (DNN) y un Clasificador Temporal Conexionista (CTC) en dispositivos móviles de huella reducida con restricción de energía.

ZHANG et al. Wake-up-Word Spotting Using End-to-End Deep Neural Network System, in 23rd International Conference on Pattern Recognition (ICPR), Cancún, México, IEEE 2016, páginas 2878-83, XP033086067, se refiere a un sistema de detección de palabras para activación (WUW) con base en una arquitectura de extremo a extremo y divulga un sistema WUW liviano con base en el proceso de refinamiento del patrón orientado a WUW y el transductor de estado finito ponderado (WFST) con base en la detección de WUW y la estimación de la puntuación de confianza.

Resumen de la invención

Las reivindicaciones independientes exponen la invención reivindicada. Las realizaciones específicas se exponen en las reivindicaciones dependientes.

Al menos una de las soluciones técnicas anteriores adoptadas en las realizaciones de esta especificación puede lograr los siguientes efectos beneficiosos: el patrón de activación de voz puede entrenarse con datos de voz generales accesibles y datos de voz específicos de palabras clave, y luego el patrón de activación de voz entrenado puede ser utilizado para la activación de voz, lo cual es favorable para mejorar la precisión de la activación de voz.

Breve descripción de los dibujos

Con el fin de describir las soluciones técnicas en las realizaciones de esta especificación o en la técnica anterior más claramente, los dibujos adjuntos que se utilizarán en la descripción sobre las realizaciones o la técnica anterior se introducirán brevemente a continuación. Es evidente que los dibujos adjuntos que se describen a continuación son simplemente algunas realizaciones divulgadas en esta especificación. Los expertos en la técnica pueden obtener además otros dibujos adjuntos de acuerdo con estos dibujos adjuntos sin esfuerzos creativos.

La Figura 1 es un diagrama esquemático de una arquitectura general involucrada en una solución de esta especificación en un escenario de aplicación real;

La Figura 2 es un diagrama de flujo esquemático de un método de activación de voz de acuerdo con un ejemplo de esta especificación;

La Figura 3 es un diagrama esquemático de un esquema conceptual del patrón de activación de voz en la Figura 2 de acuerdo con una realización de esta especificación;

La Figura 4 es un diagrama esquemático de extracción de características del módulo de extracción de características de la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación;

La Figura 5 es un diagrama estructural esquemático de una Red Neuronal Profunda (DNN) de la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación;

La Figura 6 es un diagrama estructural esquemático de un Clasificador Temporal Conexionista (CTC) en la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación; y

La Figura 7 es un diagrama estructural esquemático de un aparato de activación de voz correspondiente a la Figura 2 de acuerdo con una realización de esta especificación.

Descripción detallada

En las realizaciones de esta especificación se proporcionan un método, aparato, y dispositivo electrónico de activación de voz.

Con el fin de permitir que los expertos en la técnica comprendan mejor las soluciones técnicas en esta especificación, las soluciones técnicas en las realizaciones de esta especificación se describirán de manera clara y completamente a continuación con referencia a los dibujos adjuntos en las realizaciones de esta especificación. Es evidente que las realizaciones descritas son simplemente algunas, y no todas, las realizaciones de esta solicitud. Todas las demás realizaciones obtenidas por los expertos en la técnica con base en el alcance de las reivindicaciones adjuntas, deberán estar todas incluidas dentro del alcance de protección de esta solicitud.

Para facilitar la comprensión, la idea de las soluciones de esta especificación se explica a continuación. En esta especificación, un patrón de activación de voz que incluye una Red Neuronal Profunda (DNN) y un Clasificador Temporal Conexionista (CTC) se entrena con datos de voz generales. El patrón de activación de voz entrenado se puede utilizar para la activación de voz y soporte de palabras clave definidas por el usuario desencadenadas por la activación de voz. Además, el patrón de activación de voz se puede utilizar en dispositivos de bajo consumo tales como teléfonos móviles y electrodomésticos, debido a que la DNN que se incluye en el patrón de activación de voz puede ser relativamente no tan complicada y, por ejemplo, solo puede tener tres o cuatro capas con doscientos o trescientos nodos en cada capa. El patrón de activación de voz puede denominarse CTC-KWS, y el KWS aquí es la detección de palabras clave que se menciona en el antecedente.

La DNN es un perceptrón de múltiples capas, el cual tiene una capa oculta entre una capa de entrada y una capa de salida y puede simular relaciones complejas no lineales. El CTC es un clasificador configurado para realizar una tarea de etiquetado de etiquetas, y no requiere una alineación forzada entre la entrada y la salida.

La Figura 1 es un diagrama esquemático de una arquitectura general involucrada en una solución de esta especificación en un escenario de aplicación real. Dos partes están involucradas principalmente en la arquitectura general: datos de voz y un patrón de activación de voz. El patrón de activación de voz incluye una DNN y un CTC. La activación de voz se puede implementar introduciendo los datos de voz en el patrón de activación de voz para el procesamiento.

La solución de esta especificación se describe en detalle a continuación con base en la idea anterior y la arquitectura general.

La Figura 2 es un diagrama de flujo esquemático de un método de activación de voz de acuerdo con un ejemplo de esta especificación. A partir de la perspectiva de los programas, el cuerpo ejecutivo del flujo puede ser un programa en un servidor o un terminal, por ejemplo, un programa de entrenamiento de patrón, un programa de reconocimiento de voz, una aplicación de activación de voz, etc. A partir de la perspectiva de los dispositivos, el posible cuerpo ejecutivo del flujo es, pero no se limita a, al menos uno de los siguientes dispositivos que pueden servir como servidores o terminales: un teléfono móvil, una tableta, un dispositivo portátil inteligente, una máquina de automóvil, un ordenador personal, un ordenador de tamaño mediano, un grupo de ordenadores, etc.

El flujo de la Figura 2 puede incluir las siguientes etapas.

En S202, los datos de voz se ingresan en un patrón de activación de voz entrenado con datos de voz generales.

En la realización de esta especificación, el cuerpo ejecutivo u otro cuerpo puede monitorizar la voz para obtener los datos de voz. Cuando se monitoriza la voz, un usuario puede expresar una palabra clave predeterminada para desencadenar el patrón de activación de voz para ejecutar la activación de voz.

En S204, el patrón de activación de voz genera un resultado para determinar si se debe ejecutar la activación de voz, en donde el patrón de activación de voz incluye una DNN y un CTC.

En la realización de esta especificación, comparado con los datos de voz específicos de palabras clave mencionados en el antecedente, los datos de voz generales descritos en la etapa S202 están menos restringidos y, por lo tanto, son fácilmente accesibles. Por ejemplo, puede ser un corpus de Reconocimiento de Voz Continuo de Vocabulario Grande (LVCSR) o similar.

En la realización de esta especificación, la DNN que se incluye en el patrón de reconocimiento de voz puede predecir una distribución de probabilidad posterior de una secuencia de fonemas de pronunciación correspondiente a características de voz de entrada. La DNN puede ser seguida por el CTC para dar una puntuación de confianza correspondiente a la secuencia de fonemas de pronunciación predicha. Se puede generar un resultado para determinar si se debe ejecutar la activación de voz con base en la puntuación de confianza.

Con el método de la Figura 2, en lugar de depender de datos de voz específicos de palabras clave, el patrón de activación de voz se puede entrenar con los datos de voz generales accesibles, y además el patrón de activación de voz entrenado se puede utilizar para la activación de voz, lo cual es favorable para mejorar la precisión de la activación de voz.

El patrón de activación de voz también rompe las restricciones de los datos de voz específicos de palabras clave y soportar palabras clave activadas desencadenadas por el usuario. Por lo tanto, es más conveniente y flexible en aplicaciones reales y favorable para mejorar la experiencia del usuario.

Con base en el método de la Figura 2, se proporcionan además algunas soluciones de implementación específicas y soluciones ampliadas del método en las realizaciones de la especificación, las cuales se describen a continuación.

Para facilitar la comprensión, se proporciona un diagrama esquemático de un esquema conceptual del patrón de activación de voz de la Figura 2 en una realización de esta especificación, como se muestra en la Figura 3.

El esquema conceptual de la Figura 3 incluye un módulo de extracción de características, una DNN, y un CTC en secuencia. En la etapa S204, la generación, mediante el patrón de activación de voz, de un resultado para determinar si ejecutar la activación de voz puede incluir específicamente:

extraer características acústicas a partir de los datos de voz de entrada;

ingresar las características acústicas en la DNN para el procesamiento con el fin de obtener una probabilidad de clase de las características acústicas correspondientes respectivamente a cada fonema de pronunciación;

ingresar la probabilidad de clase en el CTC para el procesamiento con el fin de obtener una puntuación de confianza de un término de activación de voz correspondiente a una secuencia de fonema de pronunciación; y

determinar si se debe ejecutar la activación de acuerdo con la puntuación de confianza, y generar un resultado de determinación.

Diversas partes del esquema conceptual de la Figura 3 se describen en detalle en combinación adicional con la Figura 4, la Figura 5, y la Figura 6 de acuerdo con el anterior flujo de activación de voz.

La Figura 4 es un diagrama esquemático de extracción de características del módulo de extracción de características en la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación.

En la Figura 4, cuando se monitoriza actualmente una sección de voz “Zhi Ma Kai Men”, una secuencia de etiqueta objetivo correspondiente a la misma es una secuencia de fonemas de pronunciación, la cual se puede expresar como: “zhilma2kailmen2”, en donde los números representan tonos. Además de los fonemas tales como iniciales y finales, los fonemas de tono también se tienen en cuenta como una unidad de modelado. En una aplicación real, se pueden tomar en cuenta todos los fonemas independientes o dependientes del contexto, entre los cuales el último es más numeroso. Sin embargo, en consideración de reducir la carga computacional posterior de la DNN, es preferible considerar solo los fonemas independientes del contexto, específicamente 72 unidades de fonemas chinos independientes del contexto, que incluyen una unidad en blanco.

Las características acústicas se pueden extraer mediante el módulo de extracción de características a partir de los datos de voz de entrada, los cuales pueden incluir específicamente la extracción de marcos de características acústicas de los datos de voz de entrada a partir de una ventana de acuerdo con un intervalo especifico de tiempo, en donde cada uno de los marcos de características acústicas son energías del banco de filtros de registro multidimensional; apilar una pluralidad de marcos de características acústicas adyacentes respectivamente; tomar los marcos de características acústicas apiladas respectivamente como características acústicas extraídas a partir de los datos de voz generales; y además, los marcos de características acústicas apilados se pueden utilizar como entradas de la DNN respectivamente.

Las energías del banco de filtros de registro se refieren a señales de energía extraídas por un banco de filtros de registro, las cuales pueden expresarse como un vector en la solución de esta especificación para facilitar el procesamiento del patrón. La multi-dimensión en lo anterior representa múltiples dimensiones del vector.

Por ejemplo, una longitud especificada de una ventana de tiempo puede ser de 25 milisegundos, cada ventana de tiempo puede moverse por 10 milisegundos, y la multi-dimensión puede ser, por ejemplo, de 40 dimensiones. En un eje de tiempo de los datos de voz, se pueden utilizar milisegundos a partir de 0 a 25 como una ventana, y las energías del banco de filtros de registro de 40 dimensiones se extraen correspondientemente a partir de los datos de voz para que sirvan como un primer marco de características acústicas; se pueden utilizar milisegundos a partir de 10 a 35 como una ventana, y las energías del banco de filtros de registro de 40 dimensiones se extraen correspondientemente a partir de los datos de voz para que sirvan como un segundo marco de características acústicas; y se pueden extraer múltiples marcos de características acústicas de la misma manera.

Además, el objetivo de apilar una pluralidad de marcos de características acústicas adyacentes es permitir más información a partir de un contexto de un marco actual, lo cual conduce a mejorar la precisión de los resultados de predicción posteriores. Siguiendo el ejemplo anterior, el marco actual, los diez marcos consecutivos adyacentes antes del marco actual, y los cinco marcos consecutivos adyacentes después del marco actual pueden apilarse, por ejemplo, para obtener una característica de apilamiento de 640 dimensiones para ingresar a la DNN posterior. Además, el medio cepstral y la normalización de la varianza se pueden llevar a cabo en las dimensiones de la característica de apilamiento, y luego se puede llevar a cabo la entrada hacia atrás.

Se debe observar que la manera de extracción de características y los parámetros adoptados en el ejemplo anterior son solo de ejemplo, y pueden ajustarse de acuerdo como sea necesario en aplicaciones reales.

La Figura 5 es un diagrama estructural esquemático de la DNN en la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación.

En la Figura 5, diversas neuronas de la DNN están completamente conectadas. Las características acústicas extraídas por el módulo de extracción de características de la Figura 3 se ingresan en la DNN. La DNN puede describir una relación entre una característica $x_0 \in R^{n0}$ acústica de entrada y una unidad j de modelado en una capa de salida de acuerdo con el siguiente mapeo de función:

$$Z_i = x_{i-1}W_i^T + B_i, 1 \leq i \leq N+1 \text{ (fórmula I)}$$

$$X_i = \sigma(z_i), 1 \leq i \leq N; \text{ (fórmula II)}$$

$$y_j = \frac{exp(z_{N+1,j})}{\Sigma_k exp(Z_{N+1,k})}; \text{ (fórmula III)}$$

donde $x_{i,i>0} \in R^{ni}$ es una salida de una capa oculta, $W_i \in R^{n_i \times ni-1}$ y $B_i \in R^{ni}$ son pesos y parámetros de desplazamiento respectivamente, n_i es el número de nodos en la capa $i^{ésima}$, $\theta=\{W_i, B_i\}$, "T" denota la transposición de una matriz, N es el número de capas ocultas, y σ es una función de activación no lineal, por ejemplo, una función de Unidad Lineal Rectificada (ReLU) σ(z) = $max(z, 0)$. La fórmula III es una función softmax, que representa la parte posterior estimada de una unidad j de etiqueta.

En una aplicación real, también se puede utilizar una Red Neuronal Recurrente (RNN) junto con el CTC. Sin embargo, en el caso de que los datos de entrenamiento estén restringidos, tal como en el Antecedente, los requisitos mínimos de consumo de energía y computación de los dispositivos móviles se pueden cumplir más fácilmente utilizando la DNN junto con el CTC. Con el fin de reducir la complejidad en el cálculo, la DNN con aproximadamente cientos de nodos en una capa oculta es más adecuado.

La Figura 6 es un diagrama estructural esquemático del CTC en la Figura 3 en un escenario de aplicación real de acuerdo con una realización de esta especificación.

El CTC está diseñado específicamente para tareas de etiquetado secuenciales. A diferencia del criterio de entropía cruzada para la alineación a nivel de marco entre las características de entrada y las etiquetas objetivo, el CTC tiene como objetivo aprender automáticamente la alineación entre los datos de voz y las secuencias de etiquetas (por ejemplo, fonemas o caracteres, etc.), eliminando así la necesidad de alineación forzada de datos, y la entrada no es necesariamente la misma que la longitud de la etiqueta.

En la Figura 6, se extrae una unidad de modelado específica a partir de L, y el CTC se ubica en una capa softmax de la DNN. La DNN se compone de una unidad |L| y una unidad en blanco. La introducción de la unidad en blanco alivia la carga de la predicción de etiqueta, ya que los símbolos no se generan de manera correspondiente durante la incertidumbre.

y_j^t($j \in [0, |L|]$, $t \in [0, T)$) se define como una probabilidad de que la DNN genera j en una etapa t de tiempo. Se dan una secuencia x^T de entrada de una longitud T de marco y una etiqueta $l^{\leq T}$ objetivo, y $l_i \in L$. Una ruta CTC $\pi=(\pi_0,...,\pi_{T-1})$ es

una secuencia de etiquetas a nivel de marco, la cual es diferente a partir de / en que la ruta CTC permite la aparición de etiquetas repetidas que no están en blanco y unidades en blanco.

La ruta π de CTC se puede mapear con su secuencia / de etiquetas correspondiente eliminando las etiquetas repetidas y las unidades en blanco. Por ejemplo, τ("*aa-b-c*") = τ("*abbcc-*") = "*abc*". Una función de mapeo de diversos a uno se define como τ, y "-" representa un espacio en blanco. Si se da x^T, y se asume que una condición de probabilidad de salida de cada etapa de tiempo es independiente, la probabilidad de la ruta π es:

$$P(\pi \mid x; \theta) = \Pi_{t=0}^{T-1} y_{\pi_t}^t; \text{ (fórmula IV)}$$

Entonces, la probabilidad de / puede calcularse con base en τ sumando las probabilidades de todas las rutas mapeadas a /. En aplicaciones reales, es problemático sumar todas las rutas en el CTC en términos de cálculo. Con respecto a este problema, se puede adoptar un algoritmo de programación dinámica hacia adelante y hacia atrás. Todas las posibles rutas de CTC se representan de manera compacta como cuadrículas con base en el algoritmo, tal como se muestra en la Figura 6.

En el momento del entrenamiento, el CTC tiene como objetivo habilitar $\theta^* = argmin_\theta \Sigma_{(x,l)\in S} -log(p(l|x; \theta))$, donde S representa los datos de entrenamiento utilizados. En el momento de la decodificación, cuando la puntuación de confianza generada por el CTC es mayor que un umbral establecido, un motor de detección puede tomar una decisión positiva en consecuencia, y se puede considerar que se han detectado las palabras clave correspondientes. El umbral establecido se puede ajustar con base en un conjunto de datos de verificación.

En la realización de esta especificación, el patrón puede entrenarse mediante un método de descenso de gradiente, preferiblemente mediante un método de descenso de gradiente aleatorio asincrónico, para optimizar iterativamente los parámetros en el patrón de activación de voz hasta que el entrenamiento converja.

Por ejemplo, la DNN y el CTC pueden entrenarse en un servidor que tiene una Unidad de Procesamiento de Gráficos (GPU). Los parámetros de Red se inicializan aleatoriamente para distribuirse uniformemente dentro de un rango de (-0.02, 0.02), una tasa de aprendizaje inicial es 0.008, y un momento es 0.9.

Para facilitar la comprensión, la tasa de aprendizaje y su función se describen a continuación.

La tasa de aprendizaje es un parámetro utilizado en el método de descenso de gradiente. En el método de descenso de gradiente, se puede inicializar primero una solución, y con base en esta solución, se determina una dirección de movimiento y un tamaño de etapa de movimiento, de tal modo que después de que la solución inicial se mueva de acuerdo con la dirección y el tamaño de etapa, se puede reducir la salida de una función objetivo. Luego se actualiza a una nueva solución, se busca continuamente una siguiente dirección de movimiento y un tamaño de etapa siguiente, y después de que este proceso se realiza de forma iterativa, la función objetivo se reduce constantemente, para finalmente encontrar una solución, de tal modo que la función objetivo sea relativamente pequeña. En el proceso de encontrar la solución, si el tamaño de la etapa es demasiado grande y la búsqueda no es lo suficientemente cuidadosa, se puede omitir una buena solución, y si el tamaño de la etapa es demasiado pequeño, el proceso de búsqueda de la solución procederá con demasiada lentitud. Por lo tanto, es importante definir el tamaño de la etapa de manera adecuada. La tasa de aprendizaje se utiliza para ajustar el tamaño de la etapa original. En el método de descenso de gradiente, el tamaño de la etapa en cada ajuste es igual a la tasa de aprendizaje multiplicada por un gradiente.

En el entrenamiento anterior, también se puede utilizar un conjunto de datos de verificación para una verificación cruzada del patrón de activación de voz para determinar si el entrenamiento converge.

Con el fin de mejorar el rendimiento y la robustez del patrón de activación de voz, se proporcionan además más medidas en las soluciones de esta especificación.

Una medida es el entrenamiento adaptativo. Específicamente, de acuerdo con la invención, se ajusta un patrón general con datos de voz de algunas palabras clave específicas y con una tasa de aprendizaje relativamente baja. Con base en esta consideración, cuando se entrena el patrón de activación de voz, también se adquieren datos de voz específicos de palabras clave, y el patrón de activación de voz se entrena con los datos de voz específicos de palabras clave. Una tasa de aprendizaje utilizada en el entrenamiento es menor que la utilizada en el entrenamiento del patrón de activación de voz con los datos de voz generales.

Otra medida es el aprendizaje por transferencia. Específicamente, los parámetros de red pueden no inicializarse aleatoriamente, sino que se refieren a una red correspondiente existente la cual tiene la misma estructura de topología que la red objetivo, excepto por unidades detalladas en la capa de salida, y puede utilizar un criterio de entropía cruzada. La transferencia de aprendizaje se puede considerar especialmente cuando los datos de entrenamiento tienen una gran escala.

Otras medidas incluyen, por ejemplo, el uso de instrucciones vectoriales relacionadas con la arquitectura (por ejemplo, ARM's NEON) para acelerar aún más la multiplicación, y así sucesivamente.

Como se mencionó anteriormente, la solución de esta especificación puede soportar fácilmente palabras clave de activación de voz definidas por el usuario. Se puede determinar una secuencia de etiquetas objetivo correspondiente a dichas palabras clave definidas por el usuario a través de un diccionario.

Un método de activación de voz proporcionado en la realización de esta especificación es como se describe en lo anterior. Con base en la misma idea de especificación, se proporciona además un aparato correspondiente en una realización de esta especificación, como se muestra en la Figura 7.

La Figura 7 es un diagrama estructural esquemático de un aparato de activación de voz correspondiente a la Figura 2 de acuerdo con una realización de esta especificación. El aparato puede estar ubicado en el cuerpo ejecutivo del flujo de la Figura 2, que incluye un módulo 701 de entrada y un patrón 702 de activación de voz.

Los datos de voz son ingresados por el módulo 701 de entrada al patrón 702 de activación de voz entrenado con datos de voz generales, y el patrón 702 de activación de voz genera un resultado para determinar si se debe ejecutar la activación de voz, en donde el patrón de activación de voz incluye una DNN y un CTC.

Opcionalmente, los datos de voz generales incluyen un corpus LVCSR.

El aparato incluye además un módulo 703 de entrenamiento; y el entrenamiento, por el módulo 703 de entrenamiento, el patrón de activación de voz con los datos de voz generales incluye:

optimizar iterativamente, mediante el módulo 703 de entrenamiento, los parámetros en el patrón de activación de voz con los datos de voz generales mediante un método de descenso de gradiente estocástico asincrónico hasta que el entrenamiento converja.

El módulo 703 de entrenamiento adquiere además datos de voz específicos de palabras clave; y entrena el patrón de activación de voz con los datos de voz específicos de palabras clave, en donde una tasa de aprendizaje utilizada en el entrenamiento es menor que la utilizada en el entrenamiento del patrón de activación de voz con los datos de voz generales.

Opcionalmente, el módulo 703 de entrenamiento verifica de manera cruzada el patrón de activación de voz con un conjunto de datos de verificación en el entrenamiento para determinar si el entrenamiento converge.

Opcionalmente, la generación, mediante el patrón 702 de activación de voz, de un resultado para determinar si ejecutar la activación de voz incluye específicamente:

extraer, mediante el patrón 702 de activación de voz, características acústicas a partir de los datos de voz de entrada;

ingresar las características acústicas en la DNN que se incluye en el patrón 702 de activación de voz para el procesamiento con el fin de obtener una probabilidad de clase de las características acústicas correspondientes respectivamente a cada fonema de pronunciación;

ingresar la probabilidad de clase en el CTC que se incluye en el patrón 702 de activación de voz para el procesamiento con el fin de obtener una puntuación de confianza de un término de activación de voz correspondiente a una secuencia de fonemas de pronunciación; y

determinar si se debe ejecutar la activación de acuerdo con la puntuación de confianza, y generar un resultado de determinación.

Opcionalmente, la extracción, mediante el patrón 702 de activación de voz, las características acústicas a partir de los datos de voz de entrada incluyen específicamente:

extraer, mediante el patrón 702 de activación de voz, marcos de características acústicas de los datos de voz de entrada a partir de una ventana de acuerdo con un intervalo de tiempo especificado, en donde cada uno de los marcos de características acústicas son energías del banco de filtros de registro multidimensionales;

apilar una pluralidad de marcos de características acústicas adyacentes, respectivamente; y

tomar los marcos de características acústicas apiladas respectivamente como características acústicas extraídas a partir de la voz monitorizada.

Con base en la misma idea de especificación, se proporciona además un dispositivo electrónico correspondiente en un ejemplo de esta especificación, que incluye:

al menos un procesador; y

una memoria comunicativamente conectada a al menos un procesador; en donde

la memoria almacena una instrucción ejecutable por el al menos un procesador, y la instrucción es ejecutada por el al menos un procesador para permitir que el al menos un procesador:

ingrese datos de voz a un patrón de activación de voz entrenado con datos de voz generales, y generar, mediante el patrón de activación de voz, un resultado para determinar si se debe ejecutar la activación de voz, en donde el patrón de activación de voz incluye una DNN y un CTC.

Con base en la misma idea de especificación, un medio de almacenamiento informático no volátil correspondiente con una instrucción ejecutable por ordenador almacenada en el mismo se proporciona además en un ejemplo de esta especificación, en donde la instrucción ejecutable por ordenador está configurada para:

ingresar datos de voz a un patrón de activación de voz entrenado con datos de voz generales, y generar, mediante el patrón de activación de voz, un resultado para determinar si se debe ejecutar la activación de voz, en donde el patrón de activación de voz incluye una DNN y un CTC.

En lo anterior se han descrito las realizaciones específicas de esta especificación. Otras realizaciones caen dentro del alcance de las reivindicaciones adjuntas. Bajo algunas circunstancias, las acciones o etapas descritas en las reivindicaciones pueden realizarse en una secuencia diferente a la de las realizaciones y aún pueden lograr un resultado deseado. Además, los procesos representados en los dibujos adjuntos no son necesariamente requeridos para lograr el resultado deseado de acuerdo con la secuencia específica o secuencia consecutiva mostrada. El procesamiento multitarea y el procesamiento paralelo también son posibles o pueden ser ventajosos en algunas formas de implementación.

Las realizaciones en la especificación se describen progresivamente, se pueden obtener partes idénticas o similares de las realizaciones con referencia entre sí, y cada realización enfatiza una parte diferente a partir de otras realizaciones. Especialmente, las realizaciones del aparato, dispositivo electrónico, y medio de almacenamiento informático no volátil son básicamente similares a las realizaciones del método, por lo tanto, se describen de manera simple. Para las partes relacionadas, consultar las descripciones de las partes en las realizaciones del método.

El aparato, el dispositivo electrónico, y el medio de almacenamiento informático no volátil que se proporciona en las realizaciones de esta especificación corresponden al método. Por tanto, el aparato, el dispositivo electrónico, y el medio de almacenamiento informático no volátil también tienen efectos técnicos beneficiosos similares a los del método correspondiente. Como los efectos técnicos beneficiosos del método se han descrito en detalle en lo anterior, los efectos técnicos beneficiosos del aparato, el dispositivo electrónico, y el medio de almacenamiento informático no volátil no se detallarán aquí.

En la década de 1990, una mejora de una tecnología se puede distinguir obviamente como una mejora en el hardware (por ejemplo, una mejora en la estructura de un circuito tal como un diodo, un transistor, y un interruptor) o una mejora en el software (una mejora en un procedimiento del método). Sin embargo, con el desarrollo de tecnologías, las mejoras de diversos procedimientos de métodos en la actualidad pueden considerarse como mejoras directas en las estructuras de circuitos de hardware. Casi todos los diseñadores programan los procedimientos de métodos mejorados en circuitos de hardware para obtener las estructuras de circuito de hardware correspondientes. Por lo tanto, es inapropiado asumir que la mejora de un procedimiento del método no se puede implementar mediante el uso de un módulo de entidad de hardware. Por ejemplo, un Dispositivo Lógico Programable (PLD) (por ejemplo, una Matriz de Puerta Programable en Campo (FPGA)) es un dicho circuito integrado, y sus funciones lógicas están determinadas por un dispositivo de programación del usuario. Los diseñadores programan por sí mismos para “integrar” un sistema digital en un PLD, sin solicitarle al fabricante del chip que diseñe y fabrique un chip de circuito integrado dedicado. Además, en la actualidad, la programación se implementa principalmente mediante el uso de software compilador lógico en lugar de fabricar manualmente un chip de circuito integrado. El software del compilador lógico es similar a un compilador de software utilizado para desarrollar y escribir un programa, y los códigos originales antes de compilar también deben escribirse utilizando un lenguaje de programación específico, lo cual se denomina como Lenguaje de Descripción de Hardware (HDL). Existen diversos tipos de HDLs, tales como Lenguaje de Expresión Booleano Avanzado (ABEL), Lenguaje de Descripción de Hardware Altera (AHDL), Confluencia, Lenguaje de Programación de Universidad de Cornell (CUPL), HDCal, Lenguaje de Descripción de Hardware Java (JHDL), Lava, Lola, MyHDL, PALASM, y Lenguaje de Descripción de Hardware Ruby (RHDL), entre los cuales el Lenguaje de Descripción de Hardware de Circuito Integrado de Muy Alta Velocidad (VHDL) y Verilog se utilizan con mayor frecuencia en la actualidad. Los expertos en la técnica también deben saber que un circuito de hardware para implementar el procedimiento del método lógico puede obtenerse fácilmente programando de manera lógicamente ligera el procedimiento del método utilizando los diversos lenguajes de descripción de hardware anteriores y programándolo en un circuito integrado.

Se puede implementar un controlador de cualquier manera adecuada. Por ejemplo, el controlador puede emplear una forma de un microprocesador o un procesador y un medio legible por ordenador que almacena códigos de programa legibles por ordenador (tales como software o firmware) ejecutables por el microprocesador o procesador, una puerta lógica, un interruptor, un Circuito Integrado Específico de la Aplicación (ASIC), un controlador lógico programable, y un microcontrolador incorporado. Los ejemplos del controlador incluyen, pero no se limitan a, los siguientes microcontroladores: ARC 625D, Atmel AT91SAM, Microchip PIC18F26K20, y Silicone Labs C8051F320. El controlador de la memoria se puede implementar además como una parte de la lógica de control de la memoria. Los expertos en la técnica también saben que, además de implementar el controlador mediante el uso de códigos de programa legibles por ordenador, es completamente factible programar lógicamente las etapas del método para permitir que el

controlador implemente la misma función en una forma de una puerta lógica, un interruptor, un ASIC, un controlador lógico programable, y un microcontrolador incorporado. Por lo tanto, dicho controlador puede considerarse como un componente de hardware, y los aparatos que se incluyen en el controlador y configurados para implementar diversas funciones también pueden considerarse como estructuras dentro del componente de hardware. O bien, los aparatos configurados para implementar diversas funciones pueden incluso considerarse como módulos de software configurados para implementar el método y estructuras dentro del componente de hardware.

El sistema, aparatos, módulos o unidades que se ilustran en las realizaciones anteriores pueden implementarse específicamente mediante un chip de ordenador o una entidad, o implementarse mediante un producto que tiene una función específica. El dispositivo de implementación atípico es un ordenador. Específicamente, por ejemplo, el ordenador puede ser un ordenador personal, un ordenador portátil, un teléfono celular, un teléfono con cámara, un teléfono inteligente, un asistente digital personal, un reproductor multimedia, un dispositivo de navegación, un dispositivo de correo electrónico, una consola de juegos, una tableta, un dispositivo portátil, o una combinación de cualquiera de estos dispositivos.

Para facilitar la descripción, el aparato se divide en diversas unidades con base en funciones, y las unidades se describen de manera separada. En una implementación de esta especificación, las funciones de diversas unidades también se pueden implementar en una o más piezas de software y/o hardware.

Los expertos en la técnica deben comprender que las realizaciones de esta especificación pueden proporcionarse como un método, un sistema, o un producto de programa informático. Por lo tanto, las realizaciones de esta especificación pueden implementarse en una forma de una realización de hardware completa, una realización de software completa, o una realización que combina software y hardware. Además, las realizaciones de esta especificación pueden estar la forma de un producto de programa informático implementado en uno o más medios de almacenamiento utilizables por ordenador (que incluyen, pero no se limitan a, una memoria de disco magnético, un CD-ROM, una memoria óptica y similares) que incluyen los códigos de programa utilizables por ordenador.

Esta especificación se describe con referencia a diagramas de flujo y/o diagramas de bloques del método, el dispositivo (sistema) y el producto de programa informático de acuerdo con las realizaciones de esta especificación. Debe entenderse que se pueden utilizar instrucciones de programas informáticos para implementar cada proceso y/o bloque en los diagramas de flujo y/o diagramas de bloques y combinaciones de procesos y/o bloques en los diagramas de flujo y/o diagramas de bloques. Las instrucciones del programa informático se pueden proporcionar a un ordenador de propósito general, un ordenador de propósito especial, un procesador incorporado o un procesador de otro dispositivo de procesamiento de datos programable para generar una máquina, tal como el ordenador o el procesador de otro dispositivo de procesamiento de datos programable ejecutan una instrucción para generar un aparato configurado para implementar funciones designadas en uno o más procesos en un diagrama de flujo y/o uno o más bloques en un diagrama de bloques.

Las instrucciones del programa informático también pueden almacenarse en una memoria legible por ordenador que puede guiar al ordenador u otro dispositivo de procesamiento de datos programable para trabajar de una manera específica, de tal modo que la instrucción almacenada en la memoria legible por ordenador genera un artículo de fabricación que incluye un aparato de instrucción, y el aparato de instrucción implementa funciones designadas por uno o más procesos en un diagrama de flujo y/o uno o más bloques en un diagrama de bloques.

Las instrucciones del programa informático también pueden cargarse en el ordenador u otro dispositivo de procesamiento de datos programable, de tal modo que se ejecuten una serie de etapas de operación en el ordenador u otro dispositivo programable para generar un procesamiento implementado por ordenador y, por lo tanto, la instrucción ejecutada en el ordenador u otro dispositivo programable proporciona etapas para implementar funciones designadas en uno o más procesos en un diagrama de flujo y/o uno o más bloques en un diagrama de bloques.

En una configuración típica, el dispositivo informático incluye una o más unidades centrales de procesamiento (CPUs), una interfaz de entrada/salida, una interfaz de red, y una memoria.

La memoria puede incluir medios legibles por ordenador, tales como una memoria volátil, una Memoria de Acceso Aleatorio (RAM), y/o una memoria no volátil, por ejemplo, una Memoria de Solo Lectura (ROM) o una memoria RAM flash. La memoria es un ejemplo de un medio legible por ordenador.

El medio legible por ordenador incluye medios no volátiles y volátiles, así como medios móviles y no móviles, y puede implementar el almacenamiento de información a través de cualquier método o tecnología. La información puede ser una instrucción legible por ordenador, una estructura de datos, y un módulo de un programa u otros datos. Un ejemplo del medio de almacenamiento de un ordenador incluye, pero no se limita a, una memoria de cambio de fase (PRAM), una memoria estática de acceso aleatorio (SRAM), una memoria dinámica de acceso aleatorio (DRAM), otros tipos de RAM, una ROM, una memoria de solo lectura programable y borrable eléctricamente (EEPROM), una memoria flash u otras tecnologías de memoria, una memoria de disco compacto de solo lectura (CD-ROM), un disco versátil digital (DVD) u otros almacenamientos ópticos, una cinta de casete, un almacenamiento en cinta magnética/disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que no sea de transmisión, y se puede

utilizar para almacenar información accesible al dispositivo informático. De acuerdo con la definición en este texto, el medio legible por ordenador no incluye los medios transitorios, tales como una señal de datos modulada y un portador.

Cabe señalar además que los términos “incluye”, “comprende” o cualquier otra variación de los mismos están destinados a cubrir la inclusión no exclusiva, de tal modo que un proceso, método, producto o dispositivo que incluye una serie de elementos no solo incluye los elementos, sino también incluye otros elementos no enumerados expresamente, o incluye además elementos inherentes al proceso, método, producto o dispositivo. En ausencia de más limitaciones, un elemento definido por “que incluye un/una...” no excluye que el proceso, método, producto o dispositivo que incluye el elemento tenga además otros elementos idénticos.

Esta especificación se puede describir en un contexto general de una instrucción ejecutable por ordenador ejecutada por un ordenador, por ejemplo, un módulo de programa. En general, el módulo de programa incluye una rutina, un programa, un objeto, un conjunto, una estructura de datos, y similares utilizados para ejecutar una tarea específica o implementar un tipo de datos abstractos específicos. Esta especificación también se puede implementar en entornos informáticos distribuidos. En estos entornos informáticos distribuidos, se ejecuta una tarea mediante el uso de dispositivos de procesamiento remoto conectados a través de una red de comunicaciones. En los entornos informáticos distribuidos, el módulo de programa puede estar ubicado en un medio de almacenamiento informático local y remoto que incluye un dispositivo de almacenamiento.

Las realizaciones en la especificación se describen progresivamente, se pueden obtener partes idénticas o similares de las realizaciones con referencia entre sí, y cada realización enfatiza una parte diferente de otras realizaciones. Especialmente, la realización del sistema es básicamente similar a las realizaciones del método, por lo tanto, se describe de manera simple. Para las partes relacionadas, consultar las descripciones de las partes en las realizaciones del método.

La descripción anterior es simplemente de realizaciones de esta especificación, y no se utiliza para limitar esta aplicación.

REIVINDICACIONES

1. Un método de activación de voz implementado por ordenador, que comprende:

entrenar un patrón de activación de voz con datos de voz generales, en donde el entrenamiento incluye:
optimizar iterativamente los parámetros en el patrón de activación de voz con los datos de voz generales a través de un método de descenso de gradiente estocástico asíncrono hasta que el entrenamiento converja;
adquirir datos de voz específicos de palabras clave, y entrenar el patrón de activación de voz con los datos de voz específicos de palabras clave, en donde una tasa de aprendizaje utilizada en el entrenamiento con los datos de voz específicos de palabras clave es menor que la utilizada en el entrenamiento del patrón de activación de voz con los datos de voz generales;
ingresar (S202) datos de voz en el patrón de activación de voz entrenado con los datos de voz generales y los datos de voz específicos de palabras clave; y
generar (S204), mediante el patrón de activación de voz, un resultado para determinar si se ejecuta la activación de voz,
en donde el patrón de activación de voz comprende una Red Neuronal Profunda, DNN, y un Clasificador Temporal Conexionista, CTC en secuencia.

2. El método de la reivindicación 1, en donde los datos de voz generales comprenden un corpus LVCSR de Reconocimiento Continuo de Voz de Vocabulario Extenso.

3. El método de la reivindicación 1, comprendiendo además:
una verificación cruzada del patrón de activación de voz con un conjunto de datos de verificación en el entrenamiento para determinar si el entrenamiento converge.

4. El método de cualquiera de las reivindicaciones 1 a 3, en donde la generación, mediante el patrón de activación de voz, de un resultado para determinar si ejecuta la activación de voz comprende:

extraer características acústicas a partir de los datos de voz de entrada;
ingresar las características acústicas en la DNN comprendida en el patrón de activación de voz para el procesamiento con el fin de obtener una probabilidad de clase de las características acústicas correspondientes respectivamente a cada fonema de pronunciación;
ingresar la probabilidad de clase en el CTC comprendido en el patrón de activación de voz para el procesamiento con el fin de obtener una puntuación de confianza de un término de activación de voz correspondiente a una secuencia de fonemas de pronunciación;
determinar si ejecuta la activación de acuerdo con la puntuación de confianza; y
generar un resultado de determinación.

5. El método de la reivindicación 4, en donde la extracción de características acústicas a partir de los datos de voz de entrada comprende:

extraer marcos de características acústicas de los datos de voz de entrada a partir de una ventana de acuerdo con un intervalo de tiempo especificado, en donde cada uno de los marcos de características acústicas son energías del banco de filtros de registro multidimensional;
apilar una pluralidad de marcos de características acústicas adyacentes respectivamente; y
tomar los marcos de características acústicas apiladas respectivamente como características acústicas extraídas a partir de la voz monitorizada.

6. Un dispositivo electrónico, que comprende:

al menos un procesador; y
una memoria conectada comunicativamente a el al menos un procesador, en donde la memoria almacena instrucciones ejecutables por el al menos un procesador y las instrucciones son ejecutadas por el al menos un procesador para hacer que el dispositivo electrónico realice el método de cualquiera de las reivindicaciones 1 a 5.

7. Un medio legible por ordenador que almacena instrucciones las cuales, cuando son ejecutadas por al menos un procesador de un dispositivo electrónico, hacen que el dispositivo electrónico realice el método de cualquiera de las reivindicaciones 1 a 5.

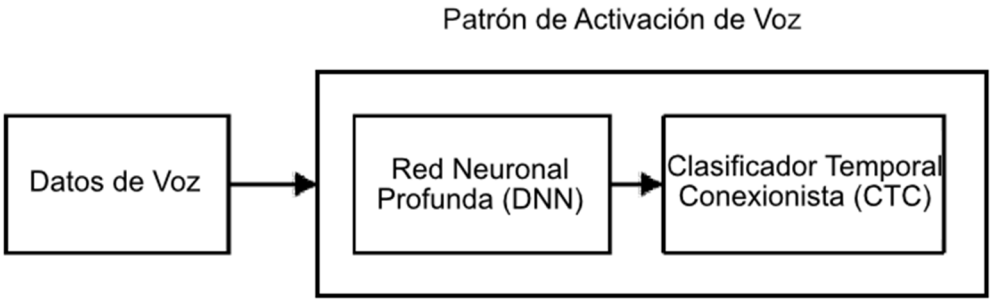
Patrón de Activación de Voz
Datos de Voz
Red Neuronal Profunda (DNN)
Clasificador Temporal Conexionista (CTC)

FIG. 1

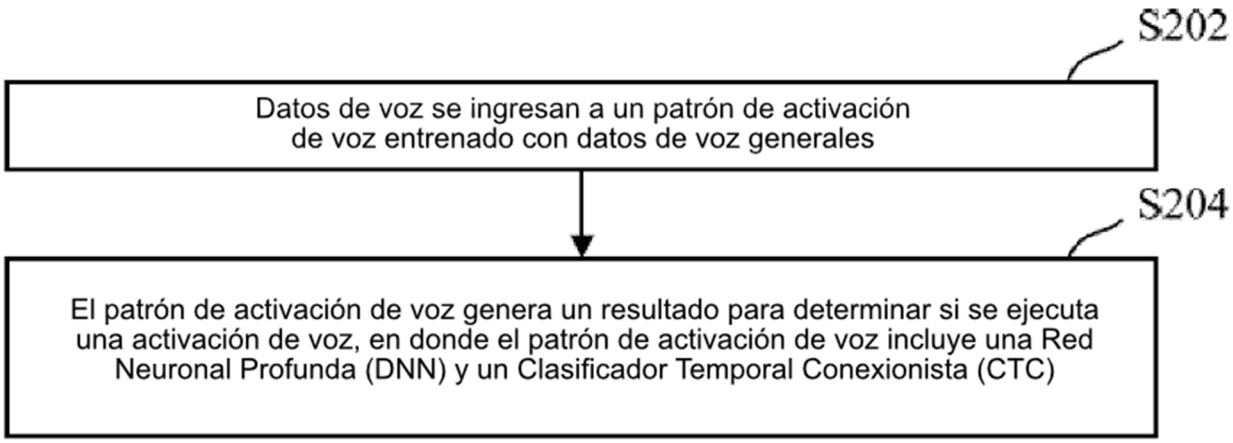
S202
Datos de voz se ingresan a un patrón de activación de voz entrenado con datos de voz generales
S204
El patrón de activación de voz genera un resultado para determinar si se ejecuta una activación de voz, en donde el patrón de activación de voz incluye una Red Neuronal Profunda (DNN) y un Clasificador Temporal Conexionista (CTC)

FIG. 2

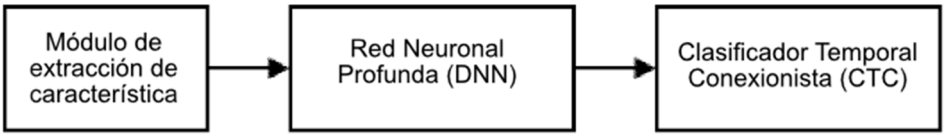

FIG. 3

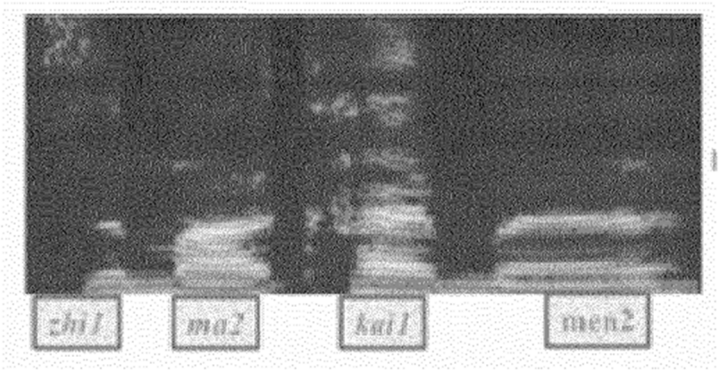

FIG. 4

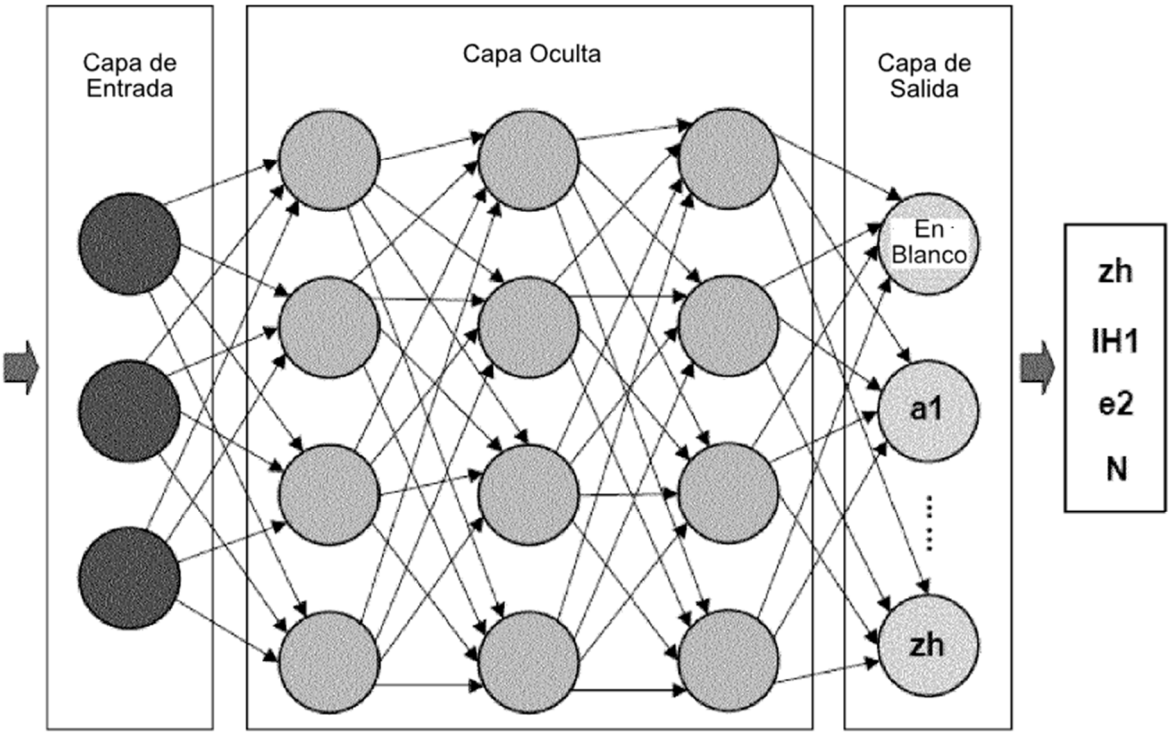

FIG. 5

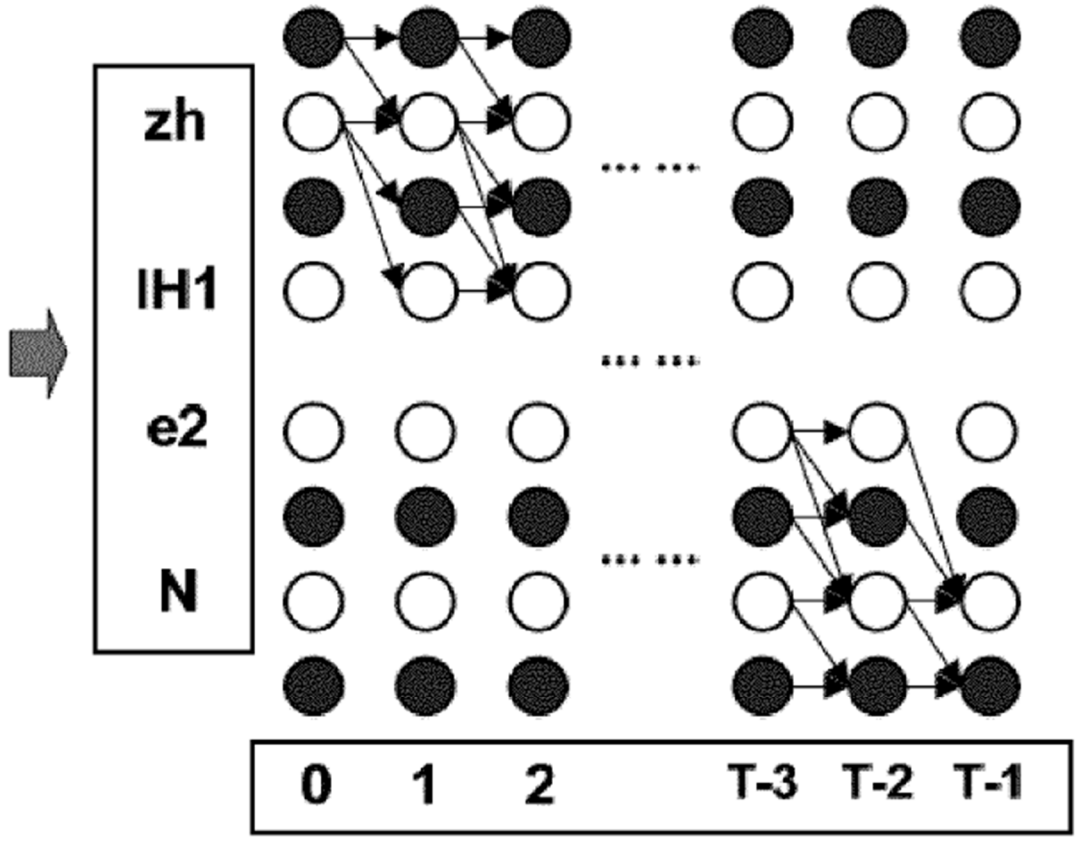
zh
IH1
e2
N
0
1
2
T-3
T-2
T-1

FIG. 6

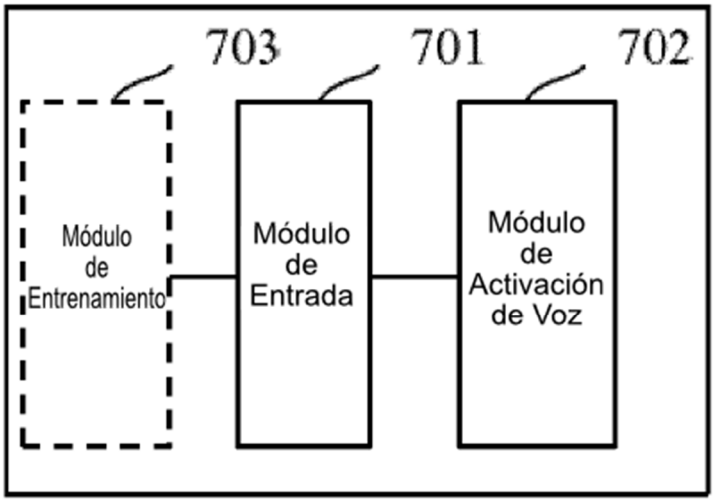
703
701
702
Módulo de Entrenamiento
Módulo de Entrada
Módulo de Activación de Voz

FIG. 7