(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 981 641**

(21) Número de solicitud: 202231062

(51) Int. CI.:

G10L 15/16 (2006.01)

(12) PATENTE DE INVENCIÓN CON EXAMEN B2

(22) Fecha de presentación:

13.12.2022

(43) Fecha de publicación de la solicitud:

09.10.2024

Fecha de concesión:

04.03.2025

(45) Fecha de publicación de la concesión:

11.03.2025

(73) Titular/es:

UNIVERSIDAD DE LEÓN (50.00%)
Avda. de la Facultad, 25
24071 León (León) ES y
S.M.E. INSTITUTO NACIONAL DE CIBERSEGURIDAD DE ESPAÑA M.P., S.A. (50.00%)

(72) Inventor/es:

VASCO CAROFILIS, Roberto Andrés;
FERNÁNDEZ ROBLES, Laura;
ALEGRE GUTIÉRREZ, Enrique y
FIDALGO FERNÁNDEZ, Eduardo

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

(54) Título: **SISTEMA, MÉTODO Y PRODUCTO DE PROGRAMA PARA LA CLASIFICACIÓN AUTOMÁTICA DEL ACENTO EN SEÑALES DE AUDIO**

(57) Resumen:

Sistema, método y producto de programa para la clasificación automática del acento en señales de audio. El método comprende recibir (110) una señal de audio (3); obtener (120) un espectrograma (4) de la señal de audio (3); obtener (130), utilizando una red neuronal (15) alimentada por el espectrograma (4) y entrenada para clasificar una pluralidad de acentos, un mapa de calor Grad-CAM (6) por cada uno de los acentos para los que la red neuronal (15) ha sido entrenada; reducir (140) la dimensionalidad de cada mapa de calor Grad-CAM (6) para obtener un respectivo tensor reducido (8); y obtener (150) una clasificación del acento (11) de la señal de audio (3) utilizando un modelo de aprendizaje automático entrenado (16), alimentado por el espectrograma (4) y los tensores reducidos (8). La presente invención permite obtener un aumento de la precisión en la clasificación de acentos de señales de audio.

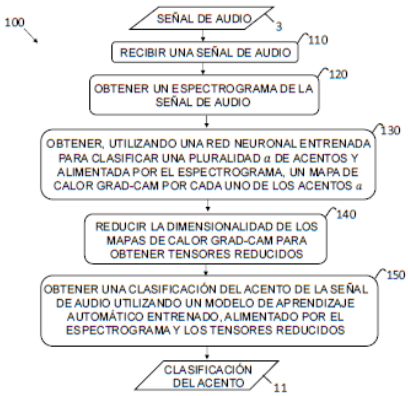


FIG. 1

DESCRIPCIÓN

SISTEMA, MÉTODO Y PRODUCTO DE PROGRAMA PARA LA CLASIFICACIÓN AUTOMÁTICA DEL ACENTO EN SEÑALES DE AUDIO

OBJETO DE LA INVENCIÓN

La presente invención se engloba en los sistemas y procedimientos para la clasificación automática del acento de hablantes de señales de audio contenidas en ficheros de audio.

ANTECEDENTES DE LA INVENCIÓN

Debido a Internet, la existencia de las redes sociales y la gran cantidad de datos que se comparten, el número de audios y videos ha crecido notablemente en los últimos años. Para poder extraer información de utilidad de ficheros de audio, o en el audio contenido en los videos, surge la necesidad de poder realizar procesos automáticos que permitan extraer y manejar dicha información.

La voz humana proporciona información valiosa sobre la persona que la genera, información que se podría utilizar en una amplia variedad de tareas. El acento particular de cada individuo proporciona parte de esta información. Un acento es una forma particular de pronunciación que se puede asociar con la localidad en la que residen sus hablantes, su etnia, la influencia de su lengua materna, o incluso de su situación socioeconómica (T. McArthur, J. Lam-McArthur, and L. Fontaine, The Oxford companion to the English language. 2 ed., 2018).

La capacidad de identificar correctamente el acento podría permitir a un sistema obtener información relevante sobre un individuo con tan solo disponer de grabaciones de su voz o dicha voz extraída de un fichero de vídeo (L. M. Arslan and J. H. Hansen, “A study of temporal features and frequency characteristics in american english foreign accent,” The Journal of the Acoustical Society of America, vol. 102, no. 1, pp. 28–40, 1997).

La clasificación automática de acentos es un campo de investigación centrado en la creación de sistemas automáticos que puedan identificar el acento de diferentes hablantes. Además de su utilidad como método para caracterizar los rasgos del hablante, tiene utilidad en una gran variedad de campos, como en la mejora de los sistemas de reconocimiento automático de voz (en inglés, “Automatic Speech Recognition”, ASR).

Los enfoques utilizados en la identificación automática de acentos son similares a los que se utilizan en la identificación de idiomas. Estos enfoques se pueden agrupar en aquellos que utilizan el modelado de la acústica y en aquellos que utilizan el modelado fonotáctico (A. Etman and A. L. Beex, “Language and dialect identification: A survey,” in 2015 SAI Intelligent Systems Conference (IntelliSys), pp. 220–231, IEEE, 2015). El modelado acústico se centra en las características espectrales de ondas sonoras, mientras que el enfoque fonotáctico se centra en reconocimiento de fonemas y su análisis posterior.

Tradicionalmente, el modo de obtener características del audio se realiza a través del uso de espectrogramas, que son el resultado de calcular el espectro de la señal de audio dividida en ventanas temporales. El resultado es una matriz que contiene información sobre el tiempo, la frecuencia y energía de cada instante (representada por color). Estas características se han utilizado con métodos basados en Máquinas de Vectores de Soporte (C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995), Modelos Ocultos de Márkov (L. Rabiner and B. Juang, “An introduction to hidden Markov models,” IEEE ASSP Magazine, vol. 3, no. 1, pp. 4–16, 1986) o Modelos Gaussianos Mezclados (D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000).

Existen múltiples trabajos sobre el modelado acústico, que trabajan procesando la onda de audio original o sus características espectrales a través de los llamados vectores de incrustación (en inglés, “embeddings“). A través de estos vectores de incrustación es posible representar frases o expresiones de un fichero de audio en un único vector incrustado. Algunos de los vectores incrustados más conocidos y utilizados en la literatura son los vectores-i y vectores-x (en inglés, “i-vectors” y “x-vectors”, respectivamente).

El uso de sistemas de clasificación basados en el modelado acústico utilizando vectores-i ha demostrado mejores resultados que los modelados fonotácticos en tareas como el reconocimiento de idiomas (E. Singer, P. Torres-Carrasquillo, D. A. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. turim, “The mitll nist lre 2011 language recognition system,” in Odyssey 2012-The Speaker and Language Recognition Workshop, 2012). Con un conjunto pequeño de atributos, se puede obtener una caracterización completa de un idioma. Los detectores robustos de atributos de voz universales pueden ser diseñados compartiendo

datos entre diferentes idiomas, como se ha demostrado en investigaciones (S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data,” IEEE transactions on audio, speech, and language processing, vol. 20, no. 3, pp. 875–887, 2011).

Los vectores-i (H. Behravan, V. Hautam¨aki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, “I-vector modeling of speech attributes for automatic foreign accent recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 29–41, 2015) se utilizan para definir un conjunto común de unidades universales fundamentales para todos los acentos hablados que sean evaluados, por lo que cualquier expresión hablada se puede transcribir con este conjunto de unidades.

Con la aparición del aprendizaje profundo y las mejoras en las redes neuronales profundas (en inglés, “Deep Neural Networks”, DNN), en los últimos años se ha vuelto muy popular el uso de vectores-x obtenidos mediante la extracción de características profundas de DNN. Dichas características se han utilizado en tareas como la clasificación del hablante (S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, “Data augmentation using deep generative models for embedding based speaker recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2598–2609, 2020) o clasificación del idioma (L. Sun, “Spoken language identification with deep temporal neural network and multi-levels discriminative cues,” in 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), pp. 153–157, IEEE, 2020), obteniendo mejores resultados que los obtenidos con vectores-i.

Incluso existen trabajos que combinan clasificadores basados en aprendizaje profundo (redes neuronales convolucionales profundas y recurrentes) con clasificadores tradicionales como las Máquinas de Vectores de Soporte. Estos trabajos (D. Honnavalli and S. Shylaja, “Supervised machine learning model for accent recognition in English speech using sequential mfcc features”, Advances in Artificial Intelligence and Data Engineering, pp. 55–66, Springer, 2021) han obtenido resultados superiores a los conseguidos por cada clasificador de un modo independiente a la hora de realizar la clasificación automática de acentos.

La mayor parte de las investigaciones anteriores se centra en la clasificación de acentos en angloparlantes con idiomas nativos diferentes al inglés. Sin embargo, la clasificación

automática de acentos de hablantes ingleses nativos es más compleja, debido a las similitudes en la acústica entre cada uno de los acentos a clasificar. Por ese motivo, surge la necesidad de tratar de mejorar la descripción de cada hablante, para hacerlo más distintivo y poder obtener mejores resultados en el proceso de clasificación automática.

La presente invención aborda esta problemática, es decir, la necesidad de mejorar las características extraídas de un audio para la clasificación de acentos, mediante la explotación de técnicas de aprendizaje profundo supervisado y la obtención de vectores incrustados a través de redes de aprendizaje profundo entrenadas previamente en un conjunto de espectrogramas de audio.

Existen investigaciones que utilizan redes neuronales convolucionales (en inglés, “Convolutional Neural Networks”, CNNs) para clasificar acentos de hablantes (Y. Zeng, H. Mao, D. Peng, and Z. Yi, “Spectrogram based multi-task audio classification,” Multimedia Tools and Applications, vol. 78, no. 3, pp. 3705–3722, 2019) (U. Singh, A. Gupta, D. Bisharad, and W. Arif, “Foreign accent classification using deep neural nets,” Journal of Intelligent & Fuzzy Systems, no. Preprint, pp. 1–6, 2020), debido a sus altas capacidades de modelado. En algunos trabajos se utiliza una arquitectura ResNet modificada para la clasificación de acentos de los hablantes.

La presente invención propone una nueva metodología que mejora el rendimiento de los modelos de clasificación de acentos existentes en la literatura.

DESCRIPCIÓN DE LA INVENCIÓN

La presente invención se refiere a un método y un sistema para la clasificación automática del acento de hablantes en ficheros de audio utilizando aprendizaje profundo supervisado y Grad-CAMs.

La invención permite clasificar automáticamente un fichero de audio en base al acento del hablante contenido en dicho fichero. Entre otras aplicaciones, dicha clasificación automática permite mejorar los sistemas de reconocimiento automático de voz (al poderlo particularizar a un entrenamiento del habla específico para dicho acento), o determinar el posible origen de dicho hablante, acotando por ejemplo una posible búsqueda de dónde podría haber sido grabado dicho fichero de audio, reduciendo así el espacio de búsqueda.

Los resultados obtenidos por las CNNs se pueden interpretar por métodos como el mapeo de activación de clases ponderado por gradiente (en inglés, “Gradient Weighted Class Activation Mapping”, Grad-CAM) (R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017). Grad-CAM permite extraer, de forma ponderada, las entradas más representativas de cada posible resultado obtenido por una CNN. Dicho de otro modo, Grad-CAM permite extraer información sobre el conocimiento que ha adquirido una CNN durante el proceso de formación.

La presente invención utiliza Grad-CAM como un método para extraer características que representen el conocimiento obtenido por una CNN durante su entrenamiento con espectrogramas de audio. Esta nueva información se utiliza en combinación con los espectrogramas para proporcionar vectores de características más distintivos que pueden aumentar el rendimiento obtenidos con modelos clásicos. Más específicamente, tras realizar el entrenamiento de una CNN con espectrogramas, se extrae para cada audio un mapa de calor, y de dicho mapa de calor se extraen una serie de características que se concatenan con la información de los espectrogramas originales. El vector resultante final se utiliza para entrenar modelos clásicos de aprendizaje automático. En este procedimiento se propone por primera vez la utilización de Grad-CAM como extractor de características utilizando una red neuronal convolucional para la clasificación de acentos. De este modo se obtienen descriptores de audio más robustos que consiguen que el sistema de clasificación de acentos de hablantes en ficheros de audio sea más eficaz.

El método para la clasificación automática del acento en señales de audio de la presente invención comprende las siguientes etapas: recibir una señal de audio; obtener un espectrograma de la señal de audio; obtener, utilizando una red neuronal alimentada por el espectrograma y entrenada para clasificar una pluralidad de acentos, un mapa de calor Grad-CAM por cada uno de los acentos para los que la red neuronal ha sido entrenada; reducir la dimensionalidad de cada mapa de calor Grad-CAM para obtener un respectivo tensor reducido; y obtener una clasificación del acento de la señal de audio utilizando un modelo de aprendizaje automático entrenado, alimentado por el espectrograma y los tensores reducidos. En las diferentes etapas del método de clasificación del acento de la presente invención se pueden utilizar tensores, que representan diferentes tipos de agrupamientos de vectores (incluyendo vectores, matrices, etc.). En la presente descripción

cuando se utilice la palabra “vector” o “vectores” se puede considerar que es aplicable también directamente a ”tensor” o “tensores” (incluyendo por ejemplo matrices).

En una realización, el procedimiento de clasificación concatena la información en forma de vector de dos fuentes diferentes. La primera fuente extrae un vector representativo tras convertir en unidimensional (aplanar) el espectrograma extraído de un audio. La segunda fuente hace uso del aprendizaje automático profundo para realizar la codificación de los mapas de calor obtenidos de los Grad-CAMs, tras haber entrenado una red neuronal convolucional utilizando los espectrogramas de los ficheros de audio. Cada audio original queda representado por un vector resultado de concatenar los dos vectores anteriores, los cuales se utilizan para entrenar un algoritmo de aprendizaje automático supervisado. Dicho algoritmo realiza la clasificación automática de audios.

Frente a una clasificación manual por un experto, la clasificación automática del acento de un hablante en audios anula la subjetividad, los errores por cansancio y falta de atención, la disparidad de criterio entre expertos, los costes asociados al tiempo del experto y disminuye el tiempo necesario para la realización de dicha clasificación. El procedimiento de clasificación de la presente invención puede ser implementado en herramientas utilizadas por empresas y FFCCSE (Fuerzas y Cuerpos de Seguridad del Estado) para realizar la clasificación automática del acento de un hablante en un fichero de audio, bien sea una grabación o el audio extraído de un video, proveniente de la red, o de manera aislada, accediendo al fichero de audio a través de medios extraíbles de almacenamiento masivo. Esta implementación podría resultar de utilidad para extraer información de audios contenidos en dispositivos de almacenamiento masivo confiscados por las FFCCSE (o descargados de la red) en muchos casos relacionados con la lucha contra el crimen, como puede ser el abuso sexual a menores o el terrorismo. Dicha información permitiría acotar la procedencia de la grabación del fichero de audio o video, dado que, conociendo el acento del mismo, se puede asociar a cierta(s) región(es) de países. Dicha información, junto con otra procedente de otros sistemas basados en, por ejemplo, aprendizaje automático permitirían acotar aún más la región e incluso zona o lugar donde se ha grabado el material.

La presente invención puede ser también aplicada a la generación de conjuntos de datos de una manera semi-supervisada. Partiendo de un conjunto de clips de audio sin información sobre su acento, la presente invención permitiría realizar una primera clasificación de los acentos contenidos en dichos clips de audio, que luego sería revisada y corregida

manualmente, ahorrando una gran cantidad de tiempo y recursos al personal asignado a la tarea de etiquetado de datos. La disposición de grandes conjuntos de ficheros de audio permitiría el entrenamiento de sistemas de clasificación de acentos más robustos y fiables, que a su vez permitirían la obtención de clasificaciones de un mayor número de acentos dentro de un mismo idioma.

Dentro del proceso previo al procedimiento de clasificación de la invención, el sistema realiza un entrenamiento de una red neuronal convolucional. Para ello, se utiliza un conjunto de ficheros de audio de entrenamiento, y cada uno de ellos se divide en fragmentos. De cada fragmento, se extrae un espectrograma, y todos los espectrogramas extraídos de los ficheros de audio de entrenamiento se utilizan para el entrenamiento de una red neuronal convolucional.

Grad-CAM se utiliza como método de interpretabilidad, siendo capaz de destacar las diferentes partes de un espectrograma en función de la clase objetivo. Grad-CAM extrae gradientes de una capa de la red neuronal convolucional y usa esta información para indicar cuáles son las regiones del espectrograma que más contribuirán a la clasificación del acento. Por ese motivo, los mapas de características convolucionales rectificados a partir de la red neuronal convolucional entrenada se combinan para obtener a partir de un fichero de audio de entrenamiento un mapa de calor Grad-CAM. Debido a que los espectrogramas se presentan en los ejes de la frecuencia y el tiempo, esto implica que la ubicación de cada elemento es útil para almacenar la información del audio original de manera espacial, por lo que los mapas de calor generados mediante Grad-CAM almacenan parte de esa información a través de su ubicación y forma resultante. Los mapas de calor Grad-CAM se pueden considerar como un mapa de localización que resalta las regiones importantes del espectrograma para predecir la clasificación del acento. Dado un fichero de audio cualquiera, un vector representativo del mismo a través de estos mapas de calor Grad-CAM contendría información representativa sobre el acento recogido en dicho fichero de audio.

Dado que el problema de la clasificación automática del acento contenido en ficheros audio se puede abordar como un problema de aprendizaje automático supervisado, a continuación, el procedimiento de la invención extrae para cada fichero de audio un vector representativo del mismo a través de la unión de dos vectores diferentes: (i) un vector del espectrograma aplanado y (ii) un vector de mapas de calor Grad-CAM. Dichos vectores representativos del audio se utilizarán para el entrenamiento de un algoritmo de aprendizaje

supervisado, que permitirá determinar el acento de un fichero de audio desconocido.

El procedimiento de la presente invención se puede aplicar a cualquier tipo de fichero de audio, tanto descargado de la Web, como suministrado al sistema a través de un dispositivo de almacenamiento externo de cualquier tipo.

En una realización, el procedimiento y sistema automatizado para para la clasificación automática del acento de hablantes en ficheros de audio utilizando aprendizaje profundo supervisado y Grad-CAMs de la presente invención comprende las siguientes etapas:

1. Obtención de ficheros de audio. Esta obtención se puede realizar en un modo en línea, a través de un ordenador con conexión a internet, o en un modo fuera de línea, obteniendo en un ordenador los ficheros de audio a través de un dispositivo de almacenamiento externo.

2. División de los ficheros de audio. Dentro del mismo ordenador, en una realización preferente de la invención, se realiza una división del fichero de audio en fragmentos de menor duración que la del fichero de audio original.

3. Extracción de espectrogramas. Dentro del mismo ordenador, para cada fichero de audio que se pretenda resumir, se realiza la extracción de espectrogramas de cada uno de los fragmentos en los que se ha dividido previamente. En una realización, los mapas de calor Grad-CAM se calculan utilizando la última capa convolucional de la arquitectura de la red neuronal.

4. Cálculo de vector del espectrograma aplanado. Para cada uno de los espectrogramas calculados para cada fragmento de un fichero, los espectrogramas se aplanan generando un vector del espectrograma aplanado. En una realización, el aplanamiento de los espectrogramas se realiza concatenando todos los elementos de todas las filas, pero se podría emplear otras alternativas de aplanamiento diferentes (por ejemplo, por columnas). La técnica de aplanado elegida no afecta al resultado obtenido, ya que la mayoría de los modelos de aprendizaje automático, a excepción de las redes convolucionales, no procesan la información espacial de las entradas, sino únicamente su posición y contenido.

5. Codificación del vector de mapas de calor Grad-CAMs. De acuerdo con una realización de la invención, se realiza un proceso de codificación de los espectrogramas en una representación vectorial basada en mapas de calor Grad-CAMs. En una realización, se diseña una red neuronal convolucional para obtener esos mapas de calor Grad-CAMs, la cual se alimenta con los espectrogramas extraídos de fragmentos de audio obtenidos de la anterior fase. En una realización, sobre los anteriores mapas de calor Grad-CAMs se aplica análisis de componentes principales (en inglés, “Principal Component Analysis”) para poder obtener los vectores reducidos de mapas de calor Grad-CAMs. Se extrae un mapa de calor Grad-CAM por cada acento que se trate de identificar, lo cual permite extraer información de todas las posibles clases sin que sea necesario conocer de antemano cuál es la clase correcta. En una realización, una vez entrenada la red neuronal, se congelan los pesos tras el entrenamiento y se utilizan los pesos obtenidos en la última capa de la red para la obtención de los mapas de calor Grad-CAMs, a partir de los cuales se extraerán los vectores de mapas de calor Grad-CAMs.

6. Generación del vector de transferencia de gradientes. En una realización, se combinan el vector de espectrograma aplanado y el vector de mapas de calor Grad-CAMs para obtener el vector de transferencia de gradientes. Dicho vector de transferencia de gradientes será la representación del fragmento de audio original.

7. Clasificación automática del fichero de audio en base al acento, utilizando los vectores de transferencia de gradientes que alimentan un modelo de aprendizaje automático previamente entrenado.

La presente invención también se refiere a un sistema para la clasificación automática del acento en señales de audio, que comprende una unidad de procesamiento de datos (e.g. un procesador) configurada para ejecutar el método de clasificación descrito. En una realización, el sistema comprende un ordenador y unos medios de almacenamiento de datos donde se almacenan los ficheros de audio a analizar.

La presente invención también se refiere a un producto de programa que comprende medios

de instrucciones de programa para llevar a la práctica el procedimiento anteriormente descrito cuando el programa se ejecuta en un procesador. El producto de programa está preferentemente almacenado en un medio de soporte de programas. Los medios de instrucciones de programa pueden tener la forma de código fuente, código objeto, una fuente intermedia de código y código objeto, por ejemplo, como en forma parcialmente compilada, o en cualquier otra forma adecuada para uso en la puesta en práctica de los procesos según la invención.

El medio de soporte de programas puede ser cualquier entidad o dispositivo capaz de soportar el programa. Por ejemplo, el soporte podría incluir un medio de almacenamiento, como una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash, un soporte de grabación magnética, por ejemplo, un disco duro o una memoria de estado sólido (SSD). Además, los medios de instrucciones de programa almacenados en el soporte de programa pueden ser, por ejemplo, mediante una señal eléctrica u óptica que podría transportarse a través de cable eléctrico u óptico, por radio o por cualquier otro medio. Cuando el producto de programa va incorporado en una señal que puede ser transportada directamente por un cable u otro dispositivo o medio, el soporte de programa puede estar constituido por dicho cable u otro dispositivo o medio. Como variante, el soporte de programa puede ser un circuito integrado en el que va incluido el producto de programa, estando el circuito integrado adaptado para ejecutar, o para ser utilizado en la ejecución de los procesos correspondientes.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

A continuación, se describen de manera muy breve una serie de figuras que ayudan a comprender mejor la invención y que se relacionan expresamente con una realización de dicha invención que se presenta como un ejemplo no limitativo de ésta.

La Figura 1 es un diagrama de flujo que representa las etapas de un método para la clasificación automática del acento en señales de audio según la presente invención.

Las Figuras 2A y 2B muestran diferentes realizaciones del método para la clasificación automática del acento en señales de audio.

La Figura 3 ilustra la obtención, a partir de un fichero de audio, de fragmentos de audio que son clasificados por el método de la presente invención.

La Figura 4 muestra, de acuerdo a una realización, un sistema empleado para clasificar el acento de señales de audio contenidas en ficheros de audio.

La Figura 5 representa el entrenamiento de la red neuronal empleada en la presente invención para obtener los mapas de calor Grad-CAM.

Las Figuras 6 y 7 muestran etapas del proceso de entrenamiento del modelo de clasificación empleado en la presente invención para clasificar los acentos de las señales de audio.

REALIZACIÓN PREFERENTE DE LA INVENCIÓN

La **Figura 1** muestra un método (100) para la clasificación automática del acento en señales de audio, esto es, el acento de los hablantes durante una locución incluida en una señal de audio.

El método (100) comprende las siguientes etapas:

- Recibir (110) una señal de audio (3).
- Obtener (120) un espectrograma de la señal de audio (3), esto es, la distribución de energía de la señal de audio (3) representada en tres dimensiones (tiempo, frecuencia y amplitud).
- Obtener (130), utilizando una red neuronal alimentada por el espectrograma y que ha sido previamente entrenada para clasificar una pluralidad a de acentos, un mapa de calor Grad-CAM (6) por cada uno de los acentos a para los que la red neuronal ha sido entrenada. En una realización, la red neuronal es una red neuronal convolucional, y los mapas de calor Grad-CAMs se extraen preferentemente de la última capa convolucional de la red neuronal.
- Reducir (140) la dimensionalidad de cada mapa de calor Grad-CAM para obtener un respectivo tensor reducido.
- Obtener (150) una clasificación del acento (11) de la señal de audio utilizando un modelo de aprendizaje automático entrenado, alimentado por el espectrograma y los tensores reducidos.

El modelo de aprendizaje automático ha sido previamente entrenado para clasificación de

acentos (en concreto, para la clasificación de la pluralidad a de acentos considerados) utilizando un algoritmo de aprendizaje supervisado. El modelo de aprendizaje automático es un clasificador que puede estar basado en un modelo de aprendizaje automático de cualquier tipo (basado en redes neuronales, Máquinas de Vectores de Soporte, etc.).

El modelo de aprendizaje automático entrenado para clasificación de acentos es un clasificador de acentos que puede obtener una puntuación para cada uno de los acentos a para los que ha sido entrenado. La clasificación del acento (11) es la salida de este clasificador y puede incluir, por ejemplo, las diferentes puntuaciones o probabilidades obtenidas para cada uno de los acentos a, o el acento concreto que tiene la mayor puntuación o probabilidad.

En la **Figura 2A** se muestra un esquema simplificado de una realización particular del método (100) para la clasificación automática del acento en señales de audio. El modelo de aprendizaje automático entrenado (16) está alimentado por dos fuentes diferentes: el espectrograma (4) y los tensores reducidos (8). El modelo de aprendizaje automático puede ser, por ejemplo, un modelo de clasificación.

Con respecto a la primera fuente que alimenta el modelo de aprendizaje automático entrenado (16), i.e. el espectrograma (4), el método (100) puede comprender obtener un tensor de espectrograma (5) a partir de la información contenida en el espectrograma (4) (esto es, representar la información contenida en el espectrograma (4) mediante un tensor de espectrograma (5)), siendo este tensor de espectrograma (5) el que alimentará el modelo de aprendizaje automático entrenado (16). El tensor de espectrograma (5) es una representación tensorial (e.g. mediante una matriz o un vector) de la información contenida en el espectrograma (4). En la realización mostrada en la Figura 2A el tensor de espectrograma (5) es un vector de espectrograma obtenido mediante aplanamiento del espectrograma (4), en concreto mediante la concatenación de los elementos del espectrograma (4) (esto es, los valores de amplitud de la distribución de energía o los valores de los píxeles del espectrograma (4)). La concatenación de los elementos del espectrograma (4) se puede realizar por filas o por columnas, unas a continuación de otras. En el ejemplo de la Figura 2A el espectrograma (4) queda representado mediante un tensor de una sola dimensión (o vector columna).

Con respecto a la segunda fuente que alimenta el modelo de aprendizaje automático

entrenado (16), i.e. los tensores reducidos (8), éstos son obtenidos a partir de cada mapa de calor Grad-CAM (6) utilizando un método cualquiera de reducción de dimensionalidad, como por ejemplo mediante un análisis de componentes principales (7). Los tensores reducidos (8) pueden representarse mediante una matriz o un vector; en el ejemplo mostrado en la Figura 2A cada tensor reducido (8) es un tensor de una sola dimensión (o vector columna). El método (100) puede comprender calcular un tensor reducido de mapas de calor Grad-CAM (8') a partir de los tensores reducidos (8), siendo este tensor reducido de mapas de calor Grad-CAM (8') el que alimentará el modelo de aprendizaje automático entrenado (16). En el ejemplo de la Figura 2A cada tensor reducido (8) es un tensor de una sola dimensión (o vector columna), y el tensor reducido de mapas de calor Grad-CAM (8') se obtiene mediante la agrupación de los diferentes tensores reducidos (8), por ejemplo concatenando los tensores reducidos (8) en diferentes columnas (formando una matriz) o en un único tensor de una sola dimensión (o vector columna, como se muestra en la Figura 2A).

El método (100) puede comprender obtener, mediante la combinación (e.g. concatenación) del tensor de espectrograma (5) y del tensor reducido de mapas de calor Grad-CAM (8'), un tensor de transferencia de gradientes (9) que alimenta al modelo de aprendizaje automático entrenado (16). El tensor de transferencia de gradientes (9) puede ser, por ejemplo, un tensor de una o dos dimensiones (un vector o matriz, respectivamente). De acuerdo a la realización mostrada en la Figura 2A, el tensor de transferencia de gradientes (9) es un tensor de una sola dimensión (o vector columna) formado mediante la concatenación de dos tensores de una sola dimensión (o vectores columna): el vector de espectrograma y el vector reducido de mapas de calor Grad-CAM.

Alternativamente a la realización de la Figura 2A, el modelo de aprendizaje automático entrenado (16) podría estar alimentado (esto es, recibir como entrada) directamente por el espectrograma (4) (o un tensor de espectrograma (5), en el caso de que se calcule éste) y por los tensores reducidos (8) (o un tensor reducido de mapas de calor Grad-CAM (8'), en el caso de que se calcule éste), como entradas separadas sin combinar. La **Figura 2B** representa una realización en la que el modelo de aprendizaje automático entrenado (16) recibe el espectrograma (4) y los tensores reducidos (8) directamente como entradas.

La señal de audio (3) se puede obtener de diferentes fuentes; por ejemplo, puede estar contenida en un fichero de audio en cualquier formato con o sin compresión (WAV, FLAC, MP3, etc.), una señal recibida en tiempo real en streaming o una señal capturada mediante

un micrófono.

La **Figura 3** representa la obtención de varias señales de audio (3) a partir de un fichero de audio (1). El método (100) es implementado por una unidad de procesamiento de datos, como por ejemplo un ordenador (2). En una realización, el método (100) puede comprender recibir un fichero de audio (1), dividir (302) el fichero de audio (1) (esto es, dividir la señal de audio contenida en el fichero de audio) en una pluralidad de señales de audio (3) de una determinada longitud (e.g. 4 segundos), y obtener la clasificación del acento (11) para cada señal de audio (3) utilizando el procedimiento de clasificación descrito anteriormente. El ordenador (2) puede obtener los ficheros de audio (1) almacenados remotamente (utilizando por ejemplo una conexión a Internet o una conexión a una red local) o localmente (e.g. en un disco duro, en un pendrive USB, en un CD-ROM, etc.).

Al trocear el fichero de audio (1) en una pluralidad de señales de audio (3) de una longitud reducida determinada, se permite por un lado el tratamiento de la señal de audio con un tamaño manejable y homogéneo de los datos de entrada (cuanto mayor sea la duración de la señal de audio (3), mayor será el tamaño del espectrograma (4) asociado, y habrá por tanto una mayor complejidad computacional en los subsiguientes procesos), y por otro lado se permite separar las locuciones de diferentes hablantes y clasificar sus respectivos acentos, lo cual es especialmente útil para clasificar acentos cuando el fichero de audio (1) contenga locuciones de hablantes con diferentes acentos. En el ejemplo de la Figura 3 las dos primeras señales de audio (3) del fichero de audio (1) se clasifican como “Acento 2” dentro de la pluralidad a de acentos (e.g. acento andaluz), mientras que la tercera señal de audio (3) se clasifica como “Acento 1” (e.g. acento gallego), lo cual implicaría que en los dos primeros tercios del fichero de audio (1) el hablante tiene acento andaluz y en el último tercio el hablante tiene acento gallego.

La **Figura 4** representa un sistema (400) para la clasificación automática del acento en señales de audio, que comprende una unidad de procesamiento de datos (e.g. un procesador (410), un ordenador, y en general cualquier dispositivo con suficiente capacidad de cómputo) configurada para ejecutar las etapas del método (100). En una realización, la unidad de procesamiento de datos recibe un fichero de audio (1) y clasifica los acentos de los hablantes contenidos en el mismo, obteniendo un fichero de audio clasificado (17). El sistema (400) puede comprender unos medios de almacenamiento de datos (e.g. una memoria (420), un disco duro o en general cualquier unidad con capacidad de

almacenamiento) configurados para almacenar unos datos de entrada, unos datos de procesamiento y/o unos datos de salida. Por ejemplo, la memoria (420) puede almacenar el fichero de audio (1) y el fichero de audio clasificado (17) generado. Por tanto, el sistema (400) clasifica automáticamente el fichero de audio (1) según el acento contenido.

Para obtener el mapa de calor Grad-CAM (6) para un acento determinado, se supone que la puntuación o probabilidad de la salida de la red neuronal es 1 para ese acento y 0 para el resto de acentos. Para poder obtener los mapas de calor Grad-CAM (6), es necesario haber realizado un entrenamiento previo de una red neuronal (18) (por ejemplo, una red neuronal convolucional, CNN) para obtener una red neuronal (15) entrenada para clasificar una pluralidad a de acentos, cuyo proceso se describe en la **Figura 5**. Para el entrenamiento se emplea un conjunto de ficheros de audio de entrenamiento (12) que contienen instancias de una pluralidad a de acentos, que pueden ser por ejemplo recibidos por el ordenador (2) a través de una conexión a Internet. Cada uno de los ficheros de audio de entrenamiento (12) se divide en fragmentos de audio (señales de audio de entrenamiento (13)), de los cuales se extraen un conjunto de espectrogramas de entrenamiento (14) de una determinada duración. Dichos espectrogramas de entrenamiento (14) se utilizan para el entrenamiento de la red neuronal (18) (e.g. CNN), obteniendo una red neuronal entrenada (15) (e.g. CNN entrenada).

Igualmente, para poder obtener la clasificación del acento (11), es necesario haber realizado un entrenamiento previo de un modelo de aprendizaje automático para obtener el modelo de aprendizaje automático entrenado (16). Para ello se pueden emplear los mismos ficheros de audio de entrenamiento (12) y sus espectrogramas de entrenamiento (14) asociados utilizados en el entrenamiento de la red neuronal (18), pero en este caso alimentando la red neuronal entrenada (15) previamente, tal y como se muestra en la **Figura 6**, obteniendo unos mapas de calor Grad-CAM de entrenamiento (26) y, a partir de ellos, los correspondientes tensores reducidos de entrenamiento (28) que se emplearán en este segundo entrenamiento. La **Figura 7** muestra un esquema simplificado del entrenamiento de un modelo de aprendizaje automático (10) (o modelo de clasificación de acentos) utilizando un conjunto de tensores de transferencia de gradientes de entrenamiento (19) calculados a partir de los tensores reducidos de entrenamiento (28) y de los espectrogramas de entrenamiento (14), obteniendo de esta forma el modelo de aprendizaje automático entrenado (16) (o modelo de clasificación de acentos entrenado). Los tensores de transferencia de gradientes de entrenamiento (19) pueden estar expresados, por ejemplo,

en forma de vector (vector fila, vector columna) o en formato matricial (tensor de dos dimensiones).

Para obtener el fichero de audio (1) cuyo acento se quiere clasificar automáticamente, o los ficheros de audio de entrenamiento (12) que se pretenden utilizar para el entrenamiento de la red neuronal (18) y del modelo de clasificación de acentos (10), el ordenador (2) puede estar conectado a Internet a través de una conexión inalámbrica o a través de un cable de red Ethernet. El fichero de audio (1) o los ficheros de audio (12) se pueden conseguir a través de un medio de soporte, que puede ser cualquier entidad o dispositivo capaz de almacenar ficheros de audio. Por ejemplo, el soporte podría incluir un medio de almacenamiento, como una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash USB, SD, mini-SD o micro-SD, un soporte de grabación magnética, por ejemplo, un disco duro o una memoria de estado sólido (SSD). El objeto de esta conexión y configuración a la red, o de la disponibilidad de soportes de medio de cualquier tipo, es la obtención del fichero de audio (1) que se va a clasificar automáticamente en base al acento de su hablante utilizando aprendizaje profundo supervisado y Grad-CAMs de la presente invención, y la obtención de los ficheros de audio (12) que se van a utilizar para el entrenamiento de la red neuronal (18), lo cual es necesario para la obtención de los mapas de calor Grad-CAMs (6), los tensores reducidos (8) y, opcionalmente, los tensores reducidos de mapas de calor Grad-CAM (8') (e.g. vectores de mapas de calor Grad-CAMs).

En una realización, para el entrenamiento de la red neuronal (18) se procede inicialmente a realizar la división de ficheros de audio de entrenamiento (12) $\{A_{t1}, A_{t2}, \ldots A_{tn}\}$que se pretenden utilizar para el entrenamiento en fragmentos dentro del ordenador (2). De modo que un fichero de audio de entrenamiento A_{tn} quedará dividido en *m*fragmentos o señales de audio de entrenamiento (13) $\{F_{t1}, F_{t2}, \ldots F_{tm}\}$. Por ejemplo, cada fichero de audio A_{tn} de entrada se puede dividir en fragmentos de 4 segundos, de manera similar a como se describe en Zeng et al. *(Zeng, Y., Mao, H., Peng, D., & Yi, Z. "Spectrogram based multi-task audio classification. Multimedia Tools and Applications", vol. 78, no 3, p. 3705-3722, 2019)*, con un software de edición de audio, o automáticamente a través de una librería de un lenguaje de programación especializada en el manejo de audio. A continuación, se realiza la extracción de espectrogramas de entrenamiento (14) $\{E_{t1}, E_{t2}, \ldots E_{tm}\}$ de cada una de las señales de audio de entrenamiento (13) $\{F_{t1}, F_{t1}, \ldots F_{tm}\}$. En una realización, a cada

fragmento de audio se le calcula la Transformada de Fourier de corta duración, posteriormente se aplica la función valor absoluto en todos sus elementos, obteniendo los espectrogramas en escala de amplitud, y los cuales se convierten en espectrogramas en escala de decibelios. Los espectrogramas extraídos E_{tm} son utilizados para el entrenamiento de una red neuronal (18) convolucional desde cero, es decir, sin entrenamiento previo.

En una realización, se utiliza la última capa convolucional de la red neuronal convolucional entrenada R_C para la extracción de mapas de calor Grad-CAMs, ya que la última capa almacena la mayor cantidad de información semántica de alto nivel e información espacial detallada (R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017).

En una realización, para la red neuronal (18) se utiliza la arquitectura DenseNet201 (F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, “Densenet: Implementing efficient convnet descriptor pyramids,” arXiv preprint arXiv:1404.1869, 2014). Si bien se puede utilizar cualquier arquitectura de redes neuronales, DenseNet aporta la ventaja de estar altamente interconectada entre las capas de sus distintos niveles, lo cual permite una alta capacidad de modelado de detalles finos, como pueden ser las frecuencias altas, y detalles de alto nivel, como pueden ser las frecuencias bajas, en un espectrograma. Dicha capacidad de modelado se refleja en que DenseNet permite obtener mapas de calor de Grad-CAM con un alto grado de detalle.

En una realización, se han utilizado los siguientes hiperparámetros para el entrenamiento de DenseNet: un tamaño de lote (en inglés, “batch size”) de 22 espectrogramas, una tasa de aprendizaje (en inglés, “learning rate”) inicial de 0.001, el cual decae en un 90% cada 25 epoch y un total de 100 epoch de entrenamiento. Se pueden utilizar, por ejemplo, los hiperparámetros propuestos por los autores del conjunto de datos (en inglés, “dataset”) VoxCeleb (A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Largescale speaker verification in the wild,” Computer Speech & Language, vol. 60, p. 101027, 2020), convirtiendo todos los audios a flujos de 16 bits de un solo canal a 16 kHz de tasa de muestreo. En una realización, los espectrogramas son generados por una ventana Hamming de *25 ms* de ancho y *10 ms* de paso, dichos valores están específicamente seleccionados para reflejar correctamente las particularidades de la voz humana, resultando en

espectrogramas de dimensiones $512\,x\,400$ píxeles para cada fragmento de audio. En una realización preferente de la invención, se redimensionan los espectrogramas a $128\,x\,100$ píxeles, generando una ventaja técnica de reducción de esfuerzo computacional requerido para su procesado. En una realización, no se eliminan los silencios de los espectrogramas para generar modelos robustos a este tipo de eventos.

En una realización, el entrenamiento de la red neuronal (18) comprende adquirir ficheros de audio de entrenamiento (12) (e.g. a través de un ordenador (2) conectado a Internet o desde un medio de almacenamiento externo), dividir los ficheros de audio de entrenamiento (12) en señales de audio de entrenamiento (13), extraer los espectrogramas de entrenamiento (14) de cada uno de los fragmentos de audio anteriores según el proceso anteriormente descrito, y entrenar una arquitectura DenseNet con los espectrogramas de entrenamiento (14) extraídos de las señales de audio de entrenamiento (13) para que pueda clasificar automáticamente un número a de acentos.

En una realización, se utiliza el conjunto de datos Voice Cloning Toolkit (VCTK) para el entrenamiento de la red neuronal (18) R_C (Yamagishi, Junichi; Veaux, Christophe; MacDonald, Kirsten. (2019). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR)).

A continuación, se explica cómo se clasifica automáticamente un fichero de audio A en base al acento del hablante que podría contener utilizando aprendizaje supervisado profundo supervisado y Grad-CAMs, empleando la red neuronal convolucional entrenada R_C, de acuerdo a una posible realización.

Se procede a realizar la división del fichero de audio A en fragmentos $\{F_1, F_2, \ldots F_p\}$ dentro del ordenador (2) para cada fichero de audio A en el que se pretenda clasificar automáticamente el acento del hablante que contiene. En una realización, cada fichero de audio A de entrada se divide en fragmentos $\{F_1, F_2, \ldots F_n\}$ de 4 segundos cada uno, al igual que en el trabajo de Zeng et al. *(Zeng, Y., Mao, H., Peng, D., & Yi, Z. "Spectrogram based multi-task audio classification. Multimedia Tools and Applications", vol. 78, no 3, p. 3705-3722, 2019)*, con un software de edición de audio, o automáticamente a través de una librería de un lenguaje de programación especializada en el manejo de audio. A

continuación, se realiza la extracción de espectrogramas $\{E_1, E_2, \ldots E_p\}$ de cada uno de los fragmentos de audio $\{F_1, F_2, \ldots F_p\}$. En una realización, se utiliza la librería de Python *librosa*. Los espectrogramas extraídos $\{E_1, E_2, \ldots E_n\}$ son utilizados para la generación del vector de espectrograma aplanado V_a (tensor de espectrograma (5)) y del vector de mapas de calor Grad-CAM V_G (tensor reducido de mapas de calor Grad-CAM (8')).

En una realización, se utilizan los hiperparámetros propuestos por los autores del conjunto de datos VoxCeleb (A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Largescale speaker verification in the wild," Computer Speech & Language, vol. 60, p. 101027, 2020), convirtiendo el fichero de audio a flujos de 16 bits de un solo canal a 16 kHz de tasa de muestreo. En una realización, los espectrogramas son generados por una ventana Hamming de 25 ms de ancho y 10 ms de paso, resultando en espectrogramas de dimensiones $512 x 400$ píxeles para cada fragmento de audio, y se redimensionan los espectrogramas a $128 x 100$ píxeles, generando una ventaja técnica de reducción de esfuerzo computacional requerido para su procesado. En una realización, no se eliminan los silencios de los espectrogramas.

Cada espectrograma E_n de cada fragmento de audio F_n se puede utilizar para generar un vector de espectrograma aplanado V_a y un vector de mapas de calor Grad-CAM V_G. En una realización, el vector de espectrograma aplanado V_a se calcula concatenando los valores de los píxeles del espectrograma E_n por filas, unas a continuación de otras, resultando en un vector de 12800 dimensiones. El vector de mapas de calor Grad-CAM V_G se calcula a partir de los mapas de calor Grad-CAM M_G que se obtienen tras alimentar la red neuronal convolucional entrenada R_C con el espectrograma E_n. En una realización, los mapas de calor Grad-CAMs M_G se extraen de la última capa convolucional de la red neuronal convolucional entrenada R_C. En una realización, se extrae un mapa de calor Grad-CAM M_G por cada uno de los acentos a en los que se haya entrenado la red neuronal convolucional R_C. Por este motivo, si la red se entrena para detectar a acentos, se extraerán un total de a mapas de calor Grad-CAM M_G. A continuación, se aplica análisis de componentes principales sobre cada mapa de calor, con objeto de obtener un vector de mapas de calor Grad-CAM V_G. En una realización, se utilizan 37 componentes principales $\{P_1, P_2, \ldots P_{37}\}$ para representar cada mapa de calor Grad-CAM M_G. En una realización, el vector de mapas

de calor Grad-CAM V_G tiene una dimensión de $37 * a$, siendo a el número de acentos para los que la red convolucional R_C ha sido entrenada.

El motivo de utilizar 37componentes fundamentales aporta una ventaja técnica a la invención, y es la retención del 95% de la varianza de la información extraída de los mapas de calor Grad-CAM M_G. Para determinar cuál es el máximo número de componentes principales que sería adecuado seleccionar, inicialmente se seleccionaron 10000 componentes principales $\{P_1, P_2, \ldots P_{10000}\}$para cada mapa de calor Grad-CAM M_G del conjunto de entrenamiento. Una vez obtenidas las 10000 componentes principales de todos los mapas de calor Grad-CAM de todos los fragmentos de los ficheros de audio de entrenamiento, se calcula un vector único de 10000 elementos, que contiene la varianza en cada posición de todas las componentes principales anteriores. Este vector de varianzas se utiliza para determinar cuántos componentes son necesarios para representar la mayor parte de la varianza de todos los Grad-CAM. A continuación, el vector de varianzas se normaliza, y dichos valores normalizados se ordenan de mayor a menor, encontrándose que las 37primeras componentes son capaces de condensar el 95% de la varianza, ya que su suma es mayor o igual a 0.95. Sería posible utilizar más componentes principales para retener un % de varianza mayor, pero aumentaría el tamaño del vector representativo, y con ello la complejidad de la presente invención. El número de componentes utilizados ha sido calculado para el conjunto de datos particular de la realización preferente, con lo que se debería realizar un análisis análogo para otros conjuntos de datos.

La utilización de Grad-CAM como método de extracción de características permite generar un mapa de calor diferenciable por cada clase a partir de un audio. Esto es debido a que los espectrogramas a partir de los cuales se generan están representados en las dimensiones de la frecuencia y el tiempo, por lo que la ubicación espacial de cada elemento tiene relevancia con respecto a su contenido, a diferencia de una imagen común, en la cual la posición de un objeto no implica un cambio del objeto en sí mismo.

Al reducir los mapas de calor diferenciables en vectores de tamaño $37 * a$ se puede resumir información relevante de un audio, con respecto a una determinada tarea, como la identificación de acentos, utilizando pocos datos. El vector de mapas de calor Grad-CAM resultante puede ser utilizado para complementar la información del espectrograma original.

Una vez obtenido los vectores de espectrograma aplanado V_a y el vector de mapas de calor Grad-CAM V_G se calcula un vector de transferencia de gradientes V_T (tensor de transferencia de gradientes (9)). En una realización, dicho vector se obtiene concatenando el vector de espectrograma aplanado V_a y el vector de mapas de calor Grad-CAM V_G. En una realización preferente de la invención, dicho vector tendría una dimensión de $12800+37*a$.

$$dim(V_T)=dim(V_a)+dim(V_G)=12800+37*a$$

El vector de transferencia de gradientes V_T obtenido para un fichero de audio de entrada A es utilizado con el modelo de clasificación automática de acentos y permite la clasificación automática del acento de los posibles hablantes contenidos en el fichero de audio A. En una realización, se ha utilizado para el entrenamiento del modelo de clasificación de audios las Máquinas de Vectores de Soporte (en inglés, “Support Vector Machines”, SVM), las cuales son altamente efectivas cuando se trabaja con entradas que poseen un gran número de dimensiones, y en los casos en que ese número de dimensiones es mayor al número de ejemplos de entrenamiento (C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995), pero podría utilizarse cualquier otro tipo de algoritmo de aprendizaje supervisado. En una realización, se han utilizado Máquinas de Vectores de Soporte con un método de escala Power Transformer de Yeo- Johnson (S. Weisberg, “Yeo-johnson power transformations,” Department of Applied Statistics, University of Minnesota. Retrieved June, vol. 1, p. 2003,2001), un kernel sigmoide y un valor de regularización $C=1$.

Se realizaron pruebas experimentales sobre el conjunto de datos VCTK (el cual contiene 88328 clips de audio de 110 hablantes nativos de inglés con múltiples acentos), y se obtuvo un aumento de la precisión en los resultados obtenidos por diferentes clasificadores (Máquinas de Vectores de Soporte, clasificadores gaussiano Naïve Bayes, clasificadores Pasivo-Agresivo) entrenados, según la presente invención, usando la concatenación del tensor de espectrograma (5) y el tensor reducido de mapas de calor Grad-CAM (8’), de hasta un $7{,}46\%$, en comparación con dichos clasificadores entrenados únicamente con espectrogramas de acuerdo al estado del arte.

REIVINDICACIONES

1. Un método para la clasificación automática del acento en señales de audio, caracterizado por que comprende:

recibir (110) una señal de audio (3);

obtener (120) un espectrograma (4) de la señal de audio (3);

obtener (130), utilizando una red neuronal (15) alimentada por el espectrograma (4) y entrenada para clasificar una pluralidad de acentos, un mapa de calor Grad-CAM (6) por cada uno de los acentos para los que la red neuronal (15) ha sido entrenada;

reducir (140) la dimensionalidad de cada mapa de calor Grad-CAM (6) para obtener un respectivo tensor reducido (8); y

obtener (150) una clasificación del acento (11) de la señal de audio (3) utilizando un modelo de aprendizaje automático entrenado (16), alimentado por el espectrograma (4) y los tensores reducidos (8).

2. El método según la reivindicación 1, que comprende obtener un tensor de espectrograma (5) a partir de la información contenida en el espectrograma (4).

3. El método según la reivindicación 2, donde el tensor de espectrograma (5) es un vector de espectrograma obtenido mediante la concatenación de los valores de amplitud de la distribución de energía del espectrograma (4).

4. El método según cualquiera de las reivindicaciones anteriores, que comprende calcular un tensor reducido de mapas de calor Grad-CAM (8') a partir de los tensores reducidos (8).

5. El método según las reivindicaciones 2 y 4, que comprende obtener, mediante la combinación del tensor de espectrograma (5) y del tensor reducido de mapas de calor Grad-CAM (8'), un tensor de transferencia de gradientes (9) que alimenta al modelo de aprendizaje automático entrenado (16).

6. El método según cualquiera de las reivindicaciones anteriores, donde reducir la dimensionalidad de cada mapa de calor Grad-CAM (6) comprende realizar un análisis de

componentes principales sobre cada mapa de calor Grad-CAM (6).

7. El método según cualquiera de las reivindicaciones anteriores, donde la red neuronal (15) es una red neuronal convolucional, y donde los mapas de calor Grad-CAMs (6) se extraen de la última capa convolucional de la red neuronal (15).

8. El método según cualquiera de las reivindicaciones anteriores, que comprende:

recibir un fichero de audio (1),

dividir el fichero de audio (1) en una pluralidad de señales de audio (3) de una determinada longitud, y

obtener la clasificación del acento (11) para cada señal de audio (3).

9. Un sistema para la clasificación automática del acento en señales de audio, caracterizado por que comprende una unidad de procesamiento de datos configurada para:

recibir una señal de audio (3);

obtener un espectrograma (4) de la señal de audio (1);

obtener, utilizando una red neuronal (15) alimentada por el espectrograma (4) y entrenada para clasificar una pluralidad de acentos, un mapa de calor Grad-CAM (6) por cada uno de los acentos para los que la red neuronal (15) ha sido entrenada;

reducir la dimensionalidad de cada mapa de calor Grad-CAM (6) para obtener un respectivo tensor reducido (8); y

obtener una clasificación del acento (11) de la señal de audio (3) utilizando un modelo de aprendizaje automático entrenado (16), alimentado por el espectrograma (4) y los tensores reducidos (8).

10. El sistema según la reivindicación 9, donde la unidad de procesamiento de datos está configurada para obtener un tensor de espectrograma (5) a partir de la información contenida en el espectrograma (4).

11. El sistema según la reivindicación 10, donde la unidad de procesamiento de datos está configurada para obtener un tensor de espectrograma (5) mediante la concatenación de los valores de amplitud de la distribución de energía del espectrograma (4).

12. El sistema según cualquiera de las reivindicaciones 9 a 11, donde la unidad de procesamiento de datos está configurada para calcular un tensor reducido de mapas de calor Grad-CAM (8') a partir de los tensores reducidos (8).

13. El sistema según las reivindicaciones 10 y 12, donde la unidad de procesamiento de datos está configurada para obtener, mediante la combinación del tensor de espectrograma (5) y del tensor reducido de mapas de calor Grad-CAM (8'), un tensor de transferencia de gradientes (9) que alimenta al modelo de aprendizaje automático entrenado (16).

14. El sistema según cualquiera de las reivindicaciones 9 a 13, donde la unidad de procesamiento de datos está configurada para reducir la dimensionalidad de cada mapa de calor Grad-CAM (6) mediante la aplicación de un análisis de componentes principales sobre cada mapa de calor Grad-CAM (6).

15. El sistema según cualquiera de las reivindicaciones 9 a 14, donde la red neuronal (15) es una red neuronal convolucional, y donde la unidad de procesamiento de datos está configurada para extraer los mapas de calor Grad-CAMs (6) de la última capa convolucional de la red neuronal (15).

16. El sistema según cualquiera de las reivindicaciones 9 a 15, donde la unidad de procesamiento de datos está configurada para:

recibir un fichero de audio (1),

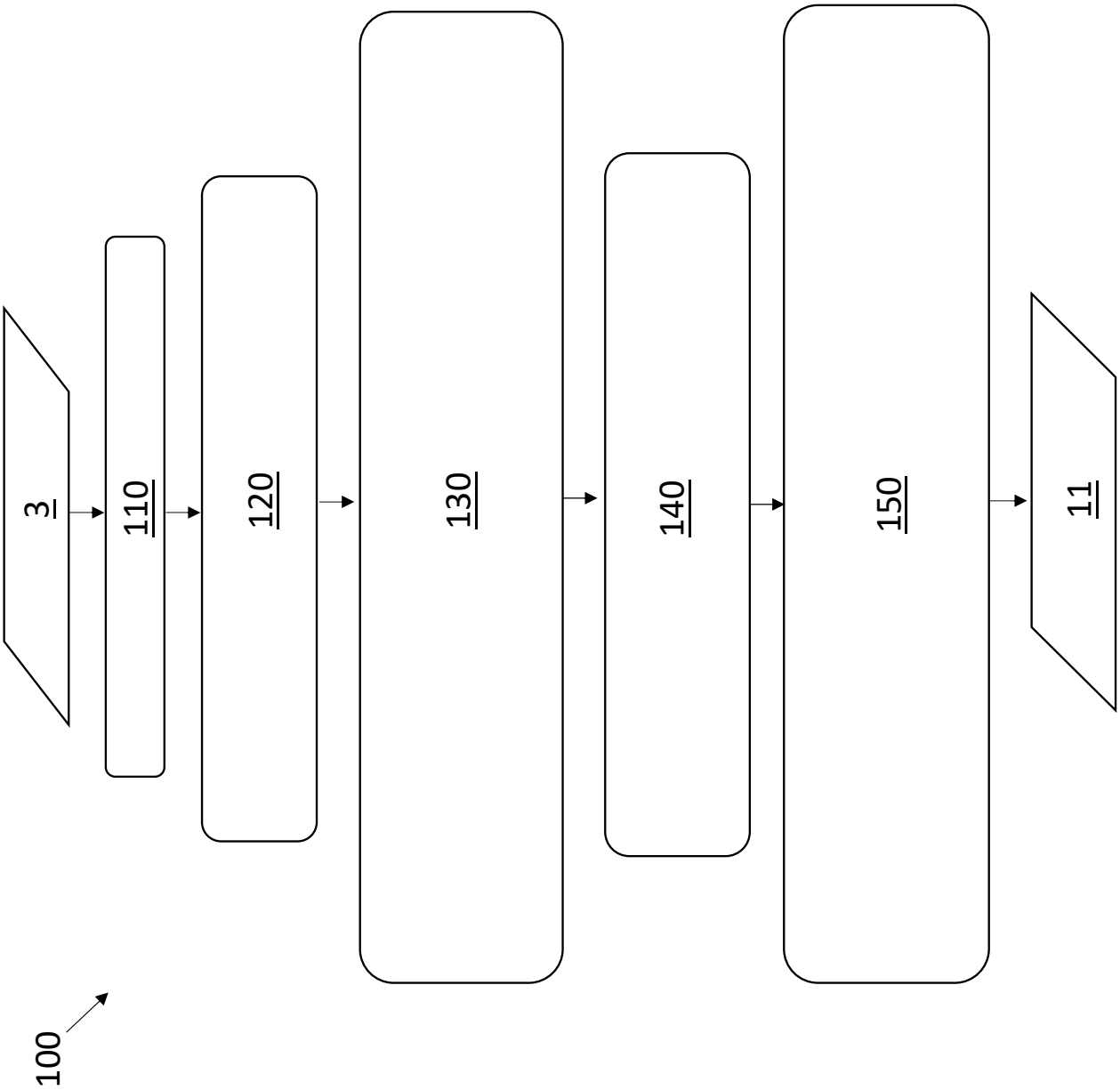
dividir el fichero de audio (1) en una pluralidad de señales de audio (3) de una determinada longitud, y

obtener la clasificación del acento (11) para cada señal de audio (3).

17. Un producto de programa que comprende medios de instrucciones de programa para llevar a cabo el método definido en cualquiera de las reivindicaciones 1 a 8 cuando el programa se ejecuta en un procesador.

18. Un medio de soporte de programas, que almacena el producto de programa según la reivindicación 17.

FIG. 1



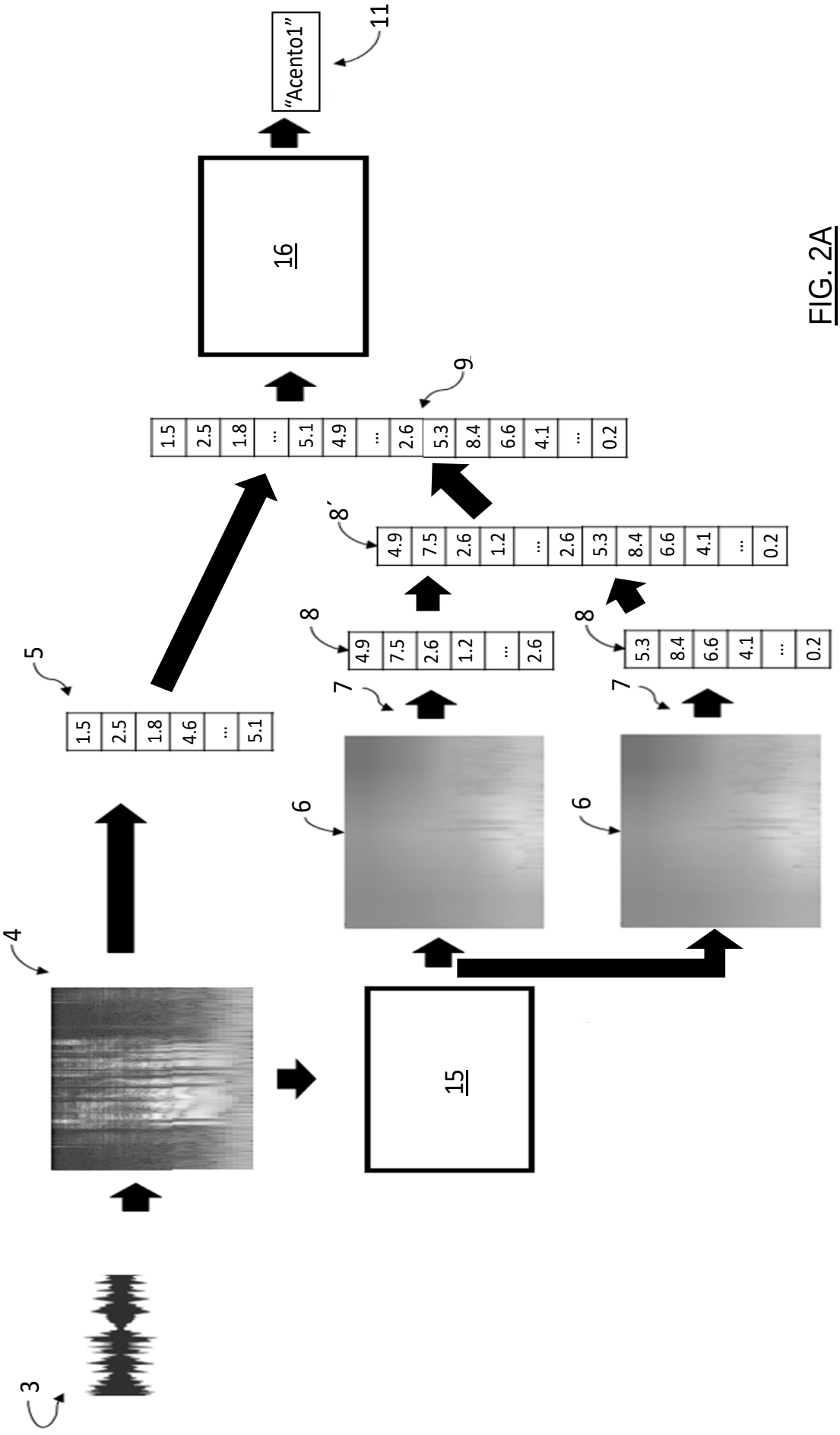

"Acento1"
11
16
9
1.5 2.5 1.8 ... 5.1 4.9 ... 2.6 5.3 8.4 6.6 4.1 ... 0.2
8´
4.9 7.5 2.6 1.2 ... 2.6 5.3 8.4 6.6 4.1 ... 0.2
8
4.9 7.5 2.6 1.2 ... 2.6
8
5.3 8.4 6.6 4.1 ... 0.2
5
1.5 2.5 1.8 4.6 ... 5.1
7
7
6
6
4
15
3

FIG. 2A

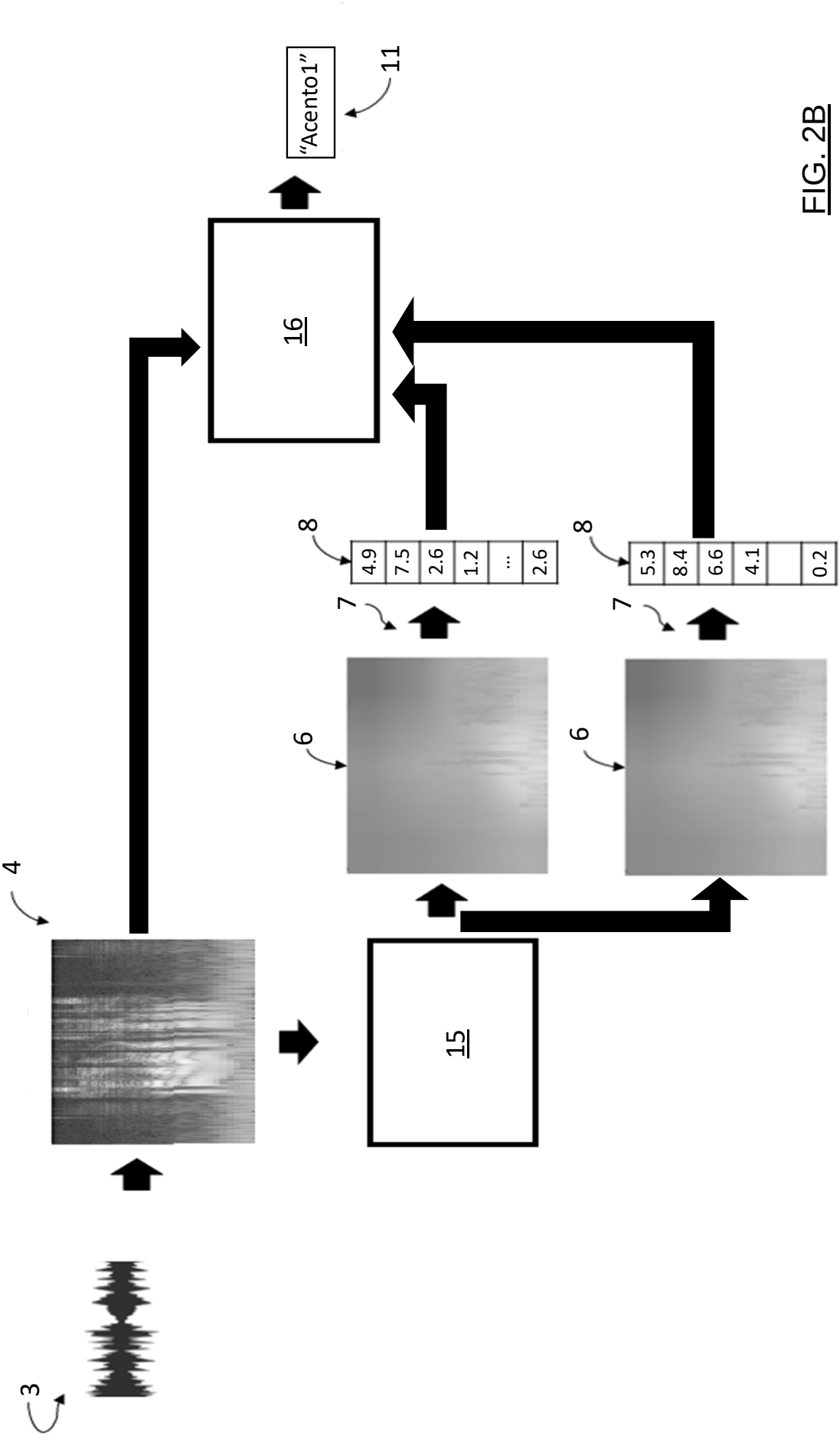


FIG. 2B

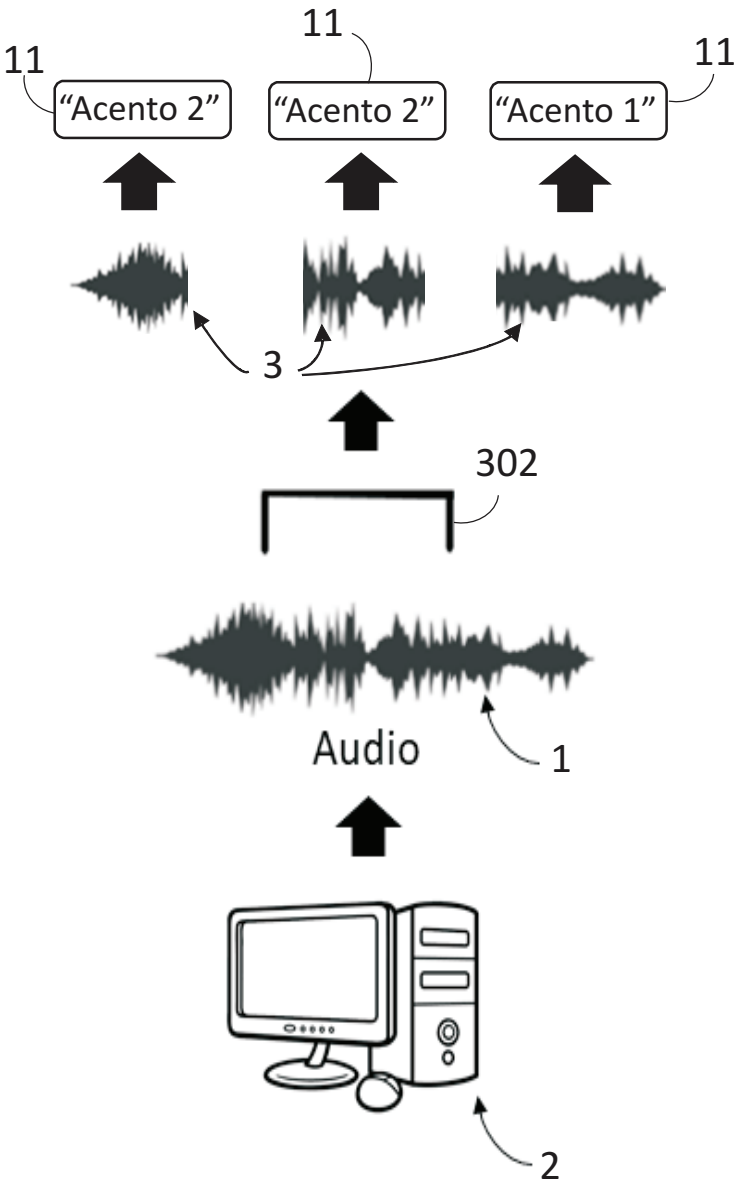


FIG. 3

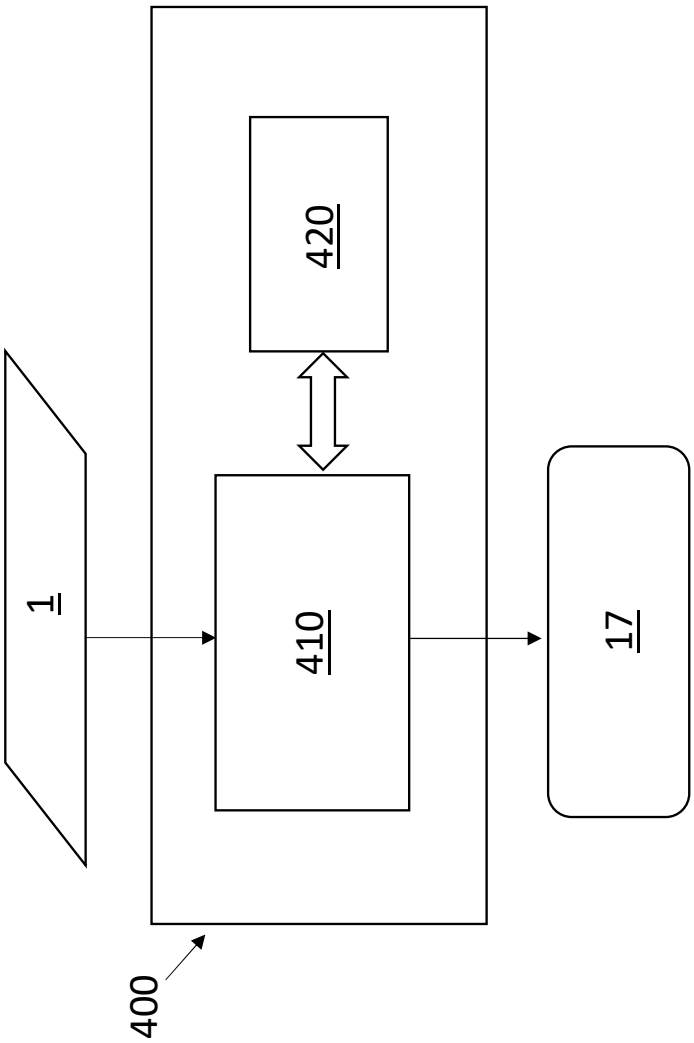


FIG. 4

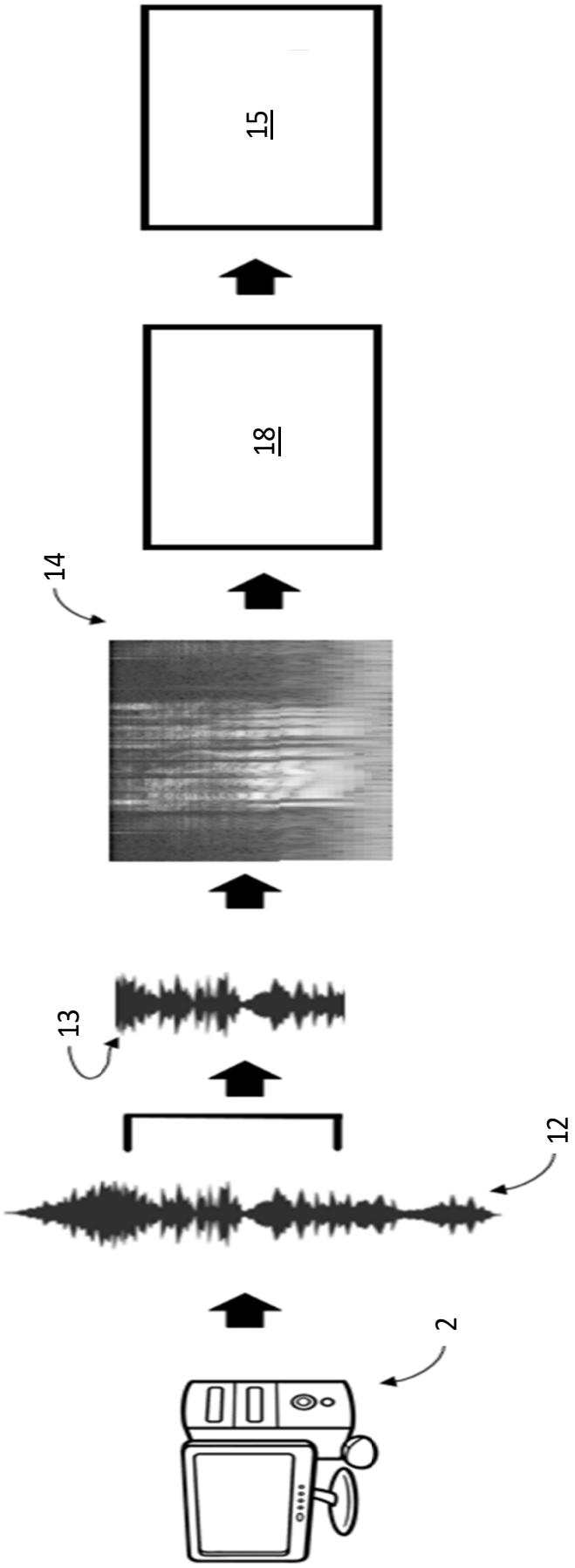


FIG. 5

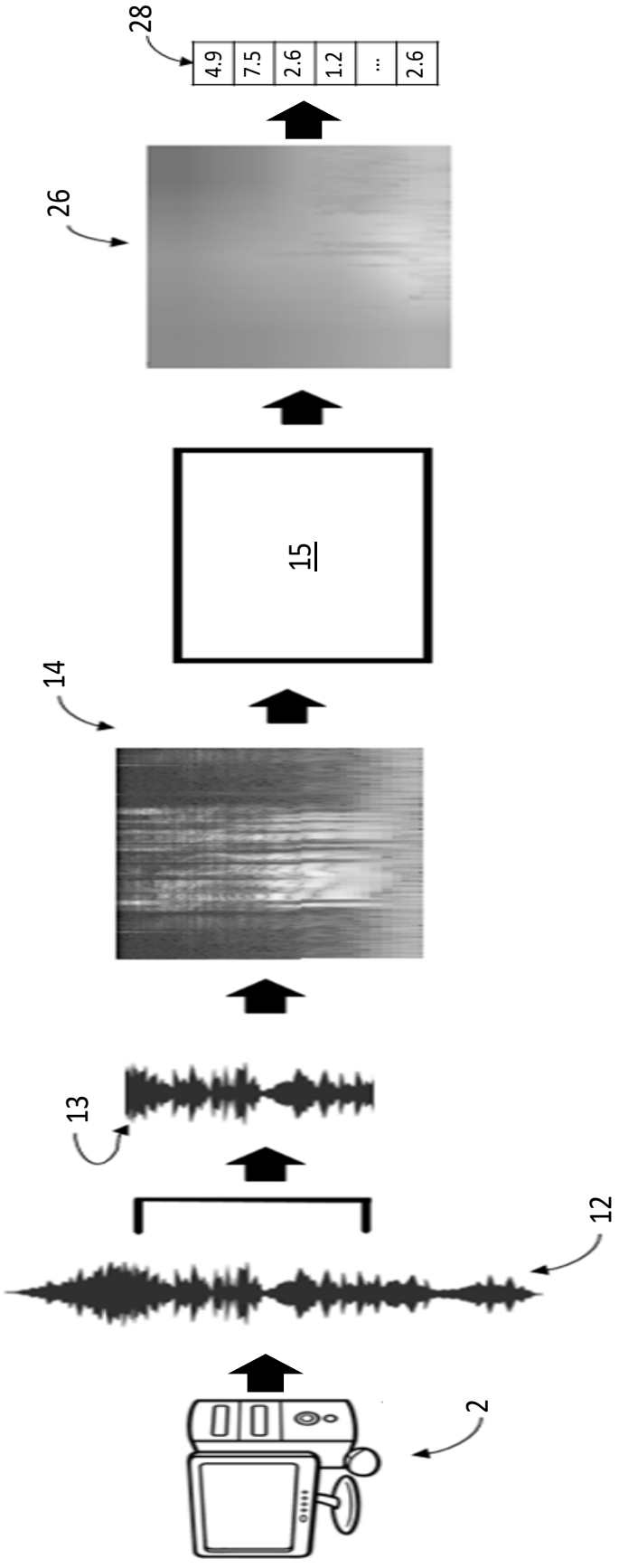


FIG. 6

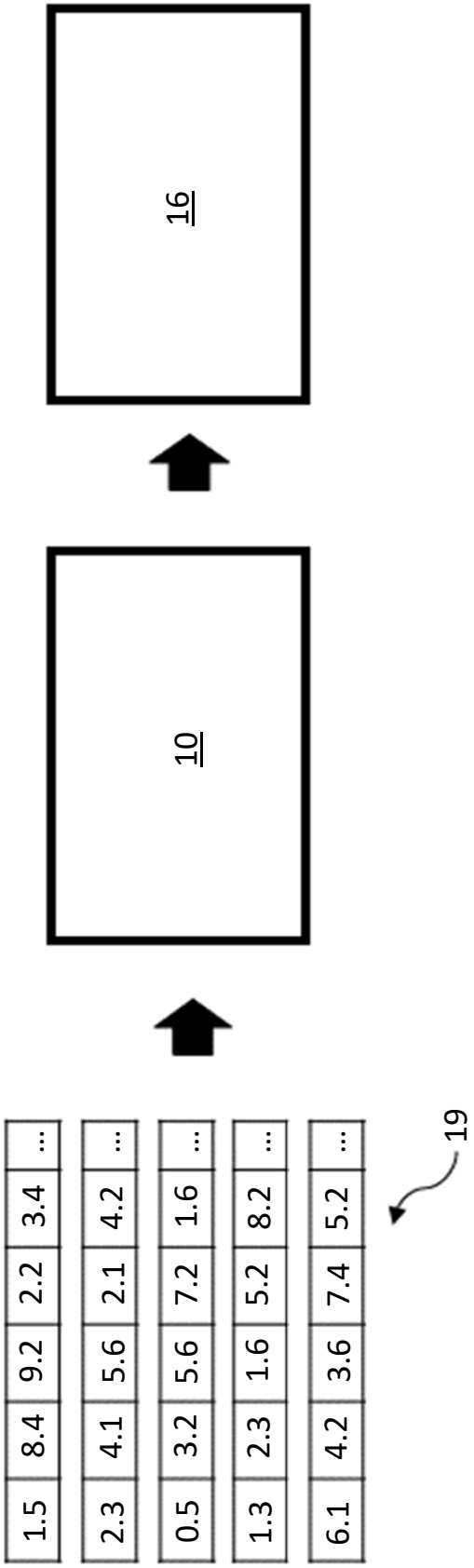

16
10
19
1.5
8.4
9.2
2.2
3.4
...
2.3
4.1
5.6
2.1
4.2
...
0.5
3.2
5.6
7.2
1.6
...
1.3
2.3
1.6
5.2
8.2
...
6.1
4.2
3.6
7.4
5.2
...

FIG. 7