(19) SUOMI - FINLAND
(FI)

PATENTTI- JA REKISTERIHALLITUS
PATENT- OCH REGISTERSTYRELSEN
FINNISH PATENT AND REGISTRATION OFFICE

(10) **FI 129402 B**
(12) **PATENTTIJULKAISU**
PATENTSKRIFT
PATENT SPECIFICATION

(45) Patentti myönnetty - Patent beviljats - Patent granted **31.01.2022**

(51) Kansainvälinen patenttiluokitus - Internationell patentklassifikation - International patent classification
G10L 15/24 (2013.01)
G10L 15/16 (2006.01)
G10L 15/06 (2013.01)
G10L 15/02 (2006.01)
G10L 25/84 (2013.01)
G10L 25/93 (2013.01)
G10L 15/20 (2006.01)
G06N 3/08 (2006.01)
G06N 20/00 (2019.01)

(21) Patenttihakemus - Patentansökning - Patent application 20206028

(22) Tekemispäivä - Ingivningsdag - Filing date **19.10.2020**

(23) Saapumispäivä - Ankomstdag - Reception date **19.10.2020**

(43) Tullut julkiseksi - Blivit offentlig - Available to the public **31.01.2022**

(73) Haltija - Innehavare - Proprietor
1 • OP Osuuskunta, Gebhardinaukio 1, 00510 HELSINKI, SUOMI - FINLAND, (FI)

(72) Keksijä - Uppfinnare - Inventor
1 • Taipale, Pauli, HELSINKI, SUOMI - FINLAND, (FI)

(74) Asiamies - Ombud - Agent
Kolster Oy Ab, Salmisaarenaukio 1, 00180 Helsinki

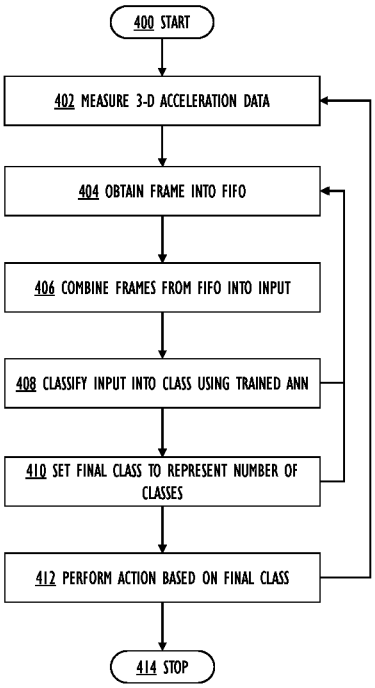
(54) Keksinnön nimitys - Uppfinningens benämning - Title of the invention
Laite ja menetelmä äänettömän puheen ilmaisemiseen
Anordning och förfarande för detektering av ljudlöst tal
APPARATUS AND METHOD FOR DETECTING SILENT SPEECH

(56) Viitejulkaisut - Anförda publikationer - References cited
US 10621973 B1, US 2015095036 A1, JP H04257899 A, CN 111326179 A,
LEE, S. et al. An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition. In: Nature Communications [online], 2019-06-18, Vol. 10, article number 2468, NLM31213598

(57) Tiivistelmä - Sammandrag - Abstract

Laite ja menetelmä hiljaisen puheen detektointiin. Menetelmä sisältää: mitataan (402) kolmiulotteista kiihtyvyysdataa käyttäjän puhe-elinten aiheuttamista vibraatioista subvokalisaation aikana käyttäen kiihtyvyysanturia, joka on konfiguroitu olemaan ulkoisesti kiinnitettävissä käyttäjän kaulan etuosaan leuan alapuolelle; saadaan (404) toistuvasti first-in first-out -puskuriin kiihtyvyysanturilta kiihtyvyysdatakehys; yhdistetään (406) kiihtyvyysdatakehykset first-in first-out -puskurista inputdatarakenteeseen; luokitellaan (408) kukin inputdatarakenne luokkaan, joka on valittu luokkasetin joukosta käyttäen opetettua keinotekoista neuroverkkoa, kukin luokka ilmaisten spesifistä sanaa tai taustamelua; asetetaan (410) lopullinen luokka edustamaan ennaltamäärättyä lukumäärää inputdatarakenteiden peräkkäisiä luokkia; ja suoritetaan (412) toiminta perustuen lopulliseen luokkaan.

An apparatus and method for detecting silent speech. The method includes: measuring (402) three-dimensional acceleration data from vibrations caused by speech organs of the user during subvocalization using an accelerometer configured to be externally attachable to an anterior part of a neck below a mandible of a user; obtaining (404) repeatedly into a first-in first-out buffer from the accelerometer an acceleration data frame; combining (406) the acceleration data frames from the first-in first-out buffer into an input data structure; classifying (408) each input data structure into a class selected among a set of classes using a trained artificial neural network, each class expressing a specific word or background noise; setting (410) a final class to represent a predetermined number of consecutive classes of the input data structures; and performing (412) an action based on the final class.



APPARATUS AND METHOD FOR DETECTING SILENT SPEECH

FIELD

Various embodiments relate to an apparatus for detecting silent speech, and to a method for detecting silent speech.

BACKGROUND

Silent speech (also known as subvocalization) is an internal speech, which typically manifests when reading, or speaking without an audible sound. During the subvocalization, speech organs of a person move, but in such a way that their movement is hardly visible. Electromyography (EMG) is used to record electrical activity of articulatory muscles during the subvocalization.

WO 2019/050881 A1 measures voltages produced by internal articulation at neuromuscular junctions over time at a set of electrodes positioned on the skin in head or neck. These signals are analyzed by a trained convolutional neural network.

US 10,126,828 B2 discloses a wearable apparatus using a bioacoustics sensor to measure a signal conducted via at least one bone of the user, and identifying a particular hand gesture based on the signal.

WO 2014/158451 A1 creates silent synthesized speech based on a vocal tract impedance profile, which is determined based on a reflected impulse response to an output signal directed to the vocal tract of the user.

US 2019/0189145 A1 interprets low amplitude or silent speech by analyzing sensor data to map the sensor data to at least one spoken syllable resulting in a string of one or more words and generating an electronic representation of the string of the one or more words. The sensor data is generated by sensors installed in a mouth (on a tongue or a palate, on teeth, mounted on a frame such as braces), by a video image capturing device placed external to the mouth, and by an airflow sensor positioned outside of the mouth.

US 2012/0075184 A1 detects silent speech command from the user by a camera-based lip-reading computer application.

US 10,621,973 B1 discloses a sub-vocal speech recognition apparatus including a headset worn over an ear and electromyography (EMG) electrodes and an inertial measurement unit (IMU, including 3-axis accelerometer, gyroscope and magnetometer) in contact with the skin in a position over the neck, under the chin and behind the ear. As the user speaks or mouths word, EMG and

IMU signals are recorded, amplified, filtered, and divided in multi-millisecond time windows, and analyzed by a neural network, a connectionist temporal classification (CTC) function and a language model. The EMG electrodes measure muscle electrical activity in response to a nerve's stimulation of the muscle. The IMU sensors collect data relating to inertial movements of the chin module. EMG and IMU signals are converted into aggregated vector representation. The IMU is used to produce chin movement signals of speech impaired people who cannot produce vibrations as they have lost their ability to form words and sounds and consequently cannot use an electrolarynx device.

US 7,574,357 B1 generates electromyographic sub-audible signals in an environment that interferes excessively with normal speech or that requires stealth communications.

All in all, the prior art subvocalization analysis requires a relatively complex setup, with EMG measurement sensors, image sensors, accelerometers, or other types of sensors placed in the head area or even inside the mouth, or even ultrasonic imaging or invasive brain-machine-interfaces.

Further sophistication (as regards to complexity, size, usability, costs etc.) is required if subvocalization is to be used as an input method in everyday life.

BRIEF DESCRIPTION

According to an aspect, there is provided subject matter of independent claims. Dependent claims define some embodiments.

One or more examples of implementations are set forth in more detail in the accompanying drawings and the description of embodiments.

LIST OF DRAWINGS

Some embodiments will now be described with reference to the accompanying drawings, in which

FIG. 1A, FIG. 1B, FIG. 2A and FIG. 2B illustrate embodiments of an apparatus for detecting silent speech;

FIG. 3 illustrates an embodiment of an accelerometer of the apparatus for detecting speech;

FIG. 4A and FIG. 4B are flow charts illustrating embodiments of a method for detecting silent speech;

FIG. 5 illustrates embodiments of three-dimensional acceleration data

measured from vibrations caused by speech organs of a user during subvocalization;

FIG. 6A and FIG. 6B illustrate embodiments of classification related to detecting silent speech; and

FIG. 7 illustrates an embodiment of a trained artificial neural network detecting silent speech.

DESCRIPTION OF EMBODIMENTS

The following embodiments are only examples. Although the specification may refer to “an” embodiment in several locations, this does not necessarily mean that each such reference is to the same embodiment(s), or that the feature only applies to a single embodiment. Single features of different embodiments may also be combined to provide other embodiments. Furthermore, words "comprising" and "including" should be understood as not limiting the described embodiments to consist of only those features that have been mentioned and such embodiments may contain also features/structures that have not been specifically mentioned.

Reference numbers, both in the description of the embodiments and in the claims, serve to illustrate the embodiments with reference to the drawings, without limiting it to these examples only.

The embodiments and features, if any, disclosed in the following description that do not fall under the scope of the independent claims are to be interpreted as examples useful for understanding various embodiments of the invention.

Let us study simultaneously FIG. 1A, which illustrates embodiments of an apparatus 100 for detecting silent speech, and FIG. 4A and FIG. 4B, which illustrate embodiments of a method for detecting silent speech. The method may be implemented as an algorithm 120 programmed as computer program code 106, executed by the apparatus 100 as a special purpose computer.

The apparatus 100 comprises an accelerometer 110 configured to be externally attachable to an anterior part of a neck below a mandible of a user 150 to measure 402 three-dimensional acceleration data from vibrations caused by speech organs (or articulators, or articulatory organs) of the user 150 during subvocalization. The accelerometer 110 may be of a similar type as used in modern smartwatches. The accelerometer 110 may be a part of an inertial measurement unit (IMU), which also comprises a magnetometer and a gyroscope.

The applicant has experimented, and the three-dimensional acceleration data provides reliable results, although less dimensions such as two, or more dimensions such as six, may be used depending on the use case and the measurement accuracy of the accelerometer 110.

In an embodiment, the accelerometer 110 is configured to measure 402 the vibrations caused by one or more of the following speech organs of the user 150: a larynx, vocal cords, a glottis, a pharynx, a velum, a tongue, lips, muscles involved in the subvocalization.

In an embodiment illustrated in FIG. 3, the accelerometer 110 is configured to be attachable in the vicinity of a laryngeal prominence of the user 150. The accelerometer 110 may be placed on top of the laryngeal prominence, or on either side of the laryngeal prominence.

FIG. 3 also illustrates an embodiment, wherein the accelerometer 110 is configured to be externally attachable by a flexible band 112 wearable around the neck of the user 150. The flexible band 112 may be manufactured of suitable natural and/or synthetic material offering suitable stretch and wearing comfort. The flexible band 112 may also include a suitable mechanism such as a snap lock or other type of fastener to open/close the loop of the flexible band 112.

In an alternate embodiment, the accelerometer 110 is configured to be externally attachable by an adhesive. The adhesive may affix the accelerometer 110 directly on the skin of the user 150, or the adhesive may affix a (textile or synthetic) label carrying the accelerometer 110 on the skin of the user 150.

In an embodiment, the accelerometer 110 is attached to an apparel so that when the apparel is worn by the user 150, the accelerometer 110 is on the anterior part of the neck below the mandible of the user 150. Such apparel may be a blouse, shirt or jacket, for example, or a strap of a helmet, for example. In some professional environments, the accelerometer 110 may be placed in an arrangement normally carrying a throat microphone, such as a flexible open-ended frame worn around the neck the of the user 150.

The apparatus 100 also comprises one or more memories 104 including the computer program code 106, and one or more processors 102 configured to execute the computer program code 106 to cause the apparatus 100 to perform the algorithm 120.

The term 'processor' 102 refers to a device that is capable of processing data. When designing the implementation of the processor 102, a person skilled in the art will consider the requirements set for the size and power

consumption of the apparatus 100, the necessary processing capacity, production costs, and production volumes, for example.

The term 'memory' 104 refers to a device that is capable of storing data run-time (= working memory) or permanently (= non-volatile memory). The working memory and the non-volatile memory may be implemented by a random-access memory (RAM), dynamic RAM (DRAM), static RAM (SRAM), a flash memory, a solid state disk (SSD), PROM (programmable read-only memory), a suitable semiconductor, or any other means of implementing an electrical computer memory.

A non-exhaustive list of implementation techniques for the processor 102 and the memory 104 includes, but is not limited to: logic components, standard integrated circuits, application-specific integrated circuits (ASIC), system-on-a-chip (SoC), application-specific standard products (ASSP), microprocessors, microcontrollers, digital signal processors, special-purpose computer chips, field-programmable gate arrays (FPGA), and other suitable electronics structures.

The computer program code 106 may be implemented by software. In an embodiment, the software may be written by a suitable programming language, and the resulting executable code may be stored in the memory 104 and executed by the processor 102.

The computer program code 106 implements the algorithm 120 for detecting the silent speech. The computer program code 106 may be coded as a computer program (or software) using a programming language, which may be a high-level programming language, such as C, C++, or Java, or a low-level programming language, such as a machine language, or an assembler, for example. The computer program code 106 may be in source code form, object code form, executable file, or in some intermediate form. There are many ways to structure the computer program code 106: the operations may be divided into modules, sub-routines, methods, classes, objects, applets, macros, etc., depending on the software design methodology and the programming language used. In modern programming environments, there are software libraries, i.e. compilations of ready-made functions, which may be utilized by the computer program code 106 for performing a wide variety of standard operations. In addition, an operating system (such as a general-purpose operating system) may provide the computer program code 106 with system services.

In an embodiment, the processor 102 may be implemented as a

microprocessor implementing functions of a central processing unit (CPU) on an integrated circuit. The CPU is a logic machine executing the computer program code 106. The CPU may comprise a set of registers, an arithmetic logic unit (ALU), and a control unit (CU). The control unit is controlled by a sequence of the computer program code 106 transferred to the CPU from the (working) memory 104. The control unit may contain a number of microinstructions for basic operations. The implementation of the microinstructions may vary, depending on the CPU design.

In an embodiment, the apparatus 100 may be implemented as a stand-alone apparatus 100 as shown in FIG. 1A and FIG. 1B, i.e., the apparatus 100 includes the one or more processors 102, the one or memories 104, the computer program code 106 and the accelerometer 110 encapsulated within a casing. As shown, the apparatus 100 may also comprise a wireless radio transceiver 108 configured to communicate with external actors, such as one or more of a wireless apparatus 160 of the user 150, a smartwatch 162 of the user, another personal apparatus (not illustrated) of the user 150, an external apparatus 164, or a networked server 166.

The wireless radio transceiver 108 may be configured to operate using one or more of the following: a cellular radio network, a wireless local area network (WLAN), or a short-range radio network (such as Bluetooth). In general, the wireless radio transceiver 108 may be interoperable with various wireless standard/non-standard/proprietary cellular radio networks such as any mobile phone network, which may be coupled with a wired network such as the Internet.

The wireless radio transceiver 108 may be implemented with a suitable cellular communication technology regardless of the generation (such as 2G, 3G, 4G, beyond 4G, 5G etc.) in their present forms and/or in their evolution forms, such as GSM, GPRS, EGPRS, WCDMA, UMTS, 3GPP, IMT, LTE, LTE-A, etc. and/or with a suitable non-cellular communication technology such as Bluetooth, Bluetooth Low Energy, Wi-Fi, WLAN, Zigbee, etc.

In an embodiment, the wireless radio transceiver 108 is coupled to a subscriber identity module (SIM), which may be an integrated circuit storing subscriber data, which is network-specific information used to authenticate and identify the subscriber on the cellular network. The subscriber identity module may be embedded into a removable SIM card. The subscriber identity module may also be an embedded-SIM (eSIM), embedded directly into the apparatus 100, and provisioned through software.

In an alternative embodiment, the apparatus 100 may be implemented as a distributed arrangement as shown in FIG. 2A and FIG. 2B: a measuring component 200 comprising the accelerometer 110, and a processing component 210, 210A, 210B, 210C, 210D comprising the one or memories 104A including the computer program code 106A and the one or more processors 102A. The communication between the distributed components may be implemented wirelessly so that the measuring component 200 comprises a wireless radio transceiver 202, and the processing component 210 comprises a wireless radio transceiver 212. A suitable standard/proprietary communication protocol is used in the communication. The communication protocol may be UDP (User Datagram Protocol), for example. Note that the measuring component 200 may also comprise one or memories 104B including computer program code 106B and one or more processors 102B. In this way, the algorithm 120 may be distributed into two parts: an algorithm part 120B performed in the measuring component 200 and an algorithm part 120A performed in the processing component 210.

The stand-alone apparatus 100 of FIG. 1A and FIG. 1B operates so that it independently detects the silent speech. Such stand-alone apparatus 100 may communicate information related to the detected silent speech with the external actors such as the wireless apparatus 160, the smartwatch 162, the other personal apparatus, the external apparatus 164, or the networked server 166.

The apparatus 100 implemented as the distributed arrangement 200, 210 operates so that at least a part of the algorithm/method for detecting the silent speech is implemented by an element other than the measuring component 200 attachable to the neck of the user 150. Consequently, at least a part of processing of the three-dimensional acceleration data, and/or an action based on the detected silent speech is performed by the external actors such as the wireless apparatus 160, 210A, the smartwatch 162, 210B, the other personal apparatus, the external apparatus 164, 210C, or the networked server 166, 210D. For example, the wireless apparatus (such as a smartphone or a pad computer) 160, 210A, the smartwatch 162, 210B or the other personal apparatus (such as a laptop) may perform a part of the processing, or accept the detected silent speech as user interface input. The external apparatus 164, 210C being a vehicle, a home automation system, or another kind of entity with which the user may interact, may perform a part of the processing, or accept the detected silent speech as user input. The networked server 166, 210D may also perform a part of the processing, or accept the detected silent speech as user input. The networked

server 166 may be a networked computer server, which interoperates with the apparatus 100 or the measuring component 200 according to a client-server architecture, a cloud computing architecture, a peer-to-peer system, or another applicable computing architecture.

An embodiment provides a computer-readable medium 170 storing the computer program code 106, which, when loaded into the one or more processors 102 and executed by the one or more processors 102, causes the one or more processors 102 to perform the algorithm/method, which will be explained with reference to FIG. 4A and FIG. 4B. The computer-readable medium 170 may comprise at least the following: any entity or device capable of carrying the computer program code 106 to the one or more processors 102, a record medium, a computer memory, a read-only memory, an electrical carrier signal, a telecommunications signal, and a software distribution medium. In some jurisdictions, depending on the legislation and the patent practice, the computer-readable medium 170 may not be the telecommunications signal. In an embodiment, the computer-readable medium 170 may be a computer-readable storage medium. In an embodiment, the computer-readable medium 170 may be a non-transitory computer-readable storage medium. The embodiment, to be described in the following, may be defined as: A computer-readable medium 170 comprising computer program code 106, which, when executed by one or more processors 102, causes performance of a method for detecting silent speech comprising: measuring 402 three-dimensional acceleration data from vibrations caused by speech organs of the user during subvocalization using an accelerometer configured to be externally attachable to an anterior part of a neck below a mandible of a user; obtaining 404 repeatedly into a first-in first-out buffer from the accelerometer an acceleration data frame containing three-dimensional acceleration data measured during a predetermined time period; after obtaining 404 each acceleration data frame, combining 406 the acceleration data frames from the first-in first-out buffer into an input data structure; classifying 408 each input data structure into a class selected among a set of classes using a trained artificial neural network, wherein the trained artificial neural network has been trained to recognize a set of words and background noise, each class expressing a specific word of the set of words or background noise; setting 410 a final class to represent a predetermined number of consecutive classes of the input data structures; and performing 412 an action based on the final class. Note that as the first operation 402 may be performed by

the measuring component 200, and the rest of the operations 404, 406, 408, 410, 412 by the processing component 210, the first operation 402 may be defined as it is for the measuring component 200, but for the processing component 210 it may combined with the operation 404 to define the first operation as follows: obtaining 404 repeatedly into a first-in first-out buffer from an accelerometer an acceleration data frame containing three-dimensional acceleration data measured during a predetermined time period, wherein the three-dimensional acceleration data has been measured 402 from vibrations caused by speech organs of the user during subvocalization using the accelerometer configured to be externally attachable to an anterior part of a neck below a mandible of a user.

Let us now study the algorithm/method with reference to FIG. 4A and FIG. 4B. Note that FIG. 4A depicts the main sequence of the algorithm/method, whereas FIG. 4B depicts optional embodiments.

The method starts in 400 and ends in 414. Note that the method may run as long as required (after the start-up of the apparatus 100 until switching off) by looping back to an operation 402.

The operations are not strictly in chronological order in FIG. 4A, and some of the operations may be performed simultaneously or in an order differing from the given ones. Other functions may also be executed between the operations or within the operations and other data exchanged between the operations. Some of the operations or part of the operations may also be left out or replaced by a corresponding operation or part of the operation. It should be noted that no special order of operations is required, except where necessary due to the logical requirements for the processing order.

As already explained, in 402, the three-dimensional acceleration data is measured by the accelerometer 110 from the vibrations caused by the speech organs of the user 150 during the subvocalization.

In 404, an acceleration data frame 124, 126, 128, 130, 132 containing three-dimensional acceleration data measured during a predetermined time period is obtained repeatedly into a first-in first-out (FIFO) buffer 122 from the accelerometer 110.

After obtaining each acceleration data frame 124, 126, 128, 130, 132 in 404, the acceleration data frames 124, 126, 128, 130, 132 are combined from the first-in first-out buffer 122 into an input data structure 134 in 406.

In 408, each input data structure 134 is classified into a class 138, 140, 142 selected among a set of classes using a trained artificial neural network 136A.

The trained artificial neural network 136A has been trained to recognize a set of words and background noise. Each class 138, 140, 142 expresses a specific word of the set of words or background noise.

FIG. 5 illustrates embodiments of two different classes, A and B, with three different samples 1-3 in each class A and B. The signal curves X, Y an Z depict acceleration (Y axis in g units) as a function of time (X axis in milliseconds). One g is the force per unit mass due to gravity at the Earth's surface defined as 9,80665 m/s^2. As shown, the acceleration signals in three different dimensions, X, Y and Z, vary considerably even within the same class between the samples, but, nevertheless, the trained artificial neural network 136A is capable of mapping the sample signal with the right class. In principle, another kind of machine learning algorithm instead of the trained artificial neural network 136A may be used, but the studies by the applicant indicate that the trained artificial neural network 136A provides the most precise recognition of the right class.

The trained artificial neural network 136A has been trained to recognize the set of words and the background noise. As shown in FIG. 5, various samples may be obtained from different persons with different characteristics, and each sample is labelled with the correct value. For example, a set of samples for the word "ONE" are obtained and thus labelled with "ONE". As the artificial neural network 136A is trained with the samples, the artificial neural network adjusts its internal mathematical models so that the predicted output for a specific sample better matches with the label of the sample. The training may be called supervised learning, which is suited for the pattern/sequence recognition, i.e., for the classification of the sequential words.

When training data (vibrations produced by subvocalized speech) is gathered in order to train the artificial neural network 136A, the acceleration signal is detected by an algorithm, which utilizes a moving average of past signal value magnitudes. If the vibration magnitude in one particular time reaches a value above the moving average, a recording for the acceleration signal is initiated. As training data is gathered, a class corresponding to “no signal” or “background noise” is gathered while moving average detector detects no signal. This gathered data is given to the artificial neural network 136A in order to train it to detect background signals without the moving average filter, producing superior results compared to when only using the moving average filter.

In an embodiment, the first-in first-out buffer 122 is dimensioned to accommodate five acceleration data frames 124, 126, 128, 130, 132, but the

number may vary depending on the implementation. The five input tensors may each be separated by 100 ms corresponding to parts of subvocalized speech, and they are classified by the artificial neural network 136A as the same class (or 3 out of 5 tensors if separation is 300 ms) in order for the signal to proceed to the majority voting phase of the classification process. For a single input tensor, there is a threshold set for the confidence level that the classifier needs to reach, otherwise the input tensor vibration is considered as the background signal.

FIG. 7 illustrates an example implementation of the trained artificial neural network 136A using the "Keras" deep learning API written in Python and running on top of the machine learning platform "TensorFlow".

An input tensor 700 (as the input data structure 134) is passed to parallel 1D convolution layers, the first convolution layer 716 with a stride length of 40, the second convolution layer 718 with a stride length of 20, the third convolution layer 720 with a stride length of 10, and the fourth convolution layer 724 with a stride length of 1. The convolution layers 716, 718, 720, 724 each create a convolution kernel that is convolved with the layer input over a single temporal dimension to produce a tensor of outputs. The fourth convolution layer 724 is preceded by a max pooling 1D layer 722, which downsamples the input representation by taking the maximum value over the window. A concatenate layer 726 takes as input a list of tensors, all of the same shape except for the concatenation axis, and returns a single tensor that is the concatenation of all inputs. A batch normalization layer 728 applies a transformation that maintains the mean output close to 0 and the output standard deviation close to 1. An activation layer 730 applies an activation function to an output. As shown in FIG. 7, the layers grouped together by a box 732 are repeated three times.

In parallel to the box 732, a left-hand branch is performed once. The branch begins with a convolution layer 702 with a stride length of 1. Next, a batch normalization layer 704 is used to normalize its inputs. A global average pooling 1D layer 706 is applied for the temporal data. A dense layer 708, 710 is applied twice in suggestion to create a densely-connected NN layer. A list of inputs are multiplied (elementwise) by a multiply layer 712. Finally, a batch normalization layer 714 is used to normalize its inputs as the final layer of the left-hand branch.

An add layer 734 takes as input a list of tensors, all of the same shape, and returns a single tensor (also of the same shape). The list of inputs stems from the once-performed left-hand branch and the three times performed box 732. As shown, all levels so far including the once performed left-hand branch and the

three times performed box 732 may be grouped into a box 736, which may be performed only once, or it may also be performed N times, N being any integer greater than 1.

Finally, a global average pooling 1D layer 734 is applied on the temporal data, and a dense layer 740 is applied to create a densely-connected NN layer, and an output tensor 742 (including the class 138, 140, 142) is outputted from the trained artificial neural network 136A.

In an embodiment illustrated in FIG. 4B, the classifying 408 includes testing of a confidence level of the class 138, 140, 142 in 420. If the confidence level of the class 138, 140, 142 selected in the classifying 408 is below a predetermined threshold value 420-YES, a class expressing the background noise is set in 424 as the class 138, 140, 142, or else the selected class 138, 140, 142 is set in 422 as the class (i.e., the original selection remains as it is).

In 410, a final class 144 is set to represent a predetermined number of consecutive classes 138, 140, 142 of the input data structures 134.

An embodiment illustrated in FIG. 4B discloses one way of implementing this with a testing of a condition in 428. If a majority of the predetermined number of the consecutive classes 138, 140, 142 of the input data structures 134 is for a specific class expressing a specific word 428-YES, the specific class is set in 430 as the final class 144, or else 428-NO a class expressing the background noise is set in 434 as the final class 144.

Optionally, an operation 432 is also executed in the 428-YES branch after the operation 430: skipping the classifying 408 of future consecutive classes in 432 for the predetermined number subtracted by one.

In an embodiment, the predetermined number is the number of acceleration data frames 124, 126, 128, 130, 132 fitting into the first-in first-out buffer 122.

FIG. 6A and FIG. 6B illustrate an elaborate example of setting the final class 144 in 410 based on the predetermined number of consecutive classes. Let the number of the consecutive classes be three, whereby a window 600 moves along the classifications as they are made. The queue 0011001110010011 represents the classes in a timeline where time advances from left to right at one classification at a time from (1) to (15). R=0 represents a class expressing the background noise, and R=1 represents some class mapped to a word from among the set of words. The resulting sequence is as follows:

(1) The three consecutive classes in the window have values 0, 0 and 1,

which results (according to the majority rule) in R=0, i.e. the final class is set as the background noise class.

(2) Window=[0, 1, 1] -> R=1, SKIP 2, i.e., the final class is set as the (some) word class. Note that if the two word classes are different, then the final class is set as the background noise class. However, let us define, for the sake of simplicity, that always when the window has two or three values of 1, they are all for the same word class, whereby according to the majority rule R=1. SKIP 2 refers to the earlier explained embodiment, wherein the classifying of the future consecutive classes is skipped for 3-1=2 future classes.

(3) Window=[1, 1, 0] skipped.

(4) Window=[1, 0, 0] skipped.

(5) Window=[0, 0, 1] -> R=0.

(6) Window=[0, 1, 1] -> R=1, SKIP 2.

(7) Window=[1, 1, 1] skipped.

(8) Window=[1, 1, 0] skipped.

(9) Window=[1, 0, 0] -> R=0.

(10) Window=[0, 0, 1] -> R=0.

(11) Window=[0, 1, 0] -> R=0.

(12) Window=[1, 0, 0] -> R=0.

(13) Window=[0, 0, 0] -> R=0.

(14) Window=[0, 0, 1] -> R=0.

(15) Window=[0, 1, 1] -> R=1.

In 412, an action is performed based on the final class 144. In an embodiment, the action comprises one or more of the following: giving 436 a command to the apparatus 100, giving 438 an input string to the apparatus 100, transmitting 440 information outside of the apparatus 100. In the embodiments of FIG. 1A and FIG. 1B, and in the embodiments of FIG. 2A and FIG. 2B, the command or input string may be given or information transmitted to the wireless apparatus 160, the smartwatch 162, the other personal apparatus, the external apparatus 164, or the networked server 166.

In an embodiment illustrated in FIG. 1A, the classifying 408 is performed in parallel using two or more trained artificial neural networks 136A, 136B, and the class is decided in 426 based on weighting classes obtained from each of the two or more trained artificial neural networks 136A, 136B. The advantage of this may in some use cases be a more reliable detection of the silent speech.

The embodiments take advantage of the physics of speech production to harness the vibrations produced by the speech organs during the silent speech. The detected silent speech may be used in human-computer interaction and user interfaces as a novel input mechanism for controlling services and devices. In practice, this means that the embodiments recognize the words the user is thinking, as the speech organs are activated and produce vibration regardless of the will when the user subvocalizes (or in optimal circumstances just thinks of) a word. The embodiments may thus detect a vibration that correlates with the word or speech the user is thinking of. The embodiments enable the user to convey personal or secret information to a machine in a data secure way. The embodiments may also be used in environments with a high level of background noise. The embodiments are especially useful for paralyzed or severely disabled people with a speech disability.

Even though the invention has been described with reference to one or more embodiments according to the accompanying drawings, it is clear that the invention is not restricted thereto but can be modified in several ways within the scope of the appended claims. All words and expressions should be interpreted broadly, and they are intended to illustrate, not to restrict, the embodiments. It will be obvious to a person skilled in the art that, as technology advances, the inventive concept can be implemented in various ways.

CLAIMS

1. An apparatus (100) for detecting silent speech comprising:

an accelerometer (110) configured to be externally attachable to an anterior part of a neck below a mandible of a user (150) to measure (402) three-dimensional acceleration data from vibrations caused by speech organs of the user (150) during subvocalization;

one or more memories (104) including computer program code (106); and

one or more processors (102) configured to execute the computer program code (106) to cause the apparatus (100) to perform at least the following:

obtaining (404) repeatedly into a first-in first-out buffer (122) from the accelerometer (110) an acceleration data frame (124, 126, 128, 130, 132) containing three-dimensional acceleration data measured during a predetermined time period;

after obtaining (404) each acceleration data frame (124, 126, 128, 130, 132) combining (406) the acceleration data frames (124, 126, 128, 130, 132) from the first-in first-out buffer (122) into an input data structure (134);

characterized by:

classifying (408) each input data structure (134) into a class (138, 140, 142) selected among a set of classes using a trained artificial neural network (136A), wherein the trained artificial neural network (136A) has been trained to recognize a set of words and background noise, each class (138, 140, 142) expressing a specific word of the set of words or background noise;

setting (410) a final class (144) to represent a predetermined number of consecutive classes (138, 140, 142) of the input data structures (134) so that if a majority of the predetermined number of the consecutive classes (138, 140, 142) of the input data structures (134) is for a specific class expressing a specific word (428-YES), setting (430) the specific class as the final class (144), or else setting (434) a class expressing the background noise as the final class (144); and

performing (412) an action based on the final class (144).

2. The apparatus of claim 1, wherein the accelerometer (110) is configured to be attachable in the vicinity of a laryngeal prominence of the user (150).

3. The apparatus of any preceding claim, wherein the accelerometer

(110) is configured to be externally attachable by a flexible band (112) wearable around a neck of the user (150).

4. The apparatus of any preceding claim, wherein the accelerometer (110) is configured to be externally attachable by an adhesive.

5. The apparatus of any preceding claim, wherein the accelerometer (110) is configured to measure (402) the vibrations caused by one or more of the following speech organs of the user (150): a larynx, vocal cords, a glottis, a pharynx, a velum, a tongue, lips, muscles involved in the subvocalization.

6. The apparatus of any preceding claim, wherein the apparatus (100) is further caused to:

if a confidence level of the class (138, 140, 142) selected in the classifying (408) is below a predetermined threshold value (420-YES), setting (424) a class expressing the background noise as the class (138, 140, 142).

7. The apparatus of any preceding claim, wherein the apparatus (100) is further caused to:

if the majority of the predetermined number of the consecutive classes (138, 140, 142) of the input data structures (134) is for the specific class expressing the specific word (428-YES), setting (430) the specific class as the final class (144) and also skipping (432) the classifying (408) of future consecutive classes for the predetermined number subtracted by one.

8. The apparatus of any preceding claim, wherein the predetermined number is the number of acceleration data frames (124, 126, 128, 130, 132) fitting into the first-in first-out buffer (122).

9. The apparatus of any preceding claim, wherein the apparatus (100) is further caused to:

performing the classifying (408) in parallel using two or more trained artificial neural networks (136A, 136B), and deciding (426) the class based on weighting classes obtained from each of the two or more trained artificial neural networks (136A, 136B).

10. The apparatus of any preceding claim, wherein the action comprises one or more of the following: giving (436) a command to the apparatus (100), giving (438) an input string to the apparatus (100), transmitting (440) information outside of the apparatus (100).

11. A method for detecting silent speech comprising:

measuring (402) three-dimensional acceleration data from vibrations caused by speech organs of the user during subvocalization using an

accelerometer configured to be externally attachable to an anterior part of a neck below a mandible of a user;

obtaining (404) repeatedly into a first-in first-out buffer from the accelerometer an acceleration data frame containing three-dimensional acceleration data measured during a predetermined time period;

after obtaining (404) each acceleration data frame, combining (406) the acceleration data frames from the first-in first-out buffer into an input data structure;

characterized by:

classifying (408) each input data structure into a class selected among a set of classes using a trained artificial neural network, wherein the trained artificial neural network has been trained to recognize a set of words and background noise, each class expressing a specific word of the set of words or background noise;

setting (410) a final class to represent a predetermined number of consecutive classes of the input data structures so that if a majority of the predetermined number of the consecutive classes of the input data structures is for a specific class expressing a specific word (428-YES), setting (430) the specific class as the final class, or else setting (434) a class expressing the background noise as the final class; and

performing (412) an action based on the final class.

12. The method of claim 11, further comprising:

if a confidence level of the class selected in the classifying is below a predetermined threshold value (420-YES), setting (430) a class expressing the background noise as the class.

13. The method of claim 11 or 12, further comprising:

if the majority of the predetermined number of the consecutive classes of the input data structures is for the specific class expressing the specific word (428-YES), setting (430) the specific class as the final class and also skipping (432) the classifying (408) of future consecutive classes for the predetermined number subtracted by one.

VAATIMUKSET

1. Laite (100) äänettömän puheen ilmaisemiseen käsittäen:

kiihtyvyysanturin (110), joka on konfiguroitu ulkoisesti kiinnitettäväksi käyttäjän (150) alaleuan alle kaulan etuosaan mittaamaan (402) kolmiulotteista kiihtyvyysdataa käyttäjän (150) puhe-elimien aiheuttamista vibraatioista subvokalisaation aikana;

yksi tai useampi muisti (104) sisältäen tietokoneohjelmakoodin (106); ja

yhden tai useamman prosessorin (102), jotka on konfiguroitu suorittamaan tietokoneohjelmakoodia (106) aiheuttamaan laitteen (100) suorittamaan ainakin seuraavan:

saadaan (404) toistuvasti first-in-first-out-puskuriin (122) kiihtyvyysanturilta (110) kiihtyvyysdatakehys (124, 126, 128, 130, 132) sisältäen ennakolta määrätyn ajanjakson aikana mitattua kolmiulotteista kiihtyvyysdataa;

kunkin kiihtyvyysdatakehyksen (124, 126, 128, 130, 132) saamisen (404) jälkeen yhdistetään (406) kiihtyvyysdatakehykset (124, 126, 128, 130, 132) first-in-first-out-puskurista (122) inputdatarakenteeksi (134);

tunnettu siitä, että:

luokitellaan (408) kukin inputdatarakenne (134) luokkaan (138, 140, 142), joka on valittu luokkien setin joukosta käyttäen opetettua keinotekoista neuroverkkoa (136A), jossa opetettu keinotekoinen neuroverkko (136A) on opetettu tunnistamaan sanojen setti ja taustamelu, kukin luokka (138, 140, 142) ilmaisten sanojen setin spesifistä sanaa tai taustamelua;

asetetaan (410) lopullinen luokka (144) edustamaan ennakolta määrättyä lukumäärää inputdatarakenteiden (134) peräkkäisiä luokkia (138, 140, 142) niin, että jos enemmistö ennakolta määrätyn lukumäärän inputdatarakenteiden (134) peräkkäisiä luokkia (138, 140, 142) on spesifille luokalle ilmaisten spesifiä sanaa (428-YES), asetetaan (430) spesifi luokka lopulliseksi luokaksi (144), tai muuten asetetaan (434) taustamelua ilmaiseva luokka lopulliseksi luokaksi (144); ja

suoritetaan (412) toiminta perustuen lopulliseen luokkaan (144).

2. Patenttivaatimuksen 1 mukainen laite, jossa kiihtyvyysanturi (110) on konfiguroitu kiinnitettäväksi käyttäjän (150) aataminomenan lähistölle.

3. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa kiihtyvyysanturi (110) on konfiguroitu ulkoisesti kiinnitettäväksi käyttäjän (150) kaulan ympäri pidettävällä joustavalla nauhalla (112).

4. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa kiihtyvyysanturi (110) on konfiguroitu ulkoisesti kiinnitettäväksi sideaineella.

5. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa kiihtyvyysanturi (110) on konfiguroitu mittaamaan (402) yhden tai useamman käyttäjän (150) seuraavan puhe-elimen aiheuttamia vibraatioita: kurkunpää, äänihuulet, äänielimet, nielu, pehmeä suulaki, kieli, huulet, subvokalisaatiossa mukana oleva lihakset.

6. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa laite (100) on lisäksi aiheutettu:

jos luokittelussa (408) valitun luokan (138, 140, 142) luotettavuustaso on alle ennakolta määrätyn kynnysarvon (420-YES), asetetaan (424) taustamelua ilmaiseva luokka luokaksi (138, 140, 142).

7. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa laite (100) on lisäksi aiheutettu:

jos enemmistö ennakolta määrätyn lukumäärän inputdatarakenteiden (134) peräkkäisiä luokkia (138, 140, 142) on spesifille luokalle ilmaisten spesifiä sanaa (428-YES), asetetaan (430) spesifi luokka lopulliseksi luokaksi (144) ja myös ohitetaan (432) tulevien peräkkäisten luokkien luokittelu (408) ennakolta määrätylle lukumäärälle vähennettynä yhdellä.

8. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa ennakolta määrätty lukumäärä on first-in-first-out-puskuriin (122) mahtuvien kiihtyvyysdatakehyksien (124, 126, 128, 130, 132) lukumäärä.

9. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa laite (100) on lisäksi aiheutettu:

suoritetaan luokittelu (408) rinnakkain käyttäen kahta tai useampaa opetettua keinotekoista neuroverkkoa (136A, 136B), ja päätetään (426) luokka perustuen kustakin kahdesta tai useammasta opetetusta keinotekoisesta neuroverkosta (136A, 136B) saadun luokan painottamiseen.

10. Jonkin edellisen patenttivaatimuksen mukainen laite, jossa toiminta käsittää yhden tai useamman seuraavista: annetaan (436) laitteelle (100) komento, annetaan (438) laitteelle (100) inputmerkkijono, lähetetään (440) laitteen (100) ulkopuolelle informaatiota.

11. Menetelmä äänettömän puheen ilmaisemiseen käsittäen:

mitataan (402) kolmiulotteista kiihtyvyysdataa käyttäjän puhe-elimien aiheuttamista vibraatioista subvokalisaation aikana käyttäen kiihtyvyysanturia, joka on konfiguroitu ulkoisesti kiinnitettäväksi käyttäjän alaleuan alle kaulan etuosaan;

saadaan (404) toistuvasti first-in-first-out-puskuriin kiihtyvyysanturilta kiihtyvyysdatakehys sisältäen ennakolta määrätyn ajanjakson aikana mitattua kolmiulotteista kiihtyvyysdataa;

kunkin kiihtyvyysdatakehyksen saamisen (404) jälkeen yhdistetään (406) kiihtyvyysdatakehykset first-in-first-out-puskurista inputdatarakenteeksi;

tunnettu siitä, että:

luokitellaan (408) kukin inputdatarakenne luokkaan, joka on valittu luokkien setin joukosta käyttäen opetettua keinotekoista neuroverkkoa, jossa opetettu keinotekoinen neuroverkko on opetettu tunnistamaan sanojen setti ja taustamelu, kukin luokka ilmaisten sanojen setin spesifistä sanaa tai taustamelua;

asetetaan (410) lopullinen luokka edustamaan ennakolta määrättyä lukumäärää inputdatarakenteiden peräkkäisiä luokkia niin, että jos enemmistö ennakolta määrätyn lukumäärän inputdatarakenteiden peräkkäisiä luokkia on spesifille luokalle ilmaisten spesifiä sanaa (428-YES), asetetaan (430) spesifi luokka lopulliseksi luokaksi, tai muuten asetetaan (434) taustamelua ilmaiseva luokka lopulliseksi luokaksi; ja

suoritetaan (412) toiminta perustuen lopulliseen luokkaan.

12. Patenttivaatimuksen 11 mukainen menetelmä, lisäksi käsittäen:

jos luokittelussa valitun luokan luotettavuustaso on alle ennakolta määrätyn kynnysarvon (420-YES), asetetaan (430) taustamelua ilmaiseva luokka luokaksi.

13. Patenttivaatimuksen 11 tai 12 mukainen menetelmä, lisäksi käsittäen:

jos enemmistö ennakolta määrätyn lukumäärän inputdatarakenteiden peräkkäisiä luokkia on spesifille luokalle ilmaisten spesifiä sanaa (428-YES), asetetaan (430) spesifi luokka lopulliseksi luokaksi ja myös ohitetaan (432) tulevien peräkkäisten luokkien luokittelu (408) ennakolta määrätylle lukumäärälle vähennettynä yhdellä.

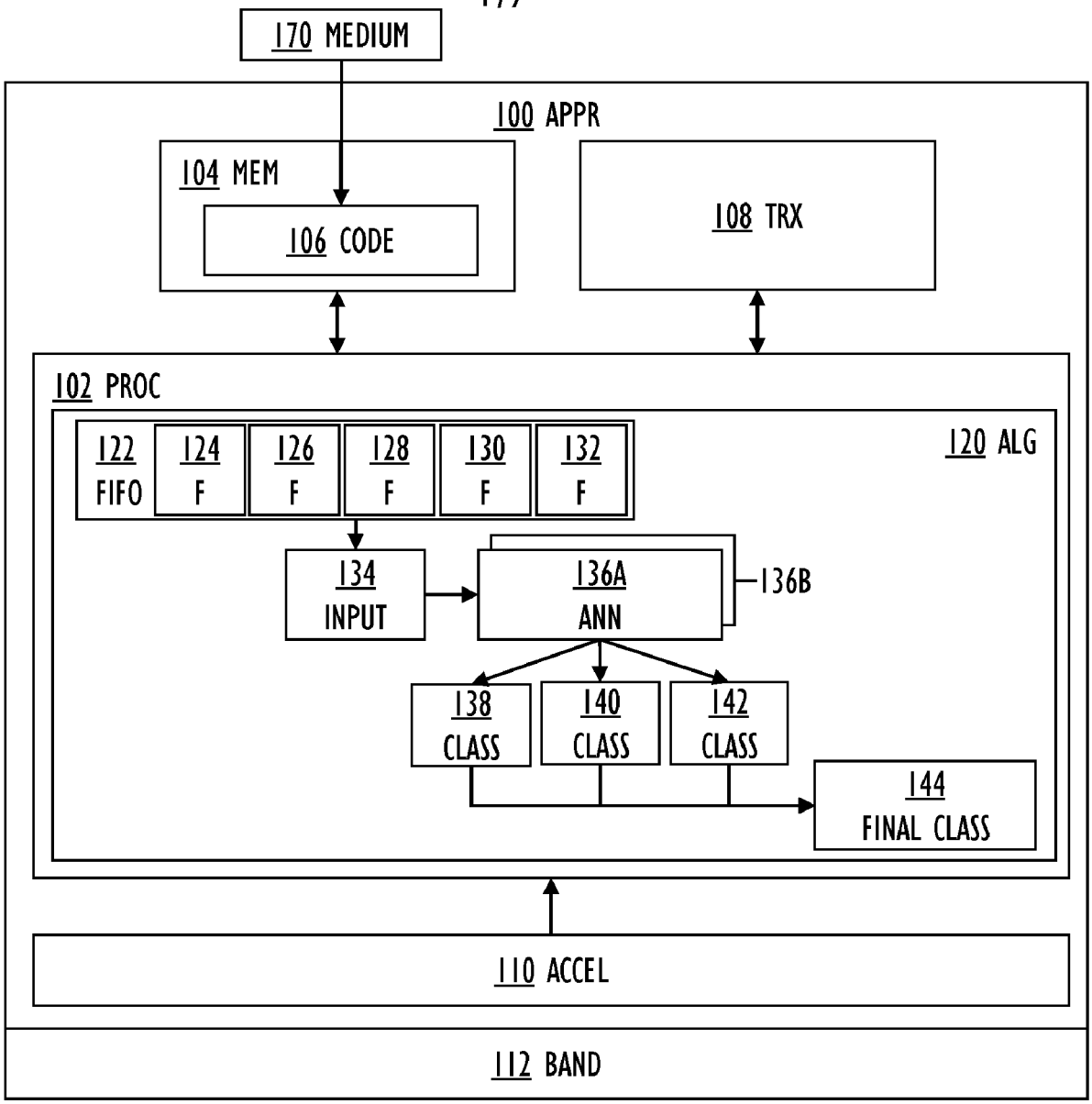


FIG. 1A

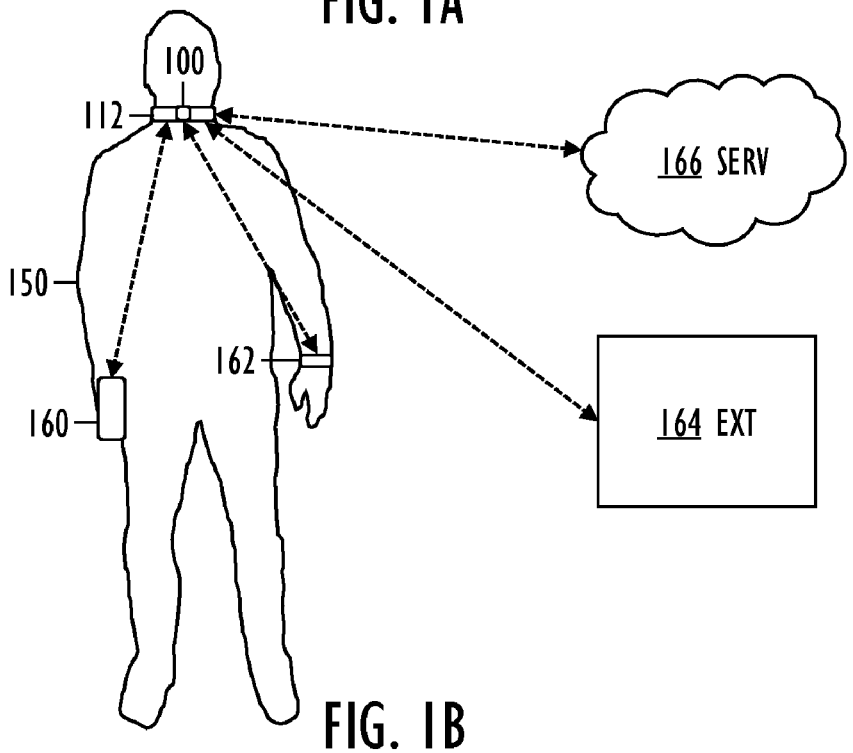


FIG. 1B

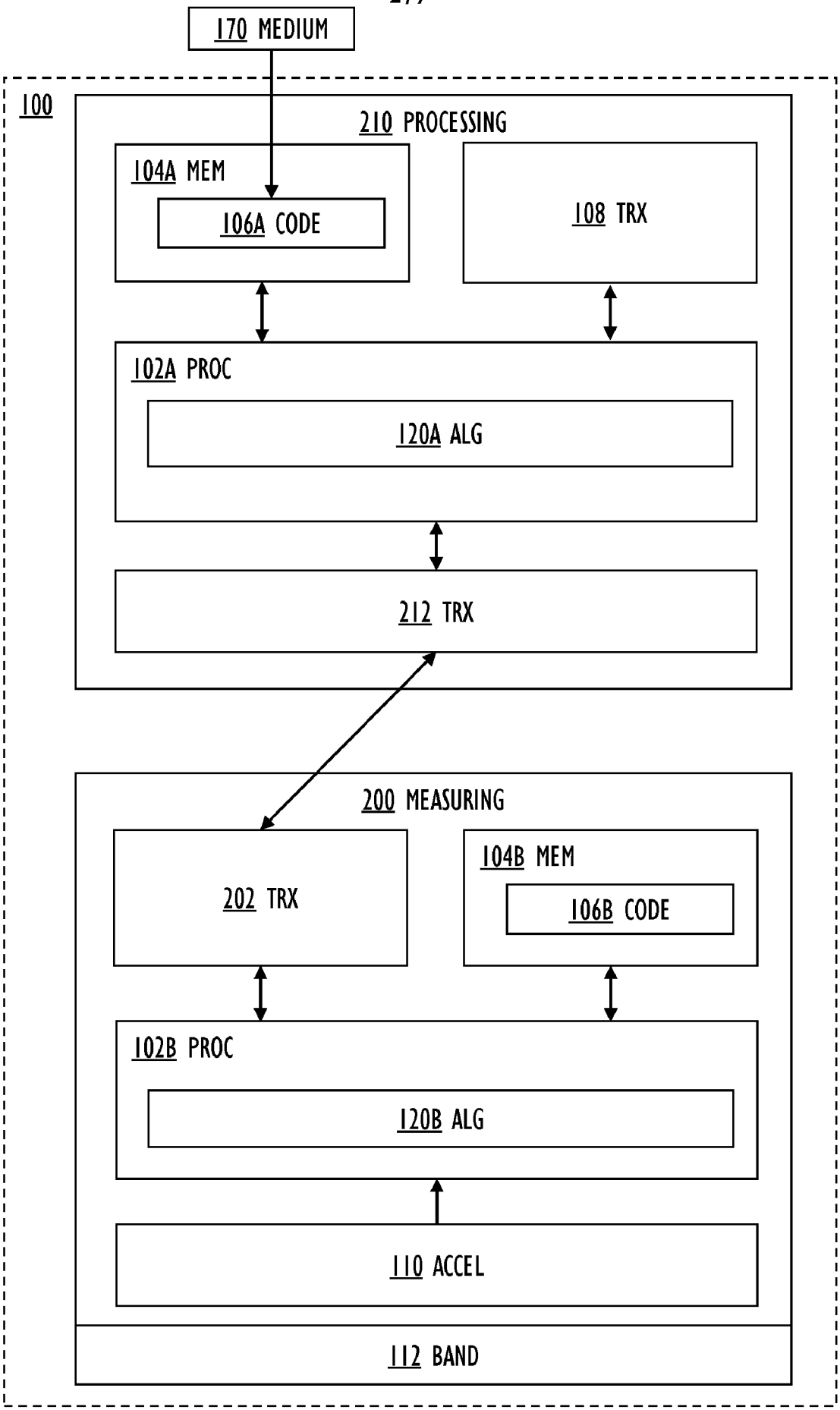


FIG. 2A

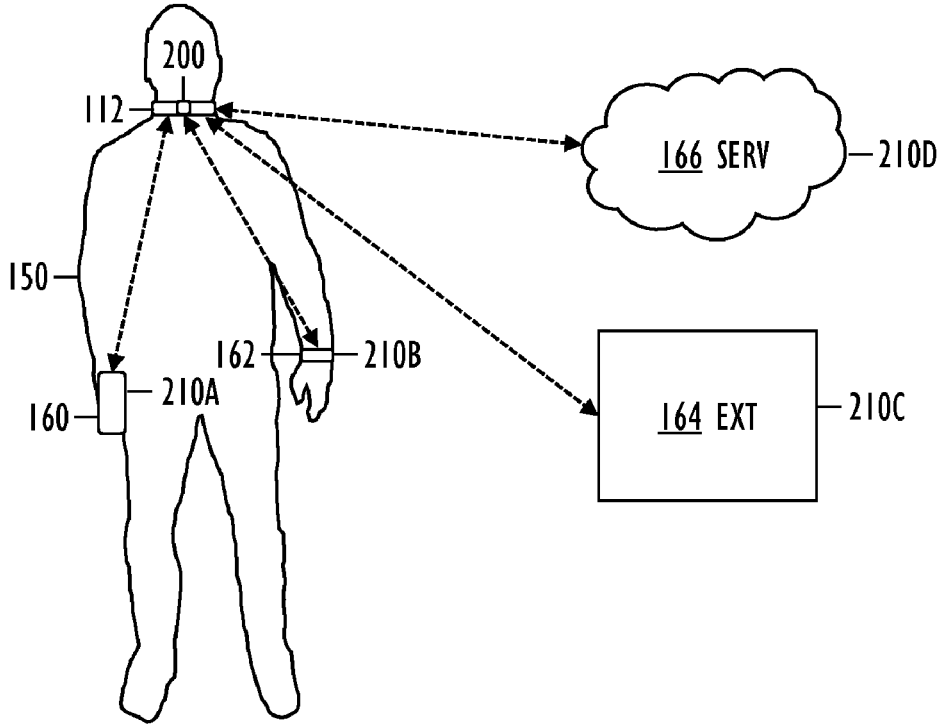


FIG. 2B

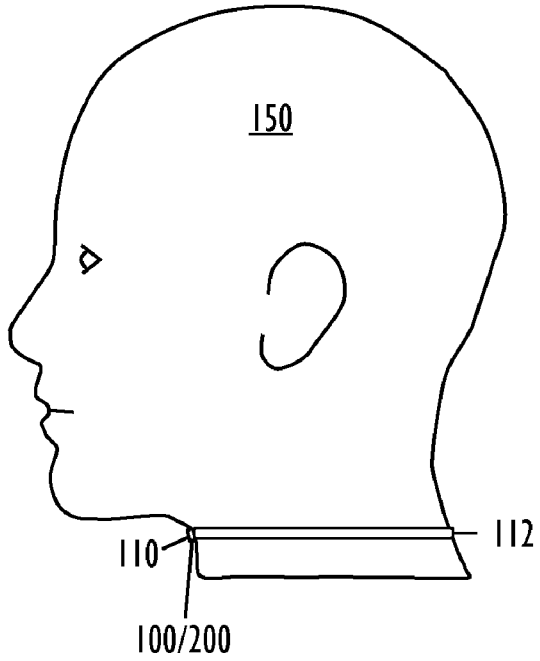


FIG. 3

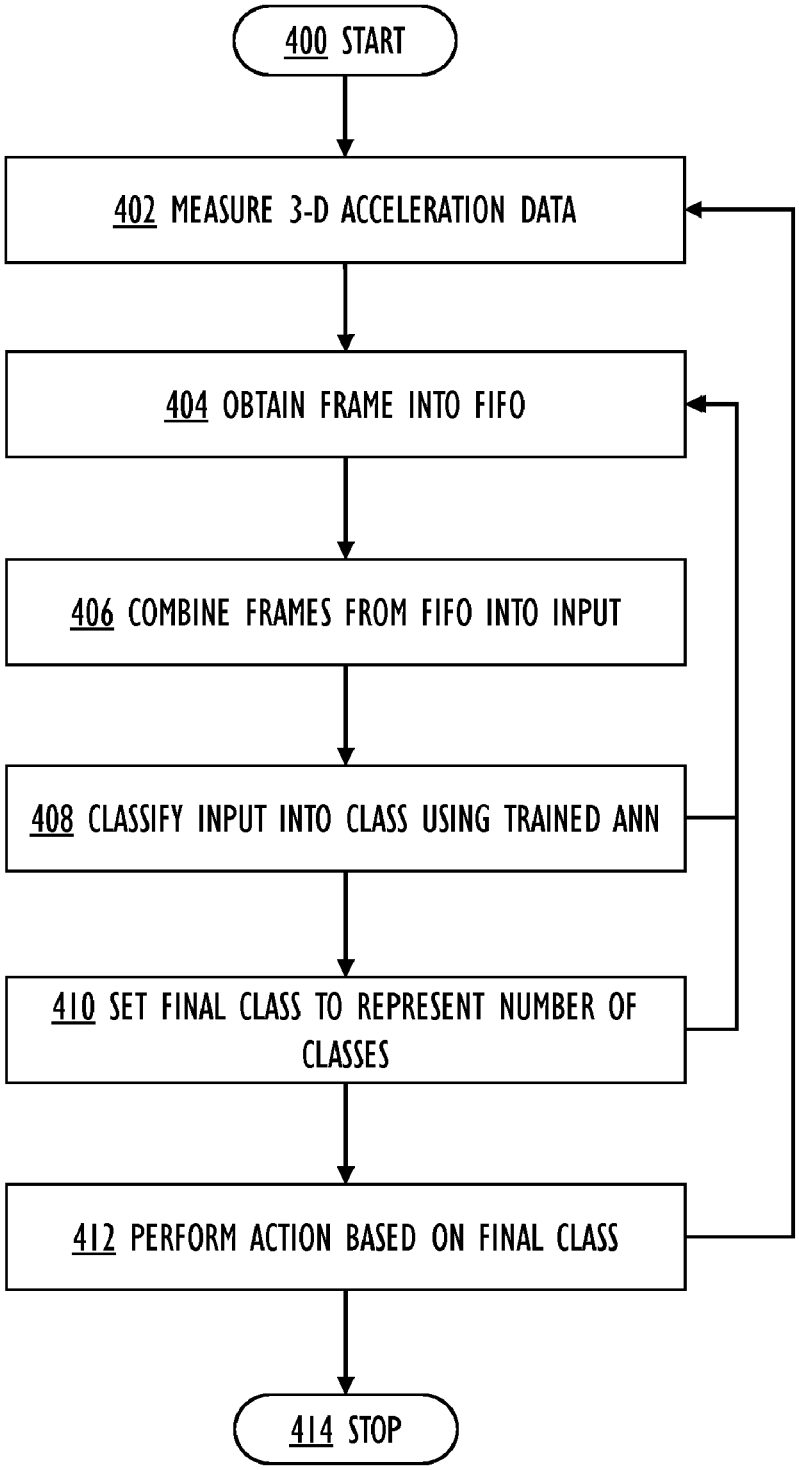


FIG. 4A

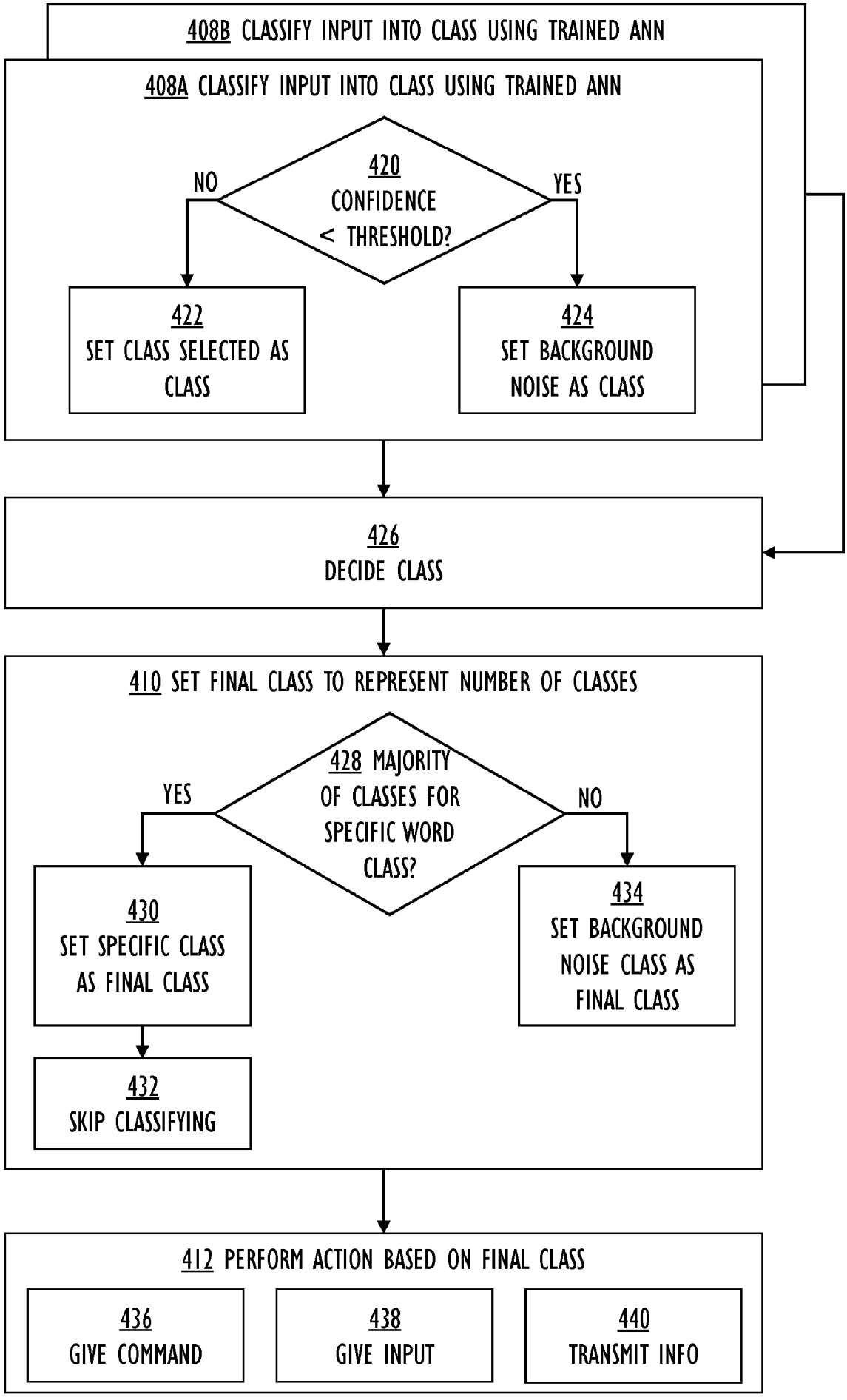
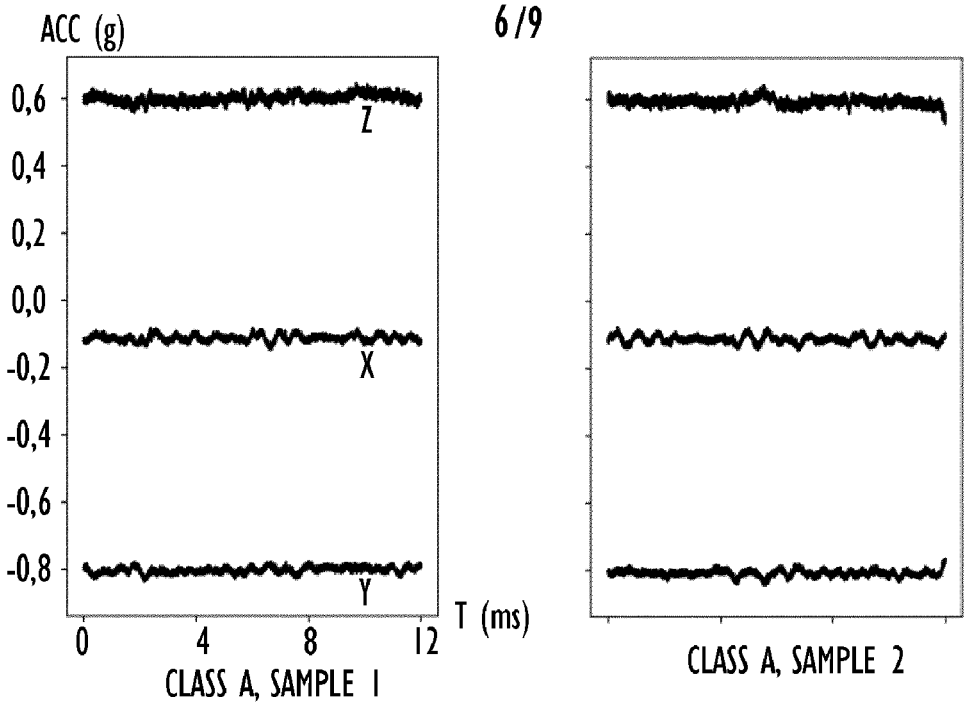
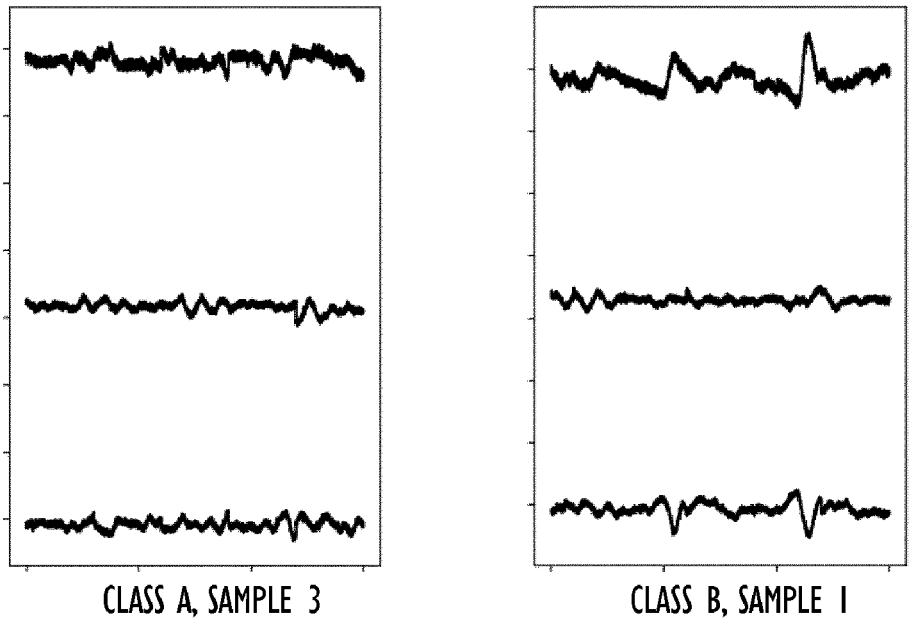


FIG. 4B





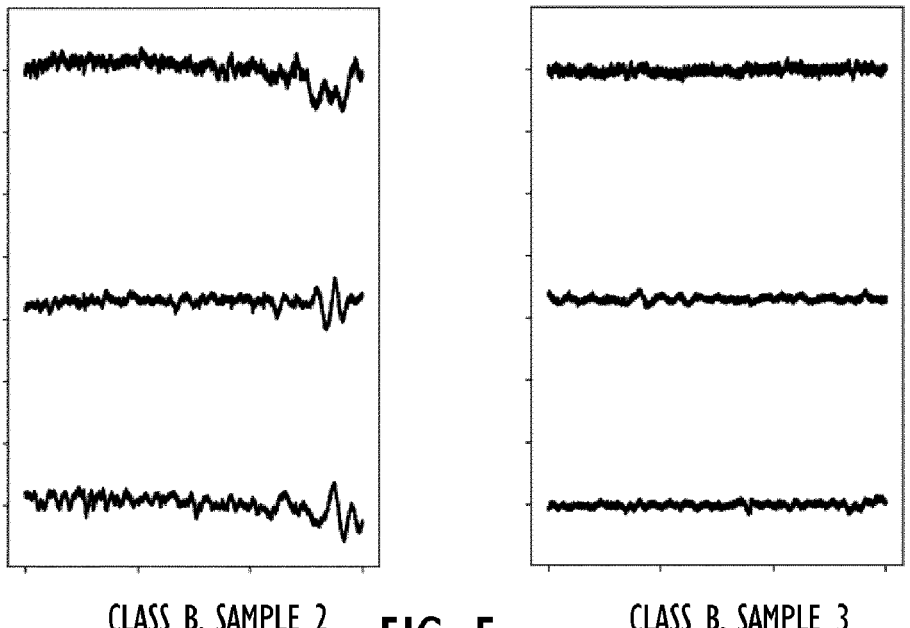


FIG. 5

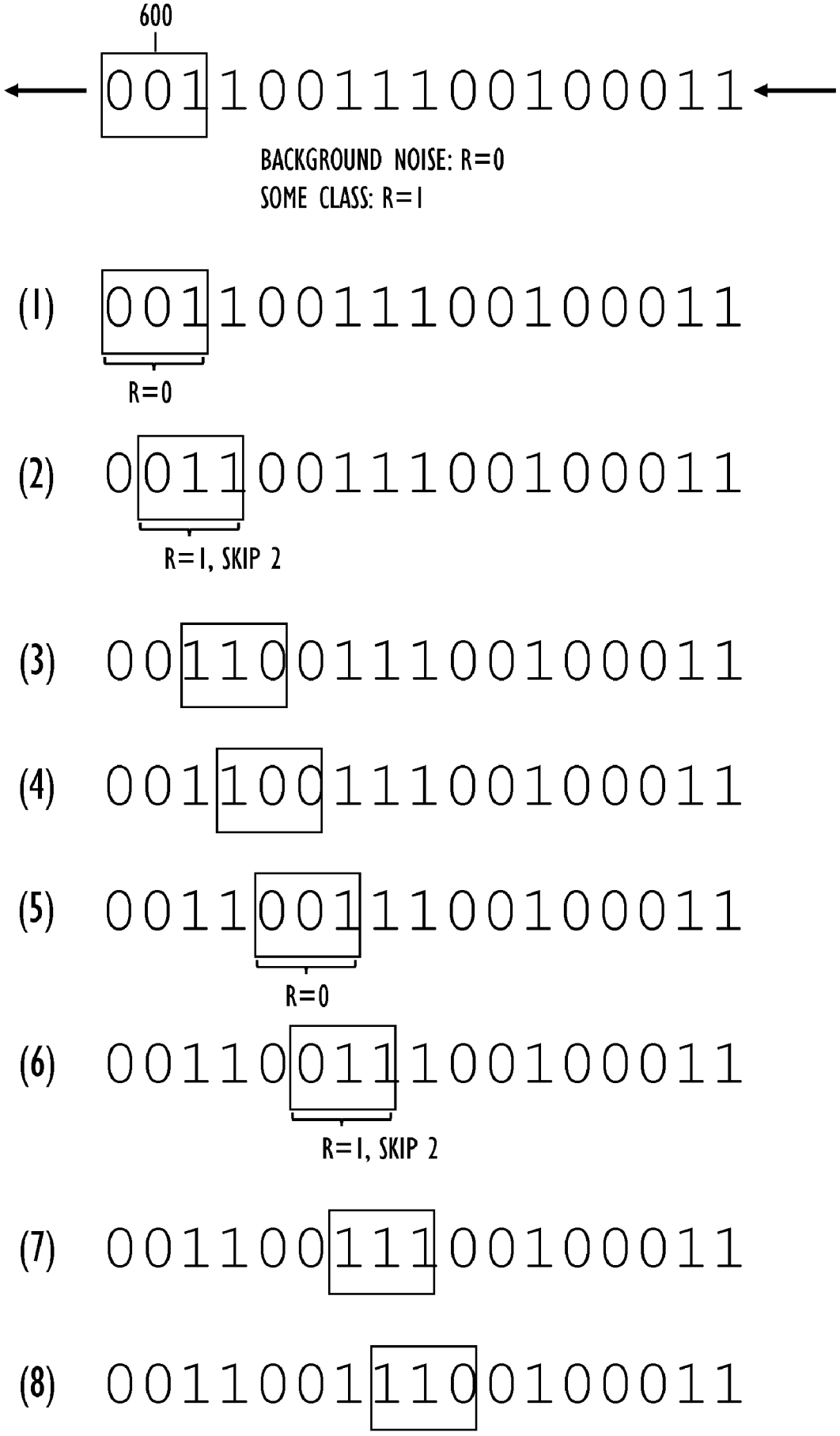


FIG. 6A

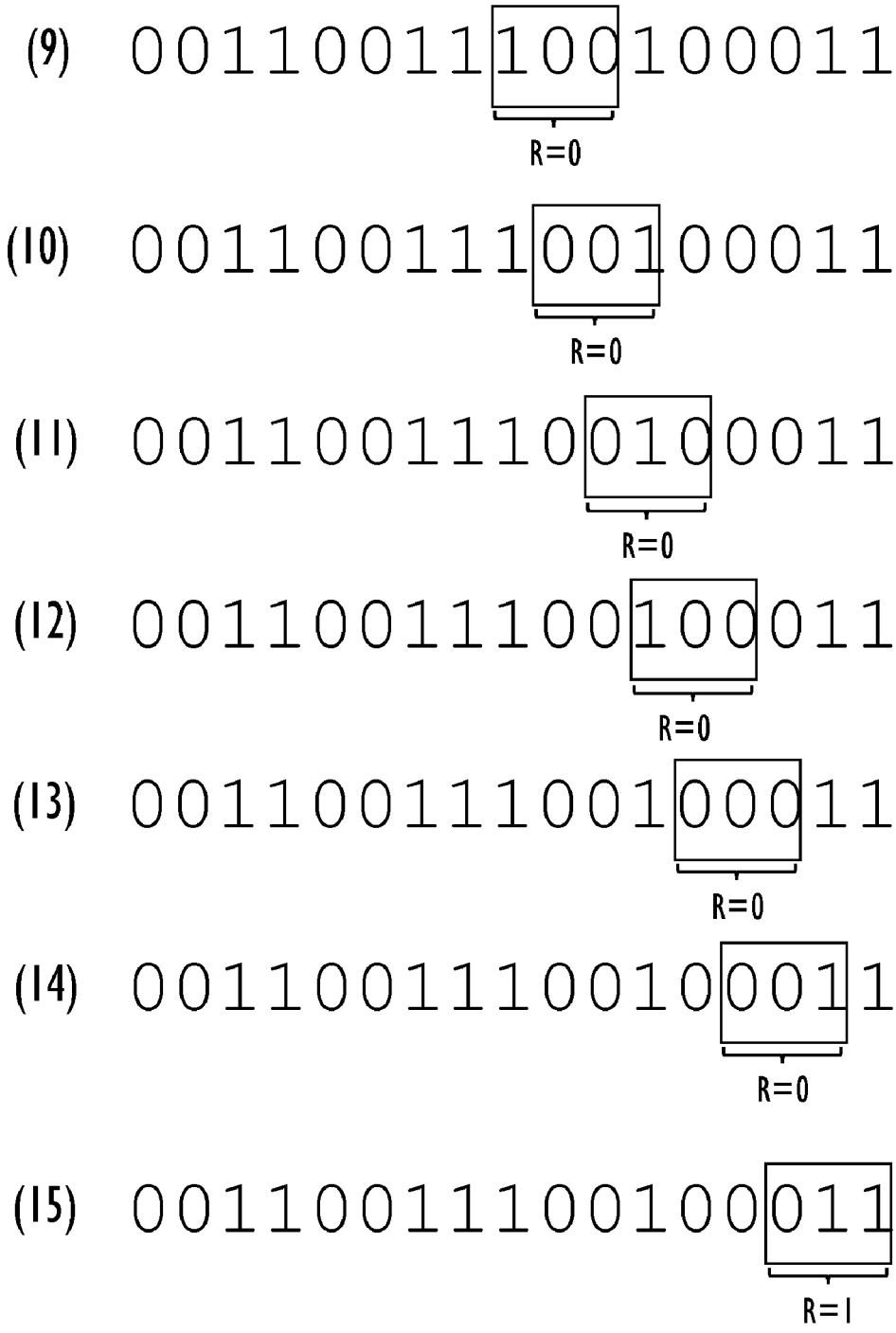


FIG. 6B

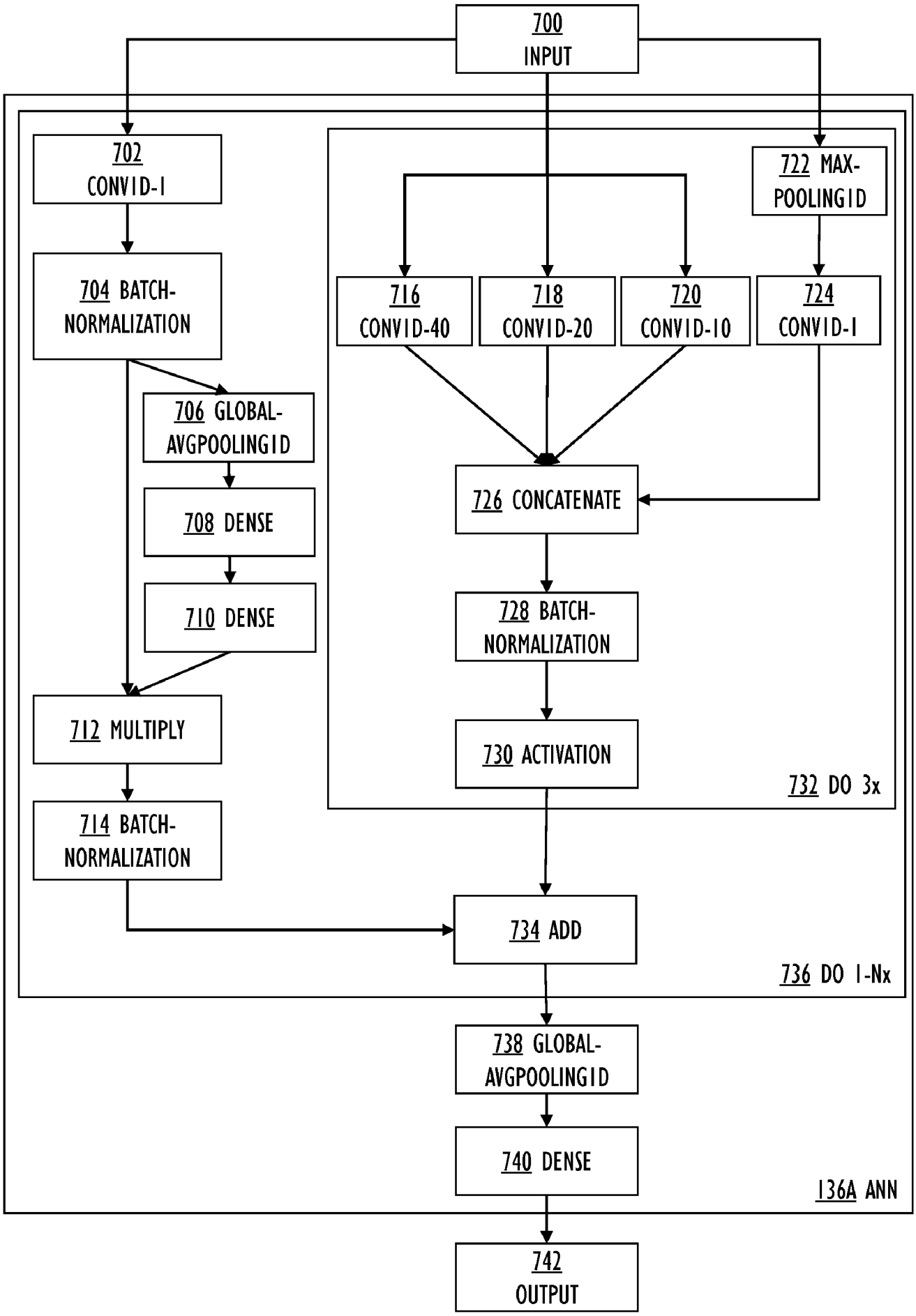


FIG. 7