(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 October 2024 (03.10.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/200071 A1

(51) International Patent Classification:
H04R 1/10 (2006.01)

(21) International Application Number:
PCT/EP2024/057153

(22) International Filing Date:
18 March 2024 (18.03.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
23163974.1 24 March 2023 (24.03.2023) EP

(71) Applicant: **SONY SEMICONDUCTOR SOLUTIONS CORPORATION** [JP/JP]; 4-14-1 Asahi-cho, Atsugi-shi, Kanagawa, 243-0014 (JP).

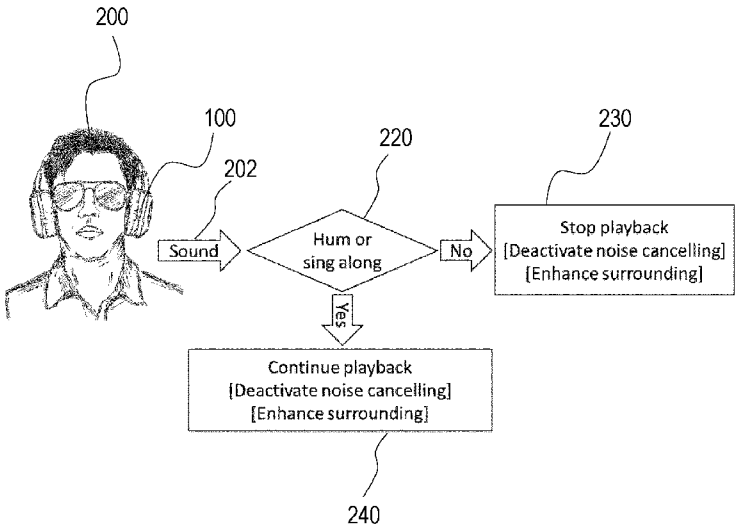
(71) Applicant *(for AL only)*: **SONY EUROPE B.V.** [GB/GB]; The Heights, Brooklands, Weybridge (Surrey) KT13 0XW (GB).

(72) Inventor: **MARKHASIN, Lev**; c/o Sony Europe B.V., Zweigniederlassung Deutschland, Stuttgart Technology Center, Hedelfinger Str. 61, 70327 Stuttgart (DE).

(74) Agent: **2SPL PATENTANWÄLTE PARTG MBB**; Landaubogen 3, 81373 München (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO,

(54) Title: APPARATUSES AND METHODS FOR CONTROLLING A SOUND PLAYBACK OF A HEADPHONE



[.] relate to optional features

FIG. 2

(57) Abstract: The present disclosure relates to a headphone comprising at least one microphone configured to capture a vocal expression of a user of the headphone, a trained machine learning processor configured to classify the user's captured vocal expression into one of a plurality of different vocal expressions, and a controller configured to control sound playback of the headphone based on the classification result of the machine learning processor. The controller may be configured to continue the sound playback in case the classification result of the machine learning processor indicates humming and/or singing as the vocal expression and stop the sound playback in case the classification result of the machine learning processor indicates speaking as the vocal expression.

WO 2024/200071 A1

[Continued on next page]

RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

APPARATUSES AND METHODS FOR CONTROLLING A SOUND PLAYBACK OF A HEADPHONE

Field

The present disclosure generally relates to headphones, and, more particularly, to apparatuses and methods for controlling a sound playback of a headphone depending on a detected sound of vocal expression of a user of the headphone.

Background

Modern headphones may recognize that a user starts to speak so that the playback can be stopped. In addition, some headphones may even deactivate noise cancelling and activate a mode in which surrounding sounds are enhanced. This may enable the user to have a conversation without stopping music or taking off the headset. Often users make sounds by humming along or whistling along a melody or singing along a song. When the headphones stop playback, the user cannot continue enjoying the music. A similar situation arises when the user wants to practice singing or beatboxing and plays some playback via the headphones and when the user starts to sing the playback stops. Some headphones might offer to deactivate the automatic “playback off functionality” but this may be annoying to the user who may wish to have to do as little as possible to switch between modes.

Thus, there may be a need to improve an automatic playback control of headphones.

Summary

According to a first aspect, the present disclosure provides a headphone which comprises at least one microphone configured to capture a vocal expression of a user of the headphone. The headphone further comprises a trained machine learning processor configured to classify the user’s captured vocal expression into one of a plurality of different categories of vocal expressions. The headphone also comprises a controller configured to control sound playback of the headphone based on the classification result of the machine learning processor.

According to a second aspect, the present disclosure provides a method for controlling a sound playback of a headphone. The method includes capturing a vocal expression of a user of the headphone via at least one built-in microphone, classifying, by a built-in trained machine learning processor, the user's captured vocal expression into one of a plurality of different categories of vocal expressions, and controlling the sound playback of the headphone based on the classification result of the machine learning processor.

Thus, the present disclosure proposes headphones with an extended functionality where the headphones not only detect sounds that the user makes but also recognize the type/category of sound. Thus, the proposed headphones may the detect when user sings along or hums and therefore do not stop playback or deactivate noise cancelling.

Brief description of the Figures

Some examples of apparatuses and/or methods will be described in the following by way of example only, and with reference to the accompanying figures, in which

Fig. 1 schematically illustrates a headphone in accordance with the present disclosure;

Fig. 2 show a flowchart method for controlling a sound playback of a headphone according to a first embodiment;

Fig. 3 show a flowchart method for controlling a sound playback of a headphone according to a second embodiment; and

Fig. 4 show a flowchart method for controlling a sound playback of a headphone according to a further embodiment.

Detailed Description

Some examples are now described in more detail with reference to the enclosed figures. However, other possible examples are not limited to the features of these embodiments described in detail. Other examples may include modifications of the features as well as

equivalents and alternatives to the features. Furthermore, the terminology used herein to describe certain examples should not be restrictive of further possible examples.

Throughout the description of the figures same or similar reference numerals refer to same or similar elements and/or features, which may be identical or implemented in a modified form while providing the same or a similar function. The thickness of lines, layers and/or areas in the figures may also be exaggerated for clarification.

When two elements A and B are combined using an “or”, this is to be understood as disclosing all possible combinations, i.e. only A, only B as well as A and B, unless expressly defined otherwise in the individual case. As an alternative wording for the same combinations, "at least one of A and B" or "A and/or B" may be used. This applies equivalently to combinations of more than two elements.

If a singular form, such as “a”, “an” and “the” is used and the use of only a single element is not defined as mandatory either explicitly or implicitly, further examples may also use several elements to implement the same function. If a function is described below as implemented using multiple elements, further examples may implement the same function using a single element or a single processing entity. It is further understood that the terms "include", "including", "comprise" and/or "comprising", when used, describe the presence of the specified features, integers, steps, operations, processes, elements, components and/or a group thereof, but do not exclude the presence or addition of one or more other features, integers, steps, operations, processes, elements, components and/or a group thereof.

Fehler! Kein gültiger Dateiname.

Fig. 1 schematically illustrates a headphone 100 in accordance with the present disclosure. The headphone 100, which may also be referred to as earphone or headset, is a personal audio device which may be worn over the ears or in the ear canal, depending on the implementation. In Fig. 1, headphone 100 is shown as an over-ear headphone, which is a type of headphone that is designed to completely cover the ears. Ear cups 102 of over-ear headphone 100 may be large and padded, and may be connected to a headband 104 that sits over the top of the head. Over-ear headphones may provide a high level of sound isolation, as the ear cups 102 create a seal around the ears that blocks out external noise. Alternatively, headphone 100 may also be implemented as an in-ear headphone, which may also be referred to as earbuds or in-ear monitors. An in-ear headphone is a type of headphone that is

designed to fit directly into the ear canal. Unlike over-ear headphones, in-ear headphones are much smaller and more compact, and they do not have a headband. In-ear headphones are typically made up of a small driver or speaker that sits inside the ear canal. In general, headphones can come in different shapes and sizes, and some models include features such as noise isolation or noise cancellation, which can help block out external sounds for a more immersive listening experience. They are commonly used for listening to music, making phone calls, and for other forms of audio entertainment. Headphone 100 may be a wireless or a wired headphone.

The headphone 100 in accordance with the present disclosure comprises one or more microphones 110 which are configured to capture a vocal expression of a user of the headphone 100. Vocal expression refers to the various ways in which a person may communicate emotions, feelings, and attitudes through their voice. For example, vocal expressions may include speaking, singing, humming, whistling, beatboxing, etc. The microphone 110 may be built into the headphone 100 to capture audio or sound waves from the user's voice. The microphone 110 can be located in different places depending on the design of the headphone 100. For example, the microphone 110 may be located at the ear cups 102 or on a cable that connects to an audio source (e.g., smartphone, tablet, etc.), or it may be located on an adjustable boom arm.

As indicated by 120, headphone 100 further comprises a trained machine learning processor which is configured to classify the user's captured vocal expression into one of a plurality of different categories of vocal expressions. In other words, the trained machine learning processor 120 implements a machine learning algorithm that takes the user's vocal expression or sound as input and processes it to analyze the user's sound. An audio sample captured by microphone 110 may be input into the machine learning processor 120, and the machine learning processor 120 may output a probability score for each category indicating the likelihood that the audio sample belongs to that category of vocal expression. The category with the highest probability score may be chosen as the classification for the audio sample. In some embodiments, the machine learning processor 120 may be configured to classify the user's captured vocal into one of speaking, humming, singing, whistling, beatboxing, etc.

There are several machine learning algorithms that can be used in machine learning processor 120 for classifying the user's captured vocal expression into different categories, including:

- Convolutional Neural Networks (CNNs): CNNs are a type of deep neural network that are well-suited for processing images and other types of multidimensional data, such as sound waves. CNNs are commonly used for sound classification tasks, such as identifying musical genres or detecting anomalies in audio signals.
- Recurrent Neural Networks (RNNs): RNNs are a type of neural network that can be used for processing sequential data, such as audio signals. RNNs are commonly used for speech recognition and other types of natural language processing tasks that involve working with audio data.
- Support Vector Machines (SVMs): SVMs are a type of supervised learning algorithm that can be used for classification tasks. SVMs are often used for sound classification tasks, such as detecting specific sound events or identifying the source of a sound.
- Random Forests: Random forests are an ensemble learning method that combine multiple decision trees to make more accurate predictions. Random forests are often used for audio classification tasks, such as detecting environmental sounds or identifying bird songs.
- Gaussian Mixture Models (GMMs): GMMs are a type of statistical model that can be used for audio classification tasks. GMMs are often used for speaker recognition and other types of audio signal processing tasks that involve working with complex audio data.

There are also many other machine learning algorithms that can be used for classifying sound, and the choice of algorithm depends on the specific task and the characteristics of the audio data.

For example, the trained machine learning processor 120 may be configured for humming and/or singing voice detection based on a Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Network (RNN). This classifier is able to take a past and future temporal context into account to decide on the presence/absence of singing voice, thus using the inherent sequential aspect of a short-term feature extraction in a piece of sound/music. The BLSTM-RNN may contain several hidden layers, so it is able to extract a simple repre-

sentation fitted to our task from low-level features. Such a BLSTM for singing voice detection is described in Simon Leglaive et al.: "Singing Voice Detection with Deep Recurrent Neural Networks".

For example, the machine learning (ML) processor 120 can be trained using a technique called supervised learning. In supervised learning, the ML algorithm is trained on a labeled dataset, where each sound is associated with a specific class or label. The process for training a sound classification algorithm may involve the following steps:

Data collection: Collect a large dataset of sound recordings, with each recording labeled with the correct class or label.

Feature extraction: Extract relevant features from each sound recording, such as frequency, amplitude, and duration.

Data preprocessing: Normalize the extracted features and prepare the data for training by splitting it into training and validation sets.

Model selection: Select an appropriate machine learning model, such as a convolutional neural network (CNN) or a support vector machine (SVM), that is capable of classifying sounds based on their features.

Training: Train the model using the labeled sound dataset and the extracted features.

Validation: Evaluate the performance of the trained model on the validation set to ensure that it is accurately classifying sounds.

Testing: Test the performance of the trained model on a new dataset of sound recordings to evaluate its ability to generalize to new data.

The headphone 100 further comprises a controller (control logic) 130 which is configured to automatically control sound playback of the headphone 100 based on the classification result of the machine learning processor 120, i.e., based on the determined category of the user's vocal expression. Sound or music playback of headphone 100 may be automatically controlled differently depending on whether speaking, humming, whistling, beatboxing, or singing was detected as the user's captured vocal expression by the trained machine learning processor 120. The controller 130 may be a control processor built in the headphone 100. Here, "automatically" may be understood as without manual interaction of the user.

In some implementations, the headphone's controller 130 may be configured to automatically restrict the headphone's sound or music playback in case the classification result of the

machine learning processor 120 indicates "speaking" as the detected category of vocal expression. For example, the sound or music playback may automatically be continued with decreased volume in case it is detected that the user is speaking. For another example, the sound or music playback may automatically be muted, stopped, or interrupted in case it is detected that the user is speaking. To further enable the user to listen to his/her conversation partner(s), the controller 130 may be configured to automatically deactivate an active noise cancelling (ANC) function of the headphone 100. For ANC, headphone 100 may comprise and use additional microphones to detect external noise, and then generate a sound wave that is 180° out of phase with the external noise. When the two sound waves meet, they cancel each other out, effectively reducing or eliminating the background noise. The result is a quieter environment that allows the user to hear audio more clearly, even at lower volume levels.

Additionally, the controller 130 may be configured to automatically enhance an active hear-through function of the headphone 100. For example, an active hear through function of headphone 100 may automatically be activated in case the classification result of the machine learning processor 120 indicates "speaking" as the category of vocal expression. Hear through refers to a feature that allows the user to hear ambient sounds from their surroundings while still wearing the headphones 100. This feature is also sometimes referred to as ambient sound mode, transparency mode, or pass-through mode. Hear through may be achieved by using external microphones on the headphones 100 to pick up sound from the user's environment, and then playing that sound through the headphones 100. This allows the user to hear important sounds like traffic, announcements, or conversations.

If the classification result of the machine learning processor 120 indicates vocal expressions other than speaking, e.g. humming, singing, or other rhythmic or melodic user sounds, the headphone's controller 130 may be configured to continue the headphone's audio or music playback. In some implementations, it is conceivable that the headphone's audio or music playback is continued with automatically decreased volume such that the user can better listen to his/her humming or singing. Additionally or alternatively, the controller 130 may be configured to automatically deactivate the headphone's ANC function such that the user can better listen to his/her humming or singing. Additionally or alternatively, the controller 130 may be configured to automatically enhance a hear-through function of the headphone 100 such that the user can better listen to his/her humming or singing.

This scenario is illustrated in **Fig. 2**. A user 200 wearing headphone 100 utters sound (vocal expression) 202. This vocal expression 202 of the user 200 is captured by built-in microphone 110 and input to built-in machine learning processor 120. At 220, machine learning processor 120 classifies captured vocal expression 202 into one of two categories: i) speaking, ii) humming or singing along (could also include whistling and/or beatboxing). If the category is i) speaking, controller 130 stops playback at 230. Optionally, controller 130 deactivates ANC and enhances surrounding sounds such that user 200 can better listen to conversation partners. If the category is ii) humming or singing along, controller 130 continues playback at 240. Optionally, controller 130 deactivates ANC and or enhances surrounding sounds such that user 200 can better listen to him-/herself.

In some example implementations of headphone 100, controller 130 or associated processing circuitry may be configured to automatically merge the user's vocal expression (user's captured voice) into the headphone's sound or audio playback (e.g., music) in case the classification result of the machine learning processor 120 indicates a melodic vocal expression such as humming and/or singing. In this way, the user can better hear him-/herself humming or singing. Two audio signals (e.g., user's voice and music playback) can be merged using a process called audio mixing. Audio mixing involves combining multiple audio signals into a single output signal (merged playback signal), which can be played in real-time. Optionally, the controller 130 may be configured to activate an ANC function of the headphone 100 such that the merged sound is not disturbed by background noise and the user can better listen to him-/herself.

Before merging the user's vocal expression (user's voice) into the headphone's audio or music playback, the controller 130 or associated processing circuitry may be configured to separate the user's captured vocal expression from surrounding or background sounds (singing voice separation) and merge the user's separated vocal expression into the (music) playback. Singing voice separation may be more challenging than speech separation because singing involves a wider range of pitch and tone variations. An example approach to singing voice separation may be to use a variant of a neural network called a deep recurrent neural network (DRNN) or a long short-term memory (LSTM) network. These networks can capture the temporal dynamics of singing or humming by modeling long-term dependencies in the input signal. Similar to speech separation, the neural network may be trained using a

dataset of mixed audio signals and their corresponding target singing voice signals. Once trained, the neural network can be used to separate the user's singing voice from background sounds in real-time. The person skilled in the art will appreciate that the accuracy of singing voice separation is dependent on the complexity of the surrounding sound and the quality of the input audio signal captured by microphone 110.

For example, Ali Aroudi et al.: "DBnet: DOA-Driven Beamforming Network for end-to-end farfield sound source separation" propose a direction-of-arrival-driven beamforming network (DBnet) consisting of direction-of-arrival (DOA) estimation and beamforming layers for end-to-end source separation. The person skilled in the art will appreciate that DOA estimation may require more than one microphone 110 in the headphone 100. DBnet may be implemented as convolutional-recurrent structure and trained using loss functions that are based on the distances between the separated speech signals and the target speech signals, without a need for the ground-truth DOAs of speakers. To improve the source separation performance, end-to-end extensions of DBnet may be used which incorporate post masking networks.

The separation of the user's sound from surrounding signals is illustrated in **Fig. 3**. The user 200 wearing headphone 100 utters sound (vocal expression) 202. This vocal expression 202 is captured by one, preferably a plurality of built-in microphones 110 and input to built-in machine learning processor 120. At 220, machine learning processor 120 classifies captured vocal expression 202 into one of two categories: i) speaking, ii) melodic vocal expression such as humming or singing along. If the category is i) speaking, controller 130 stops playback at 230, Optionally, controller 130 deactivates ANC and enhances surrounding sounds such that user 200 can better listen to conversation partners. If the category is ii) humming or singing along, controller 130 causes the separation of the user's vocal expression (voice) 202 from surrounding sounds at 340. As explained above, this may be performed via an additional machine learning network (not shown), for example. At 350, controller 130 causes the playback to continue by merging or mixing the separated vocal expression 202 into the playback (e.g., music). Here, controller 130 need not deactivate ANC and or enhance surrounding sounds. In other words, ANC may be kept activated.

The proposed concept is summarized in **Fig. 4**.

Fig. 4 shows a flowchart of a method for controlling a sound playback of a headphone 100. Method 400 includes capturing 410 a vocal expression of a user of the headphone 100 via at least one built-in microphone 110. Method 400 further includes classifying, by a built-in trained machine learning processor 120, the user's captured vocal expression into one of a plurality of different vocal expressions (e.g., speaking and non-speaking). Method 400 further includes controlling the sound playback of the headphone based on the classification result of the machine learning processor 120.

Embodiments of the present disclosure provide an opportunity to the user to hum or sing along while listening to music without having the headphone's playback stopped.

In the following, some examples of the proposed concept are presented:

An example (e.g., example 1) relates to a headphone comprising at least one microphone which configured to capture a vocal expression of a user of the headphone, a trained machine learning processor configured to classify the user's captured vocal expression into one of a plurality of different vocal expressions, and a controller configured to control sound playback of the headphone based on the classification result of the machine learning processor.

In another example (e.g., example 2) relating to a previous example (e.g., example 1) or to any other example, the machine learning processor is configured to classify the user's captured vocal into one of speaking, humming, and singing.

In another example (e.g., example 3) relating to a previous example (e.g., one of the examples 1 or 2) or to any other example, the controller is configured to restrict (e.g., stop) the sound playback in case the classification result of the machine learning processor indicates speaking as the vocal expression.

In another example (e.g., example 4) relating to a previous example (e.g., example 3) or to any other example, the controller is configured to deactivate a (active) noise cancelling function of the headphone.

In another example (e.g., example 5) relating to a previous example (e.g., one of the examples 3 or 4) or to any other example, the controller is configured to enhance a (active) hear-through function of the headphone.

In another example (e.g., example 6) relating to a previous example (e.g., one of the examples 1 to 5) or to any other example, the controller is configured to continue the sound playback in case the classification result of the machine learning processor indicates humming and/or singing as the vocal expression.

In another example (e.g., example 7) relating to a previous example (e.g., example 6) or to any other example, the controller is configured to deactivate a (active) noise cancelling function of the headphone.

In another example (e.g., example 8) relating to a previous example (e.g., one of the examples 6 or 7) or to any other example, the controller is configured to enhance a (active) hear-through function of the headphone.

In another example (e.g., example 9) relating to a previous example (e.g., one of the examples 1 to 8) or to any other example, the controller is configured to merge or mix the user's vocal expression into the sound playback of the headphone in case the classification result of the machine learning processor indicates humming and/or singing as the vocal expression.

In another example (e.g., example 10) relating to a previous example (e.g., example 9) or to any other example, the controller is configured to activate a (active) noise cancelling function of the headphone.

In another example (e.g., example 11) relating to a previous example (e.g., one of the examples 9 or 10) or to any other example, the controller is configured to separate the user's vocal expression from background sounds and merge or mix the user's separated vocal expression into the sound playback.

An example (e.g., example 12) relates to a method for controlling a sound playback of a headphone, the method comprising capturing a vocal expression of a user of the headphone

via at least one built-in microphone, classifying, by a built-in trained machine learning processor, the user's captured vocal expression into one of a plurality of different vocal expressions, and controlling the sound playback of the headphone based on the classification result of the machine learning processor.

The aspects and features described in relation to a particular one of the previous examples may also be combined with one or more of the further examples to replace an identical or similar feature of that further example or to additionally introduce the features into the further example.

Examples may further be or relate to a (computer) program including a program code to execute one or more of the above methods when the program is executed on a computer, processor or other programmable hardware component. Thus, steps, operations or processes of different ones of the methods described above may also be executed by programmed computers, processors or other programmable hardware components. Examples may also cover program storage devices, such as digital data storage media, which are machine-, processor- or computer-readable and encode and/or contain machine-executable, processor executable or computer-executable programs and instructions. Program storage devices may include or be digital storage devices, magnetic storage media such as magnetic disks and magnetic tapes, hard disk drives, or optically readable digital data storage media, for example. Other examples may also include computers, processors, control units, (field) programmable logic arrays ((F)PLAs), (field) programmable gate arrays ((F)PGAs), graphics processor units (GPU), application-specific integrated circuits (ASICs), integrated circuits (ICs) or system-on-a-chip (SoCs) systems programmed to execute the steps of the methods described above.

It is further understood that the disclosure of several steps, processes, operations or functions disclosed in the description or claims shall not be construed to imply that these operations are necessarily dependent on the order described, unless explicitly stated in the individual case or necessary for technical reasons. Therefore, the previous description does not limit the execution of several steps or functions to a certain order. Furthermore, in further examples, a single step, function, process or operation may include and/or be broken up into several sub-steps, -functions, -processes or -operations.

If some aspects have been described in relation to a device or system, these aspects should also be understood as a description of the corresponding method. For example, a block, device or functional aspect of the device or system may correspond to a feature, such as a method step, of the corresponding method. Accordingly, aspects described in relation to a method shall also be understood as a description of a corresponding block, a corresponding element, a property or a functional feature of a corresponding device or a corresponding system.

The following claims are hereby incorporated in the detailed description, wherein each claim may stand on its own as a separate example. It should also be noted that although in the claims a dependent claim refers to a particular combination with one or more other claims, other examples may also include a combination of the dependent claim with the subject matter of any other dependent or independent claim. Such combinations are hereby explicitly proposed, unless it is stated in the individual case that a particular combination is not intended. Furthermore, features of a claim should also be included for any other independent claim, even if that claim is not directly defined as dependent on that other independent claim.

Claims

1. A headphone, comprising

at least one microphone configured to capture a vocal expression of a user of the headphone;

a trained machine learning processor configured to classify the user's captured vocal expression into one of a plurality of different vocal expressions; and

a controller configured to control sound playback of the headphone based on the classification result of the machine learning processor.

2. The headphone of claim 1, wherein the machine learning processor is configured to classify the user's captured vocal into one of speaking, humming, and singing.

3. The headphone of claim 1, wherein the controller is configured to restrict the sound playback in case the classification result of the machine learning processor indicates speaking as the vocal expression.

4. The headphone of claim 3, wherein the controller is configured to deactivate a noise cancelling function of the headphone.

5. The headphone of claim 3, wherein the controller is configured to enhance a hear-through function of the headphone.

6. The headphone of claim 1, wherein the controller is configured to continue the sound playback in case the classification result of the machine learning processor indicates humming and/or singing as the vocal expression.

7. The headphone of claim 6, wherein the controller is configured to deactivate a noise cancelling function of the headphone.

8. The headphone of claim 6, wherein the controller is configured to enhance a hear-through function of the headphone.

9. The headphone of claim 1, wherein the controller is configured to merge the user's vocal expression into the sound playback of the headphone in case the classification result of the machine learning processor indicates humming and/or singing as the vocal expression.

10. The headphone of claim 9, wherein the controller is configured to activate a noise cancelling function of the headphone.

11. The headphone of claim 9, wherein the controller is configured to separate the user's vocal expression from background sounds and merge the user's separated vocal expression into the sound playback.

12. A method for controlling a sound playback of a headphone, the method comprising

capturing a vocal expression of a user of the headphone via at least one built-in microphone;

classifying, by a built-in trained machine learning processor, the user's captured vocal expression into one of a plurality of different vocal expressions; and

controlling the sound playback of the headphone based on the classification result of the machine learning processor.

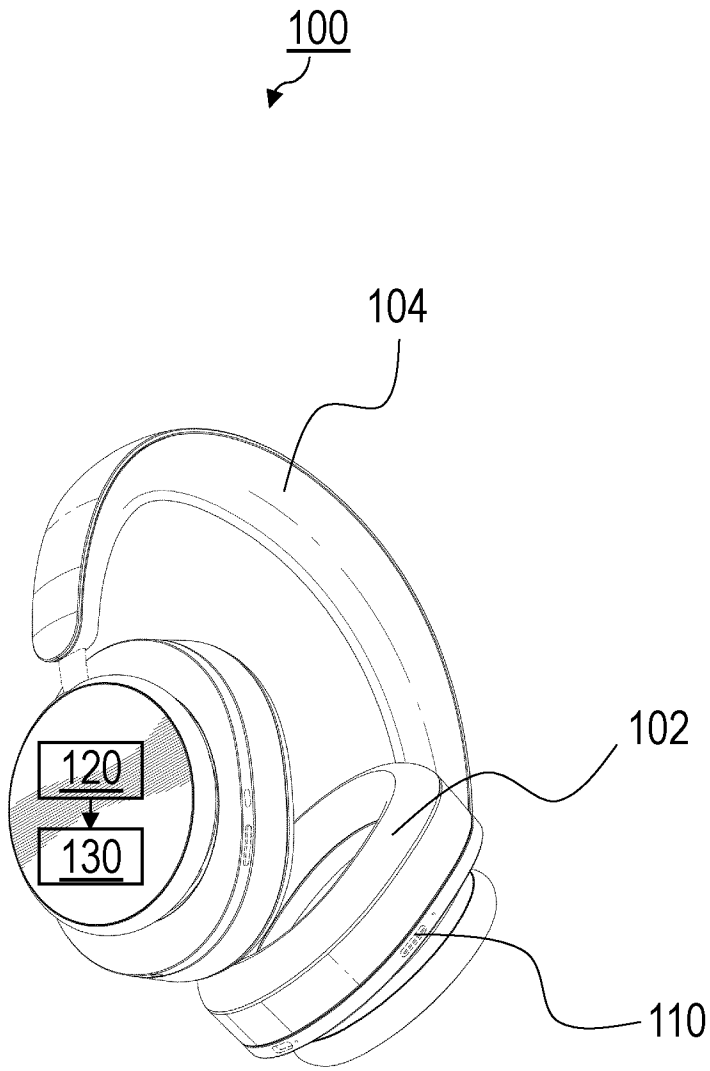
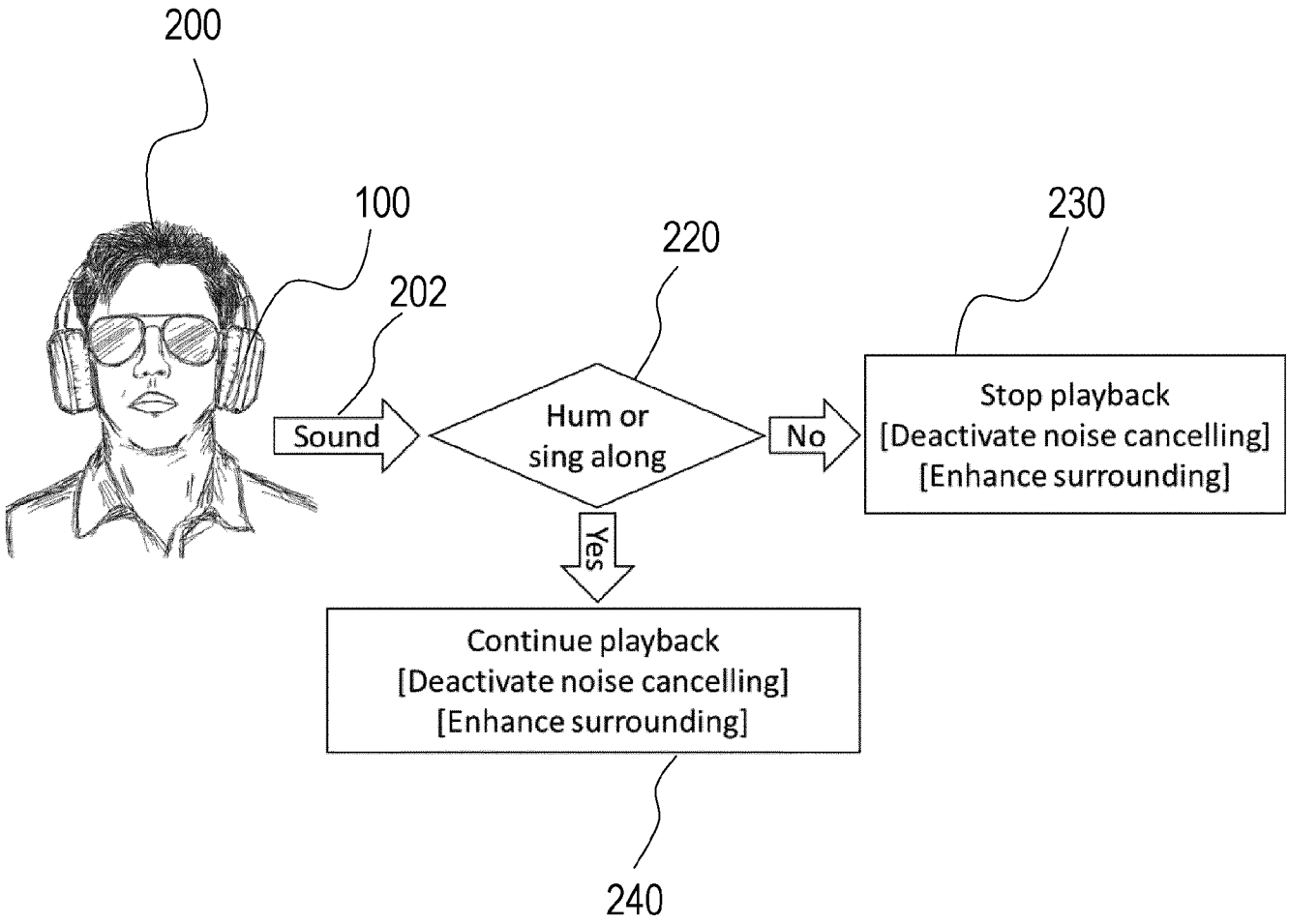
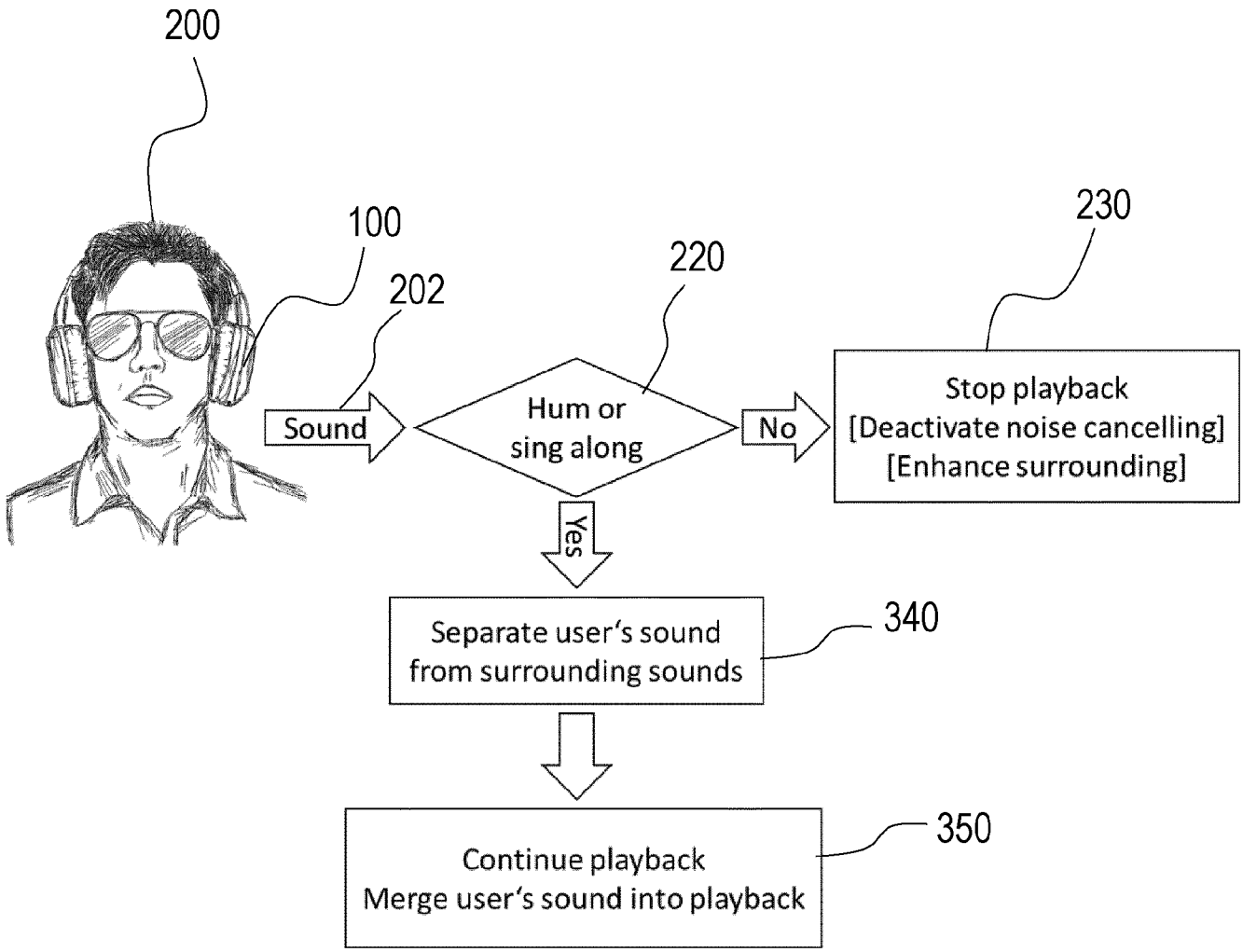


FIG. 1



[.] relate to optional features

FIG. 2



[.] relate to optional features

FIG. 3

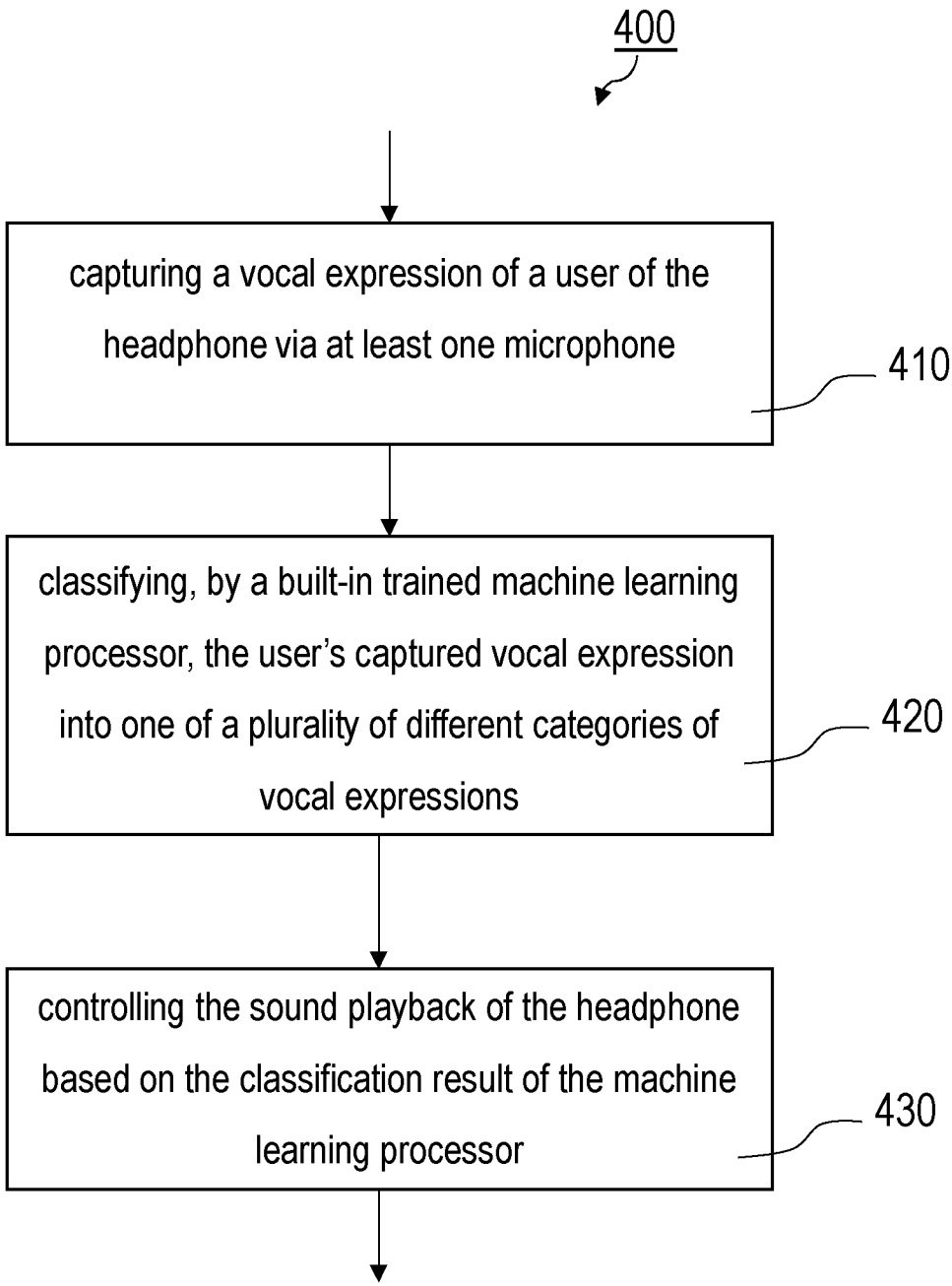

400
capturing a vocal expression of a user of the headphone via at least one microphone
410
classifying, by a built-in trained machine learning processor, the user's captured vocal expression into one of a plurality of different categories of vocal expressions
420
controlling the sound playback of the headphone based on the classification result of the machine learning processor
430

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/057153

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04R1/10
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04R G10K G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2022/366932 A1 (LESSO JOHN P [GB] ET AL) 17 November 2022 (2022-11-17) paragraph [0042] - paragraph [0043]; figures 1-3 paragraph [0049] - paragraph [0055]; figure 4 paragraph [0065]; figure 6 paragraph [0067]	1,2,6-12
X	US 2022/303688 A1 (ZYSKOWSKI JAMIE ALEXANDER [US] ET AL) 22 September 2022 (2022-09-22) paragraph [0034]; figure 1 paragraphs [0037], [0041], [0053]; figure 2 paragraph [0080] - paragraph [0085]; figure 3	1-5,12

-/--

[x] Further documents are listed in the continuation of Box C.

[x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 May 2024	15/05/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Betgen, Benjamin

1

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017/318374 A1 (DOLENC ANDRÉ [FI] ET AL) 2 November 2017 (2017-11-02) paragraph [0021] - paragraph [0022]	5,9-11
A	US 2015/195641 A1 (DI CENSO DAVIDE [US] ET AL) 9 July 2015 (2015-07-09) paragraph [0049]; figure 5 paragraph [0054]; figure 6	10,11

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2024/057153

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2022366932 A1	17-11-2022	GB 2616166 A	30-08-2023
		US 2022223168 A1	14-07-2022
		US 2022366932 A1	17-11-2022
		WO 2022153022 A1	21-07-2022
US 2022303688 A1	22-09-2022	US 2022116707 A1	14-04-2022
		US 2022303688 A1	22-09-2022
US 2017318374 A1	02-11-2017	US 2017318374 A1	02-11-2017
		WO 2017192365 A1	09-11-2017
US 2015195641 A1	09-07-2015	CN 106062746 A	26-10-2016
		EP 3092583 A1	16-11-2016
		JP 6600634 B2	30-10-2019
		JP 2017507550 A	16-03-2017
		KR 20160105858 A	07-09-2016
		US 2015195641 A1	09-07-2015
		WO 2015103578 A1	09-07-2015