



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
13.12.2017 Bulletin 2017/50

(51) Int Cl.:
H04S 7/00 (2006.01)

(21) Application number: **16173264.9**

(22) Date of filing: **07.06.2016**

(84) Designated Contracting States:
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR**
Designated Extension States:
BA ME
Designated Validation States:
MA MD

(72) Inventors:
• **ERONEN, Antti**
33820 Tampere (FI)
• **LEHTINIEMI, Arto**
33880 Lempäälä (FI)

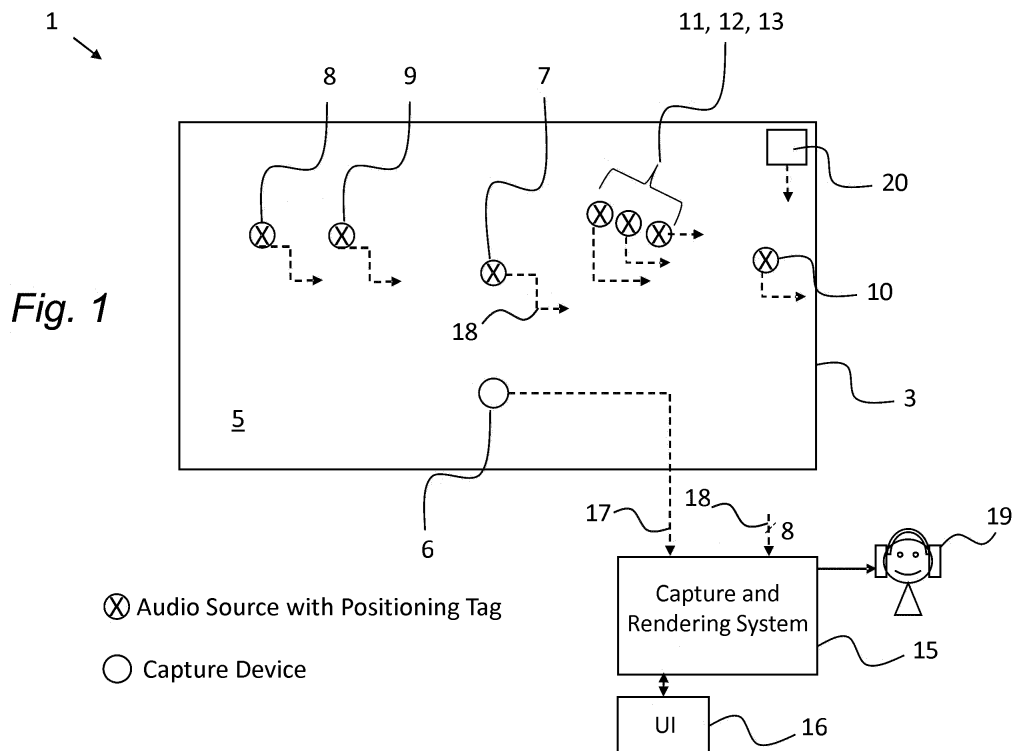
(74) Representative: **Nokia EPO representatives**
Nokia Technologies Oy
Karaportti 3
02610 Espoo (FI)

(71) Applicant: **Nokia Technologies Oy**
02610 Espoo (FI)

(54) **DISTRIBUTED AUDIO MIXING**

(57) An apparatus comprises at least one processor and at least one memory including computer program code, the at least one memory and the computer program code configured with the processor to cause the apparatus to perform the following steps. In one step, the apparatus receives a plurality of audio signals representing captured audio from respective audio sources in a space. In another step, the apparatus processes the received

audio signals to generate a spatial audio signal for playback. The processing comprises identifying a subset of the audio sources, and responsive to a detected change in spatial position of one or more of said subset of audio sources, generating the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.



Description

Field

[0001] This relates to distributed audio mixing. The further relates to, but it not limited to, methods and apparatus for distributed audio capture, mixing and rendering of spatial audio signals to enable spatial reproduction of audio signals.

Background

[0002] Spatial audio refers to playable audio data that exploits sound localisation. In a real world space, for example in a concert hall, there will be multiple audio sources, for example the different members of an orchestra or band, located at different locations on the stage. The location and movement of the sound sources is a parameter of the captured audio. In rendering the audio as spatial audio for playback such parameters are incorporated in the data using processing algorithms so that the listener is provided with an immersive and spatially oriented experience.

[0003] Nokia's Spatial Audio Capture (SPAC) is an example technology for processing audio captured via a microphone array into spatial audio; that is audio with a spatial percept. The intention is to capture audio so that when it is rendered to a user the user will experience the sound field as if they are present at the location of the capture device.

[0004] An example application of spatial audio is in virtual reality (VR) whereby both video and audio data is captured within a real world space. In the rendered version of the space, i.e. the virtual space, the user, through a VR headset, can view and listen to the captured video and audio which has a spatial percept.

[0005] Rendering of captured audio can be complex and the quality of spatial audio produced in terms of listener experience can be degraded in current systems, for example in situations where one or more audio sources change position or move during capture.

Summary

[0006] According to one aspect, a method comprises: receiving a plurality of audio signals representing captured audio from respective audio sources in a space; processing the received audio signals to generate a spatial audio signal, in which the processing comprises: detecting spatial positions of the audio sources; and responsive to a detected change in spatial position of one or more of said audio sources, generating the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.

[0007] The method may further comprise: defining a group of audio sources from the plurality of audio signals; and wherein the change in the spatial position in the gen-

erated spatial audio signal depends on the position of one or more of the other audio sources in said group of audio sources.

[0008] The method may further comprise receiving a spatial audio signal associated with a microphone array configured to provide spatial audio capture, and wherein the processing comprises generating an updated version of the spatial audio signal associated with the microphone array.

[0009] The processing may further comprise identifying a subset of the audio sources, and the detecting and generating steps are performed only in relation to the subset. The method may further comprise identifying the subset manually through a user interface. Identifying the subset may be performed automatically based on attributes of the audio sources. Identifying the subset may be based on the position of one or more sources in the space. Identifying the subset may be based on the position of the one or more sources relative to a primary source whose position is permitted to change in the signal. Identifying the subset may be based on the position of the one or more sources relative to the position of the microphone array. The method may further comprise receiving for each audio source positional data received from, or derived from, a positioning tag carried by the audio source. Identifying the subset may be determined based on the type of audio source. The type of audio source may be determined by signal analysis of the audio signals from each audio source. Analysis of the audio signals may identify a type of vocal performance.

[0010] Analysis of the audio signals may identify a type of instrument.

[0011] Determination of the audio source type may be performed by receiving identification data from a tag associated with the audio source.

[0012] Generating the spatial audio signal may comprise controlling the change in spatial position in accordance with one or more position modification rules. Selection of a position modification rule may be received from a user through a user interface. The user interface may display a spatial representation of each audio source together with the one or more selectable position modification rules for each. The, or one of said position modification rules may prevent a change in spatial position. The, or one of said position modification rules may comprise calculating an average detected change in spatial position over a predetermined time interval and wherein the change in spatial position in the signal may be based on the calculated average.

[0013] Identifying the subset may be determined by first identifying a group of associated audio sources based on spatial position, identifying a substantially common or average change in position for members of the group, and responsive to detecting a deviation of one or more members of the group from the average change in position, generating the spatial audio signal such that said deviation is not present therein. Generating the spatial audio signal may comprise changing the position of

the deviating members to substantially track the common or average change in position. Generating the spatial audio signal may comprise setting the change in position at a single location, substantially at the centre of the group. Generating the spatial audio signal may comprise setting the change in position for all members of the group to be equidistant from the centre of the group.

[0014] According to a second aspect, an apparatus comprises at least one processor and at least one memory including computer program code, the at least one memory and the computer program code configured with the processor to cause the apparatus to: receive a plurality of audio signals representing captured audio from respective audio sources in a space; process the received audio signals to generate a spatial audio signal, in which the processing comprises: detecting spatial positions of the audio sources; and responsive to a detected change in spatial position of one or more of said audio sources, generate the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.

[0015] The processor may be configured to define a group of audio sources from the plurality of audio signals, and wherein the change in the spatial position in the generated spatial audio signal depends on the position of one or more of the other audio sources in said group of audio sources.

[0016] The processor may further cause the apparatus to receive a spatial audio signal associated with a microphone array configured to provide spatial audio capture, and to generate an augmented version of the spatial audio signal associated with the microphone array.

[0017] The apparatus may identify the subset is performed by receiving identification data entered manually through a user interface. Identifying the subset may be performed automatically based on attributes of the audio sources. Identifying the subset may be based on the position of one or more sources in the space. Identifying the subset may be based on the position of the one or more sources relative to a primary source whose position is permitted to change in the signal. Identifying the subset may be based on the position of the one or more sources relative to the position of the microphone array.

[0018] The processor may be further configured to receive for each audio source positional data received from, or derived from, a positioning tag carried by the audio source.

[0019] Identifying the subset may be determined based on the type of audio source. The type of audio source may be determined by means of the processor performing signal analysis of the audio signals from each audio source. The analysis of the audio signals may identify a type of vocal performance. The analysis of the audio signals may identify a type of instrument. Determination of the audio source type may be performed by receiving identification data from a tag associated with the audio source.

[0020] The processor may be arranged to generate the spatial audio signal by controlling the change in spatial position in accordance with one or more position modification rules.

5 **[0021]** Selection of the position modification rule may be received from a user through a user interface provided by the processor.

[0022] The processor may be configured to display a spatial representation of each audio source together with the one or more selectable position modification rules for each.

[0023] The, or one of said position modification rules may prevent a change in spatial position.

10 **[0024]** The, or one of said position modification rules may comprise calculating an average detected change in spatial position over a predetermined time interval and wherein the change in spatial position in the signal is based on the calculated average.

15 **[0025]** The processor may be configured to identify the subset by first identifying a group of associated audio sources based on spatial position, to identify a substantially common or average change in position for members of the group, and responsive to detecting a deviation of one or more members of the group from the average change in position, to generate the spatial audio signal such that said deviation is not present therein.

20 **[0026]** The processor may be configured to generate the spatial audio signal by changing the position of the deviating members to substantially track the common or average change in position.

25 **[0027]** The processor may be configured to generate the spatial audio signal by setting the change in position at a single location, substantially at the centre of the group.

30 **[0028]** The processor may be configured to generate the spatial audio signal by setting the change in position for all members of the group to be equidistant from the centre of the group.

35 **[0029]** The apparatus may be further configured to receive spatial video signals in association with the audio signals.

40 **[0030]** The processor may be configured to identify the subset of audio sources based on one or both of the audio signals and the video signals.

45 **[0031]** According to a third aspect, an apparatus comprises at least one processor and at least one memory including computer program code, the at least one memory and the computer program code configured with the processor to cause the apparatus to: receive a plurality of audio signals representing captured audio from respective audio sources in a space; process the received audio signals to generate a spatial audio signal for playback, in which the processing comprises: detecting spatial positions of the audio sources; and responsive to a detected change in spatial position of one or more of said audio sources, generating the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for

the respective audio sources.

[0032] According to a fourth aspect, a computer program comprises instructions that when executed by a computer apparatus control it to perform the method of any preceding definition

[0033] According to a fifth aspect, a non-transitory computer-readable storage medium is provided having stored thereon computer-readable code, which, when executed by at least one processor, causes the at least one processor to perform a method, comprising: receiving a plurality of audio signals representing captured audio from respective audio sources in a space; processing the received audio signals to generate a spatial audio signal for playback, in which the processing comprises: detecting spatial positions of the audio sources; and responsive to a detected change in spatial position of one or more of said audio sources, generating the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.

[0034] According to a sixth aspect, apparatus is provided, the apparatus having at least one processor and at least one memory having computer-readable code stored thereon which when executed controls the at least one processor: to receive a plurality of audio signals representing captured audio from respective audio sources in a space; to process the received audio signals to generate a spatial audio signal for playback, in which the processing comprises:

to detect spatial positions of the audio sources; and responsive to a detected change in spatial position of one or more of said audio sources, to generate the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.

Brief Description of the Drawings

[0035] Embodiments will now be described, by way of non-limiting example, with reference to the accompanying drawings, in which:

Figure 1 is a schematic representation of a distributed audio capture scenario, including use of a rendering apparatus according to embodiments;
Figure 2 is a schematic diagram illustrating components of the Figure 1 rendering apparatus;
Figure 3 is a flow diagram showing method steps of audio capture and rendering according to an embodiment;
Figure 4a is a graphical representation of a real-world space in which audio sources may move;
Figure 4b is a screenshot of a User Interface (UI) indicating resulting filtered movement in the spatial sound field resulting from use of embodiments herein;

Figure 5 is a flow diagram showing method steps in one example filtering operation;

Figure 6 is a screenshot of a UI indicating how a user can select the Figure 5 filtering;

Figure 7 is a screenshot of a UI indicating another example filtering operation;

Figure 8 is a flow diagram showing method steps in the Figure 7 filtering operation;

Figure 9 is a screenshot of a UI indicating another example filtering operation;

Figure 10 is a flow diagram showing method steps in the Figure 9 filtering operation;

Figure 11 is a screenshot of a UI indicating another example filtering operation;

Figures 12a and 12b are schematic representations indicating different methods of filtering movement in virtual space for associated groups of audio sources; Figure 13 is a schematic diagram indicating functional components used in capture and rendering stages of an apparatus according to embodiments;

Figure 14 is an overview of an example method of determining characteristics of an audio signal, which may be useful in the method steps of any embodiment herein;

Figure 15 is a flowchart of a method according to Figure 14; and

Figure 16 is an overview of a process for obtaining multiple types of acoustic features in the method of claim Figure 15.

Detailed Description of Embodiments

[0036] Embodiments herein relate to systems and methods relating to the capture and rendering of spatial audio data for playback. In particular, the embodiments relate to distributed spatial and rendering methods in which there are multiple audio sources which may move within a virtual space over time. Each audio source generates respective audio signals and, in some embodiments, positioning information for use by the system. An example application is in a VR capture and rendering system in which video is also captured and rendered to provide an immersive user experience. Nokia's OZO (RTM) VR camera is used as an example of a VR capture device which comprises a microphone array to provide a spatial audio signal, but it will be appreciated that the embodiments are not limited to VR applications nor the use of microphone arrays at the capture point.

[0037] Referring to Figure 1, an overview of a VR capture scenario 1 is shown together with a first embodiment capture and rendering system (CRS) 15 with associated user interface 16. The Figure shows in plan-view a real world space 3 which may be for example a concert hall or other music venue. The CRS 15 is applicable to any real world space, however. A VR device 6 for video and spatial audio capture is supported on a floor 5 of the space 3 in front of multiple audio sources, in this case a band; the position of the VR device 6 is known, e.g. through

predetermined positional data or signals derived from a positioning tag on the VR device (not shown). The VR device 6 comprises a microphone array configured to provide spatial audio capture.

[0038] The band may be comprised of multiple members each of which has an associated external microphone or (in the case of guitarists) a pick-up feed providing audio signals. Each may therefore be termed an audio source for convenience. In other embodiments, other types of audio source can be used. The audio sources in this case comprise a lead vocalist 7, a drummer 8, lead guitarist 9, bass guitarist 10, and three members of a choir or backing singers 11, 12, 13 which members are spatially close together in a group.

[0039] As well as having an associated microphone or audio feed, the audio sources 7 - 13 carry a positioning tag which can be any module capable of indicating through data its respective spatial position to the CRS 15. For example the positioning tag may be a high accuracy indoor positioning (HAIP) tag which works in association with one or more HAIP locators 20 within the space 3. HAIP systems use Bluetooth Low Energy (BLE) communication between the tags and the one or more locators 20. For example, there may be four HAIP locators mounted on, or placed relative to, the VR device 6. A respective HAIP locator may be to the front, left, back and right of the VR device 6. Each tag sends BLE signals from which the HAIP locators derive the tag, and therefore, audio source location.

[0040] In general, such direction of arrival (DoA) positioning systems are based on (i) a known location and orientation of the or each locator, and (ii) measurement of the DoA angle of the signal from the respective tag towards the locators in the locators' local co-ordinate system. Based on the location and angle information from one or more locators, the position of the tag can be calculated using geometry.

[0041] The CRS 15 is a processing system having an associated user interface (UI) 16 which will be explained in further detail below. As shown in Figure 1, it receives as input from the capture device 6 spatial audio and video data, and positioning data, through a signal line 17. Alternatively, the positioning data can be received from the HAIP locator 20. The CRS 15 also receives as input from each of the audio sources 7 - 13 audio data and positioning data from the respective positioning tags, or the HAIP locator 20, through separate signal lines 18. The CRS 15 generates spatial audio data for output to a user device 19, such as a VR headset with video and audio output.

[0042] The input audio data can be multichannel audio in loudspeaker format, e.g. stereo signals, 4.0 signals, 5.1 signals, Dolby Atmos (RTM) signals or the like. Instead of loudspeaker format audio, the input can be in the multi microphone signal format, such as the raw eight signal input from the OZO VR camera, if used for the capture device 6.

[0043] Figure 2 shows an example schematic diagram of components of the CRS 15. The CRS 15 has a con-

troller 22, a touch sensitive display 24 comprised of a display part 26 and a tactile interface part 28, hardware keys 30, a memory 32, RAM 34 and an input interface 36. The controller 22 is connected to each of the other components in order to control operation thereof. The touch sensitive display 24 is optional, and as an alternative a conventional display may be used with the hardware keys 30 and/or a mouse peripheral used to control the CRS 15 by conventional means.

[0044] The memory 32 may be a non-volatile memory such as read only memory (ROM) a hard disk drive (HDD) or a solid state drive (SSD). The memory 32 stores, amongst other things, an operating system 38 and software applications 40. The RAM 34 is used by the controller 22 for the temporary storage of data. The operating system 38 may contain code which, when executed by the controller 22 in conjunction with RAM 34, controls operation of each of hardware components of the terminal.

[0045] The controller 22 may take any suitable form. For instance, it may be a microcontroller, plural microcontrollers, a processor, or plural processors.

[0046] In embodiments herein, the software application 40 is configured to provide video and distributed spatial audio capture, mixing and rendering to generate a VR environment, or virtual space, including the rendered spatial audio. The software application 40 also provides the UI 16 shown in Figure 1, through its output to the display 24 and receives user input through the tactile interface 28 or other input peripherals such as the hardware keys 30 or a mouse (not shown). The mixing stage, which forms a sub-part of the rendering stage in this case, may be performed manually through the UI 16 or all or part of said mixing stage may be performed automatically. The software application 40 may render the virtual space, including the spatial audio, using known signal processing techniques and algorithms based on the mixing stage.

[0047] The input interface 36 receives video and audio data from the capture device 6, such as Nokia's OZO (RTM) device, and audio data from each of the audio sources 7 - 13. The input interface 36 also receives the positioning data from (or derived from) the positioning tags on each of the capture device 6 and the audio sources 7 - 13, from which can be made an accurate determination of their respective positions in the real world space 3.

[0048] The software application 40 may be configured to operate in any of real-time, near real-time or even offline using pre-stored captured data.

[0049] One example aspect of the mixing stage of software application 40 is controlling how audio sources move, or change position, in the rendered virtual space responsive to detected movement in the captured real world space 3. In this respect, during capture it is sometimes the case that audio sources move. For example, in the Figure 1 situation, any one of the audio sources 7 - 13 may move over time, as therefore will their respective

audio position with respect to the capture device 6 and also to each other. Commonly, users may be used to static sources where the audio source is generally central. When audio sources move, the rendered result may be overwhelming and distracting.

[0050] To handle real world movement, the software application 40 may be configured to restrict or filter movement of a subset of the audio sources 7 - 13 when rendering the spatial audio part of the virtual space. This restriction / filtering is performed based on a number of attributes, to be described below, some of which can be set manually through the UI 16 and some of which can be handled automatically by predetermined algorithms.

[0051] Figure 3 shows an overview flow diagram of the capture, mixing and rendering stages of software application 40. As mentioned, the mixing and rendering stages may be combined. First, video and audio capture is performed in step 3.1; next mixing is performed in step 3.2, followed by rendering in step 3.3. Mixing (step 3.2) may be dependent on a manual or automatic control step 3.4 which may be based on attributes of the captured video and/or audio. Other attributes may be used.

[0052] Examples of how the mixing step 3.2 may be performed will now be described.

[0053] In a first example, a determination may be made as to the one or most important audio sources 7 - 13 at each point in time. The remaining audio sources are assigned to a subset subject to filtering or restriction, meaning that their position in virtual space does not track real-world movement. The most important audio sources are permitted to move, meaning that the spatial position of their corresponding audio signals also move, typically to closely track the real world movement. Filtering or restriction of the remaining subset may mean preventing their movement completely (freezing / no movement) or limiting movement by filtering, e.g. using time-based averaging.

[0054] Referring to Figure 4a, which is similar to Figure 1, in this real world scenario the CRS 15 may determine that the lead vocalist 7 is the most important and therefore its captured sound is allowed to move in the rendered virtual space 42, represented graphically in Figure 4b. The remaining audio sources 8 - 13 are in this case frozen or restricted in terms of rendered movement. Figure 4b shows the result of the lead vocalist 7 and the lead guitarist 9 moving during capture, which is indicated in Figure 4a by the dotted arrows. The other audio sources 8, 10, 11, 12, 13 are also frozen from movement.

[0055] Figure 5 shows the processing steps performed by the CRS 15 through software application 40. The capturing step 5.1 is followed by the step 5.2 of identifying the most important source (MIS). Step 5.3 mixes the audio so as to allow movement of the MIS and to prevent movement of the other audio sources. Rendering takes place in step 5.4 in accordance with said mixing step 5.3. The feedback arrow indicates that the steps 5.1 - 5.4 may be repeated over time.

[0056] Determination of the MIS can be done using

various methods. One method is for the CRS 15 to analyse the position of each audio source 7 - 13 by means of the positioning tags, and/or using the captured video data from capture device 6. The front and centre positioning of the audio source 7 in this case is indicative of its importance relative to the others. Alternatively, or additionally, the spatial audio signal from the capture device 6 and/or audio signals from each of the audio sources 7 - 13 can be analysed by the CRS 15 to determine the most important sound. For example, the sound indicative of the lead melody or vocals may be indicative of the most important source.

[0057] The MIS may also be identified manually through the UI 16. Figure 6 shows an example screenshot of the UI 16 which shows the detected positions of the audio sources within the space 3, the UI permitting manual selection of whether a source is movable (M) or fixed/frozen (F). Clicking or touching on a source may toggle between M and F or a drop down menu can provide the options. Further filtering options can be provided.

[0058] In a second example, one or more anchor sources is or are determined and their position is fixed or filtered in a manner similar to the first example. For example, referring to the UI 16 representation of Figure 7, audio sources corresponding to drums 8, lead guitar 9 and bass guitar 10 are identified and restricted. Restriction here may mean freezing or fixing the position. The respective positions of these audio sources 8 - 10 may still be obtained and tracked using the tag-based positioning, and rather than fixing or freezing their position, their position may be filtered using a long term average position of each tag. In this context, long term may mean one minute, but shorter or longer periods may be used. The average position after one minute, to take this example, may indicate an allowed degree of movement in the rendered version. Figure 8 shows the main processing steps, comprised of capture (step 8.1), identifying anchor sources (step 8.2), mixing (step 8.3) and rendering (step 8.4).

[0059] As before, setting the anchor sources may be performed automatically by position, audio analysis or by other means, or manually through the UI 16. Filtering parameters may also be set through the UI 16, e.g. to set the averaging period.

[0060] In a third example, source movement may depend on the input signal type over time. Referring to Figure 9, for example, the lead guitarist 9 may be playing a guitar solo at a point in time and so is considered an important source and allowed to move, as is the lead vocalist 7. When the guitar solo ends, and reverts back to an accompanied performance, the source may become frozen or filtered as before.

[0061] As another example, the vocalisation type for the lead vocalist 7 may change over time; speaking or whistling can be identified by signal analysis and at this time the source position is frozen or time-filtered as explained above. When the vocalisation type changes to singing, normal movement may be enabled.

[0062] A method for detecting singing from non-singing type vocalisation is described in Berenzweig, A.L. and Ellis, D.P.W. "Locating Singing Voice Segments within Music Signals" Applications of Signal Processing to Audio and Acoustics, 2001, IEEE Workshop on the, 2001, 119-122.

[0063] This may be used to identify a speech-type monologue from singing, which may be the basis of MIS determination.

[0064] A further method of identifying audio signal type, e.g. for identifying singing, speaking, a change in vocalisation or the presence of a particular instrument or musical genre, is described in PCT/F12014/051036 the contents of which are incorporated herein by reference. An overview of the system and method described therein is described later on.

[0065] A generalised method may comprise training a model, for example a Gaussian mixture model or a neural network for the classes to be classified, e.g. singing, speaking, shouting etc. For the training stage, a set of audio files is obtained where the vocal type is known. From the audio files, a set of representative features are extracted; suitable features may be, for example, mel-frequency cepstral coefficients and the fundamental frequency (fo). Also, other features related to, for example, voice characteristics can be used.

[0066] Using the features, the models are then trained to represent typical feature values of different vocalisation types, in the case of a Gaussian mixture model, or a neural network is trained to discriminate between the different vocalisation types given the features. Online classification is then performed by extracting the same features from the input signal, and then evaluating which model has the highest likelihood of having generated the features (Gaussian mixture model) or which category gives the largest output value (neural network).

[0067] The actions to be taken responsive to signal type can be set using the mixing UI 16. The UI 16 can enable the operator to identify input signal types and associate each with different restrictions or filtered movements, for example. The manner in which the mixing operator sets the restrictions can cater for a variety of situations. For creating fast mixes, e.g. for live situations, the system may have a set of presets which can be applied. The behaviour indicated in Figure 9 may be one example preset. Figure 10 shows the main processing steps, comprised of capture (step 10.1), identification of signal type (step 10.2), mixing based on presets or rules (step 10.3) and rendering (step 10.4). In this case, the rules in step 10.3 assign certain types of signal to be moved and other types to be restricted; this represents a first preset, which can itself be modified over time to apply different rules.

[0068] In a fourth example, the CRS 15 is configured to filter or restrict movement within a group of associated audio sources. Referring to Figure 11, the choir members 11 - 13 are spatially close to one another and identified to the CRS 15 manually or automatically as a group. Each of the choir members 11 - 13 has a positioning tag, but

the CRS 15 may be configured to receive, or generate, a single source position which may be the middle or mean position of the group members, notwithstanding any movement or change in position. This may be desirable if we wish to have a single sound source that represents the whole group rather than using individual spatial positions. Alternatively, the individual sounds from each group member 11 - 13 may be adjusted so that they are equidistant from the group centre.

[0069] In a fifth example, the CRS 15 is again configured to filter or restrict movement with a group of associated audio sources, and in this case the CRS selectively filters the group movement so that coherent movement (all members moving in generally the same direction) is allowed in the rendered output, but movement in incoherent (different) directions is filtered, frozen and/or adjusted. This may be performed by determining a reference movement direction and subsequently filtering movement based on deviation of one or more group members. As an example, the CRS 15 may measure the most common movement direction and identifies if one or more group members deviate from the common direction.

[0070] Referring to Figure 12a, which shows only the group of choir audio sources 11 - 13 from Figure 1, the CRS 15 when detecting coherent movement allows that movement in the rendered version as shown. Referring to Figure 12b, the CRS 15 when detecting that one or more members (in this case the choir source 13) deviates away from the common movement, moves that member in the direction of common movement in contrast to their actual movement.

[0071] The CRS 15 may be user adjustable. For example, when there is a sufficient amount of coherent movement, all movement from the selected group is rendered. Other movement functions may be provided.

[0072] In the UI 16, the locator tags as before can be visualised and grouped. The mixer operator may select tags to form a group and then select an appropriate function for that group. Another example is to draw a rectangle or other shape to select tags within the area and then select an appropriate function for that group.

[0073] In a sixth example, the CRS 15 can filter movement of audio sources 7 - 13 based on the listener location within the virtual space. The CRS 15 can in real time determine the movement direction and/or velocity with reference to the listener location and through audio processing determine if it is audibly acceptable or pleasing, or not. For example, movement of a particular source when the listener is positioned directly in front of that source may be acceptable. However, if the listener were to move closer to, and to one side of, the same source, this may result in significant sideways movement producing unacceptable results. In this case, filtering can be applied, for example to reposition the audio source front and centre of the listener's position.

[0074] Referring now to Figure 13, an example schematic representation of capture and rendering parts of

the Figure 1 scenario is shown, some parts of which are provided by the CRS 15, in terms of functional components for mixing and rendering. For ease of explanation, the Figure 13 shows separate capture and rendering parts 101, 103 with mixing being performed as part of the rendering part.

[0075] In this example, there is shown only one audio signal source, indicated by reference numeral 111; however in practice there will be multiple such sources each of which may move in real world space. The audio source is in this example a Lavalier microphone 111 associated with a movable person, e.g. a singer.

[0076] The Lavalier microphone 111 is an example of a 'close' audio source capture apparatus and may in some embodiments be a boom microphone or similar neighbouring microphone capture system. Although the following examples are described with respect to a Lavalier microphone and thus a Lavalier audio signal the concept may be extended to any microphone external or separate to e.g. the microphones or array of microphones of the capture device 5 configured to capture the spatial audio signal. Thus the concept is applicable to any external/additional microphones in addition to a spatial audio capture (SPAC) microphone array, used in the OZO device, be they Lavalier microphones, hand held microphones, mounted mics, etc. The external microphones can be worn/carried by persons or mounted as close-up microphones for instruments or a microphone in some relevant location which the designer wishes to capture accurately. The Lavalier microphone 111 may in some embodiments be a microphone array. The Lavalier microphone typically comprises a small microphone worn around the ear or otherwise close to the mouth. For other sound sources, such as musical instruments, the audio signal may be provided either by a Lavalier microphone or by an internal microphone system of the instrument (e.g., pick-up microphones in the case of an electric guitar).

[0077] The Lavalier microphone 111 may be configured to output the captured audio signals to a variable delay compensator 117. The Lavalier microphone may be connected to a transmitter unit (not shown), which wirelessly transmits the audio signal to a receiver unit (not shown).

[0078] Furthermore the capture part 101 comprises a Lavalier (or close source) microphone position tag 112. The Lavalier microphone position tag 112 may be configured to determine information identifying the position or location of the Lavalier microphone 111 or other close microphone. It is important to note that microphones worn by people may be freely moved in the acoustic space and the system supporting location sensing of wearable microphone may support continuous sensing of user or microphone location. The Lavalier microphone position tag 112 may be configured to output this determination of the position of the Lavalier microphone to a position tracker 115.

[0079] The capture part 101 comprises a spatial audio

capture device 113 which corresponds to the capture device 5 in Figure 2. The spatial audio capture device 113 is an example of an 'audio field' capture apparatus and may in some embodiments be a directional or omnidirectional microphone array. The spatial audio capture device 113 may be configured to output the captured audio signals to a variable delay compensator 117.

[0080] Furthermore the capture part 101 may comprise a spatial capture position tag 114. The spatial capture position tag 114 may be configured to determine information identifying the position or location of the spatial audio capture device 113. The spatial capture position tag 114 may be configured to output this determination of the position of the spatial capture microphone to the position tracker 115. In the case the position tracker is co-located with the capture apparatus or the position of the capture apparatus with respect to the position tracker is otherwise known, and location data is obtained in relation to the capture apparatus, the capture apparatus does not need to comprise a position tag.

[0081] In some embodiments the spatial audio capture device 113 is implemented within a mobile device. The spatial audio capture device is thus configured to capture spatial audio, which, when rendered to a listener, enables the listener to experience the sound field as if they were present in the location of the spatial audio capture device. The Lavalier microphone 111 in such embodiments may be configured to capture high quality close-up audio signals (for example from a key person's voice, or a musical instrument). When mixed to the spatial audio field, the attributes of the key source such as gain, timbre and spatial position may be adjusted in order to provide the listener with a much more realistic immersive experience. In addition, it is possible to produce more point-like auditory objects, thus increasing the engagement and intelligibility.

[0082] The capture part 101 furthermore may comprise a position tracker 115. The position tracker 115 may be configured to receive the positional tag information identifying positions of the Lavalier microphone 111 and the spatial audio capture device 113 and generate a suitable output identifying the relative position of the Lavalier microphone 111 relative to the spatial audio capture device 113 and output this to the render apparatus 103 and specifically in this example to a location data filtering model 120. Furthermore in some embodiments the position tracker 115 may be configured to output the tracked position information to a variable delay compensator 117.

[0083] Thus in some embodiments the locations of the Lavalier microphones (or the persons carrying them) with respect to the spatial audio capture device 113 may be tracked and used for mixing the sources to correct spatial positions. In some embodiments the position tags namely, the microphone position tag 112 and the spatial capture position tag 114 may be implemented using High Accuracy Indoor Positioning (HAIP) or another suitable indoor positioning technology. In some embodiments, in addition to or instead of HAIP, the position tracker may

use video content analysis and/or sound source localization, as will be explained below.

[0084] The capture part 101 furthermore may comprise a variable delay compensator 117 configured to receive the outputs of the Lavalier microphone 111 and the spatial audio capture device 113. Furthermore in some embodiments the variable delay compensator 117 may be configured to receive source position and tracking information from the position tracker 115. The variable delay compensator 117 may be configured to determine any timing mismatch or lack of synchronisation between the close audio source signals and the spatial capture audio signals and determine the timing delay which would be required to restore synchronisation between the signals. In some embodiments the variable delay compensator 117 may be configured to apply the delay to one of the signals before outputting the signals to the render apparatus 103 and specifically in this example to the audio renderer 121. The timing delay may be referred to as being a positive time delay or a negative time delay with respect to an audio signal. For example, denote a first (spatial) audio signal by x , and another (Lavalier) audio signal by y . The variable delay compensator 117 is configured to try to find a delay T , such that $x(n)=y(n-T)$. Here, the delay T can be either positive or negative.

[0085] The capture stage 101 furthermore may comprise a camera 118 configured to capture video signals; the camera 118 may be a single camera or array of cameras, as in the OZO device example, in which spatial video signals are produced. In some embodiments, the camera 118 may form part of the same physical unit as the spatial audio capture device 113.

[0086] The capture stage 101 further may comprise a content analysis module 119 which receives both the video signals from the camera 118 and audio from the variable time delay compensator 117, which corrects for any timing mismatch or lack of synchronisation between the close audio signals, the spatial capture audio signals and the video signals. The output from the content analysis module 119 is provided to the rendering stage 103.

[0087] Also provided to the rendering stage 103 are signals from the position tracker 115 and the signals from the time delay compensator 117.

[0088] In some embodiments, the render part 103 comprises a head tracker 123. The head tracker 123 may be any suitable means for generating a positional input, for example a sensor attached to a set of headphones configured to monitor the orientation of the listener, with respect to a defined or reference orientation and provide a value or input which can be used by the audio renderer 121. The head tracker 123 may in some embodiments be implemented by at least one gyroscope and/or digital compass.

[0089] In some embodiments, the render part 103 comprises a location data filtering module 120 which takes as input the unfiltered HAIP location data from the position tracker 115, and the audio-visual content analysis data from the content analysis module 119.

[0090] The render part 103 comprises an audio renderer 121. The audio renderer 121 is configured to receive output for the location data filtering module 120. The audio renderer 121 can furthermore be configured to receive an input from the head tracker 123. Furthermore the audio renderer 121 can be configured to receive other user inputs for example entering manually through the UI 16. The audio renderer 121, as described herein in further detail later, can be configured to mix together the audio signals, namely the Lavalier microphone audio signals and the spatial audio signals based on the positional information and/or the head tracker inputs in order to generate a mixed audio signal in accordance with the methods and rules described above to cater for detected movement over time. The mixed audio signal can for example be passed to headphones 125. However the output mixed audio signal can be passed to any other suitable audio system for playback (for example a 5.1 channel audio amplifier).

[0091] In some embodiments the audio renderer 121 may be configured to perform spatial audio processing on the audio signals from the microphone array and from the close microphone.

[0092] The Lavalier audio signal is from the Lavalier microphones and the spatial audio captured by the microphone array and processed with the spatial analysis may in some embodiments be combined by the audio renderer to a single binaural output which can be listened through headphones.

[0093] The spatial audio signal may be converted into a multichannel signal. The multichannel output may then be binaurally rendered, and summed with binaurally rendered Lavalier source signals.

[0094] The rendering may be described initially with respect to a single (mono) channel, which can be one of the multichannel signals from the spatial audio signal or one of the Lavalier sources. Each channel in the multichannel signal set may be processed in a similar manner, with the treatment for Lavalier audio signals and multichannel signals having the following differences:

- 1) The Lavalier audio signals have time-varying location data (direction of arrival and distance) whereas the multichannel signals are rendered from a fixed location.
- 2) The ratio between synthesized "direct" and "ambient" components may be used to control the distance perception for Lavalier sources, whereas the multichannel signals are rendered with a fixed ratio.
- 3) The gain of Lavalier signals may be adjusted by the user whereas the gain for multichannel signals is kept constant.

[0095] The render part 103 in some embodiments may comprise headphones 125. The headphones can be used by the listener to generate the audio experience using the output from the audio renderer 121.

[0096] Thus based on the location tracking, the Lavalier

lier microphone signals can be mixed to suitable spatial positions in the spatial audio field in accordance with predetermined rules or user input. The rendering can be done by rendering the spatial audio signal using virtual loudspeakers with fixed positions, and the captured Lavalier source is rendered from a time varying position. Thus, the audio renderer 121 is configured to control the azimuth, elevation, and distance of the Lavalier or close source based on the tracked position data and the filtering rules.

[0097] Moreover, the user may be allowed to adjust the gain and/or spatial position of the Lavalier source using the output from the head-tracker 123. For example by moving the listeners head the head-tracker input may affect the mix of the Lavalier source relative to the spatial sound. This may be changing the 'spatial position' of the Lavalier source based on the head-tracker or by changing the gain of the Lavalier source where the head-tracker input is indicating that the listener's head is 'towards' or 'focussing' on a specific source. Thus the mixing/rendering may be dependent on the relative position/orientation of the Lavalier source and the spatial microphones but also be dependent on the orientation of the head as measured by the head-tracker. In some embodiments the user input may be any suitable user interface input, such as an input from a touchscreen indicating the listening direction or orientation.

[0098] Alternatively to a binaural rendering (for headphones), a spatial downmix into a 5.1 channel format or other format could be employed. In this case, the Lavalier or close source can in some embodiments mixed to its 'proper' spatial position using known amplitude panning techniques.

[0099] For completeness, an overview of one method for identifying audio signal type is now given with reference to Figures 14 to 16. Further details are provided in in PCT/FI2014/051036. The method may be used in the MIS identification step previously described. This method generally involves:

- (i) determining one or more acoustic features of audio data,
- (ii) generating at least one first classification based on the one or more determined acoustic features using a first classifier,
- (iii) generating at least one second classification based on the one or more determined acoustic features using at least one second classifier, which is different from the first,
- (iv) generating a third classification based on said first and second classifications using a third classifier, and
- (v) providing classification data, or tags, for the audio data based on said third classification.

[0100] Any one of the classifications can be used alongside selection rules to determine the MIS.

[0101] The first classifier and the third classifier may

be a non-probabilistic classifier, such as a support vector machine (SVM) classifier. The second classifier may be a probabilistic classifier, for example based on one or more Gaussian Mixture Models (GMMs). The classification data (or tags, as they are described in the reference) may indicate one or more of the following characteristics for determining the MIS using the predefined rules: a musical instrument, the presence or absence of vocals and/or a vocalist gender, presence or absence of music and a musical genre. This assumes that the audio data is music. In other embodiments, the audio data may comprise spoken word elements or a combination of music and spoken word elements.

[0102] Figure 14 is an overview of a method of determination of tag information for the audio track. Acoustic features 131 of the audio are extracted and input to first level classifiers 132 to generate first level classifications for the audio track. In this example, first classifiers 133 and second classifiers 134 are used to generate first and second classifications respectively. In the embodiments to be described below, the first classifiers 133 are non-probabilistic classifiers, while the second classifiers 134 are probabilistic classifiers. The first and second classifications generated by the first level classifiers 132 are provided as inputs to a second level classifier 135. One or more second level classifications are generated by the second level classifier 135, based at least in part on the first and second classifications. In the embodiments to be described below, the second level classifiers 135 include a third classifier 136, which outputs a third classification. One or more tags 137 are generated, based on the second level classifications. Such tags 137 may be stored by a tagging module 138 for determining the MIS.

[0103] The method will now be described in more detail, with reference to Figures 15 and 16.

[0104] Beginning at step s15.0 of Figure 15, if the received input signal is in a compressed format, such as MPEG-1 Audio Layer 3 (MP3), Advanced Audio Coding (AAC) and so on, the input signal is decoded into pulse code modulation (PCM) data (step s15.1). In this particular example, the samples for decoding are taken at a rate of 44.1 kHz and have a resolution of 16 bits.

[0105] Next, acoustic features 131 or descriptors are extracted which indicate characteristics of the audio track (step s13.2). In this particular embodiment, the features 131 are based on mel-frequency cepstral coefficients (MFCCs). In other embodiments, other features such as fluctuation pattern and danceability features, beats per minute (BPM) and related features, chorus features and other features may be used instead of, or as well as MFCCs. An example method for extracting acoustic features 131 from the input signal at step s15.2 is described in detail in PCT/FI2014/051036. As described therein, the audio features 131 produced at step 15.2 may include one or more of:

- a MFCC matrix for the audio track;
- first and, optionally, second time derivatives of the

- MFCCs, also referred to as "delta MFCCs";
- a mean of the MFCCs of the audio track;
 - a covariance matrix of the MFCCs of the audio track;
 - an average of mel-band energies over the audio track, based on output from the channels of the filter bank obtained;
 - a standard deviation of the mel-band energies over the audio track;
 - an average logarithmic energy over the frames of the audio track, obtained as an average of $c_{mel}(O)$ over a period of time obtained;
 - a standard deviation of the logarithmic energy;
 - a median fluctuation pattern over the song;
 - a fluctuation pattern bass feature;
 - a fluctuation pattern gravity feature;
 - a fluctuation pattern focus feature;
 - a fluctuation pattern maximum feature;
 - a fluctuation pattern sum feature;
 - a fluctuation pattern aggressiveness feature;
 - a fluctuation pattern low-frequency domination feature;
 - a danceability feature (detrended fluctuation analysis exponent for at least one predetermined time scale); and
 - a club-likeness value.
 - an average of an accent signal in a low, or lowest, frequency band;
 - a standard deviation of said accent signal;
 - a maximum value of a median or mean of periodicity vectors;
 - a sum of values of the median or mean of the periodicity vectors;
- tempo indicator for indicating whether a tempo identified for the input signal is considered constant, or essentially constant, or is considered non-constant, or ambiguous;
- a first BPM estimate and its confidence;
 - a second BPM estimate and its confidence;
 - a tracked BPM estimate over the audio track and its variation;
 - a BPM estimate from a lightweight tempo estimator.

[0106] Example techniques for beat tracking, using accent information, are disclosed in US published patent application no. 2007/240558 A1, US patent application no. 14/302,057, International (PCT) published patent application nos. WO2013/164661 A1 and WO2014/001849 A1, the disclosures of which are hereby incorporated by reference. In one example beat tracking method, described in GB 1401626.5, one or more accent signals are derived from the input signal, for detection of events and/or changes in the audio track. A first one of the accent signals may be a chroma accent signal based on fundamental frequency F_0 salience estimation, while a second one of the accent signals may be based on a multi-rate filter bank decomposition of the input signal. A BPM estimate may be obtained based on a periodicity analysis for extraction of a sequence of periodicity vectors on the

basis of the accent signals, where each periodicity vector includes a plurality of periodicity values, each periodicity value describing the strength of periodicity for a respective period length, or "lag". A point-wise mean or median of the periodicity vectors over time may be used to indicate a single representative periodicity vector over a time period of the audio track. For example, the time period may be over the whole duration of the audio track. Then, an analysis can be performed on the periodicity vector to determine a most likely tempo for the audio track. One example approach comprises performing k-nearest neighbours regression to determine the tempo. In this case, the system is trained with representative music tracks with known tempo. The k-nearest neighbours regression is then used to predict the tempo value of the audio track based on the tempi of k-nearest representative tracks. More details of such an approach have been described in Eronen, Klapuri, "Music Tempo Estimation With k -NN Regression", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, Issue 1, pages 50-57, the disclosure of which is incorporated herein by reference.

[0107] Chorus related features that may be extracted at step s15.2 include: a chorus start time; and a chorus end time.

[0108] Example systems and methods that can be used to detect chorus related features are disclosed in US 2008/236371 A1, the disclosure of which is hereby incorporated by reference in its entirety.

[0109] Other features that may be used as additional input include:

- a duration of the audio track in seconds;
- an A-weighted sound pressure level (SPL);
- a standard deviation of the SPL;
- an average brightness, or spectral centroid (SC), of the audio track, calculated as a spectral balancing point of a windowed FFT signal magnitude in frames of, for example, 40 ms in length;
- a standard deviation of the brightness;
- an average low frequency ratio (LFR), calculated as a ratio of energy of the input signal below 100Hz to total energy of the input signal, using a windowed FFT signal magnitude in 40 ms frames; and
- a standard deviation of the low frequency ratio.

[0110] Figure 16 is an overview of a process of extracting multiple acoustic features 131, some or all of which may be obtained in step s15.2. Figure 16 shows how some input features are derived, at least in part, from computations of other input features. The features 131 shown in Figure 16 include the MFCCs, delta MFCCs and mel-band energies discussed above, indicated using bold text and solid lines.

[0111] Returning to Figure 15, in steps s15.3 to s15.10, the method produces the first level classifications, that is the first classifications and the second classifications, based on the features 131 extracted in step s15.2. Al-

though Figure 15 shows steps s15.3 to s15.10 being performed sequentially, in other embodiments, steps s15.3 to s15.7 may be performed after, or in parallel with, steps s15.8 to s15.10.

[0112] The first and second classifications are generated using the first classifiers 133 and the second classifiers 134 respectively, where the first and second classifiers 133, 134 are different from one another. For instance, the first classifiers 133 may be non-probabilistic and the second classifiers 134 may be probabilistic classifiers, or vice versa. In this particular embodiment, the first classifiers 133 are support vector machine (SVM) classifiers, which are non-probabilistic. Meanwhile, the second classifiers 134 are based on one or more Gaussian Mixture Models (GMMs).

[0113] In step s15.3, one, some or all of the features 131 or descriptors extracted in step s15.2, to be used to produce the first classifications 133, are selected and, optionally, normalised. For example, a look up table or database may be stored for each of the first classifications to be produced, that provides a list of features to be used to generate each first classifier and statistics, such as mean and variance of the selected features, that can be used in normalisation of the extracted features 131. In such an example, the list of features is received from the look up table, and accordingly the method selects and normalises the listed features for each of the first classifications to be generated. The normalisation statistics for each first classification in the database may be determined during training of the first classifiers 133.

[0114] As noted above, in this example, the first classifiers 133 are SVM classifiers. The SVM classifiers 133 are trained using a database of audio tracks for which information regarding musical instruments and genre is already available. The database may include tens of thousands of tracks for each particular musical instrument that might be tagged.

[0115] Examples of musical instruments for which information may be provided in the database include:

Accordion;
Acoustic guitar;
Backing vocals;
Banjo;
Bass synthesizer;
Brass instruments;
Glockenspiel;
Drums;
Eggs;
Electric guitar;
Electric piano;
Guitar synthesizer;
Keyboards;
Lead vocals;
Organ;
Percussion;
Piano;
Saxophone;

Stringed instruments;
Synthesizer; and
Woodwind instruments.

[0116] The training database includes indications of genres that the audio tracks belong to, as well as indications of genres that the audio tracks do not belong to. By analysing features 131 extracted from the audio tracks in the training database, for which instruments and/or genre are known, a SVM classifier 133 can be trained to determine whether or not an audio track includes a particular instrument, for example, an electric guitar. Similarly, another SVM classifier 133 can be trained to determine whether or not the audio track belongs to a particular genre.

[0117] As mentioned above, the training process can include determining a selection of one or more features 131 to be used as a basis for particular first classifications and statistics for normalising those features 131. The number of features available for selection, M, may be much greater than the number of features selected for determining a particular first classification, N; that is, $M \gg N$. The selection of features 131 to be used is determined iteratively, based on a development set of audio tracks for which the relevant instrument or genre information is available. The normalisation of the selected features 131 at step s15.3 is optional. Where provided, the normalization of the selected features 131 in step s15.3 may potentially improve the accuracy of the first classifications.

[0118] At step s15.4, the method defines a single "feature vector" for each set of selected features 131 or selected combination of features 131. The feature vectors may then be normalized to have a zero mean and a variance of 1, based on statistics determined and stored during the training process.

[0119] At step s15.5, the method generates one or more first probabilities that the audio track has a certain characteristic, corresponding to a potential tag 137, based on the normalized transformed feature vector or vectors. A first classifier 133 is used to calculate a respective probability for each feature vector defined in step s15.4. In this manner, the number of SVM classifiers 133 corresponds to the number of characteristics or tags 137 to be predicted for the audio track.

[0120] In this particular example, a probability is generated for each instrument tag and for each musical genre tag to be predicted for the audio track, based on mean MFCCs and a MFCC covariance matrix. In addition, a probability may be generated based on whether the audio track is likely to be an instrumental track or a vocal track. Also, for vocal tracks, another first classification may be generated based on whether the vocals are provided by a male or female vocalist. In other embodiments, the controller may generate only one or some of these probabilities and/or calculate additional probabilities at step 15.5. The different classifications may be based on respective selections of features from the available features 131 ex-

tracted in step s15.2.

[0121] Optionally, a logarithmic transformation may be applied to the probabilities output by the SVM classifiers 133 (step s15.6), so that the probabilities of all the first classifications are on the same scale and the optimal predicated probability threshold may correspond to a pre-determined value, such as 0.5.

[0122] The first classifications are then output (step s15.7). The first classifications correspond to the normalized probability p_{norm} that a respective one of the tags 137 to be considered applies to the audio track. The first classifications may include probabilities p_{inst1} that a particular instrument is included in the audio track and probabilities p_{gen1} that the audio track belongs to a particular genre.

[0123] Returning to Figure 15, in steps s15.8 to s15.10, second classifications for the input signal are based on the MFCCs and other parameters produced in step s15.2, using the second classifiers 134. In this particular example, the features 131 on which the second classifications are based are the MFCC matrix for the audio track and the first time derivatives of the MFCCs.

[0124] In steps s15.8 to s15.10, the probabilities of the audio track including a particular instrument or belonging to a particular genre are assessed using probabilistic models that have been trained to represent the distribution of features extracted from audio signals captured from each instrument or genre. As noted above, in this example the probabilistic models are GMMs. Such models can be trained using an expectation maximisation algorithm that iteratively adjusts the model parameters to maximise the likelihood of the model for a particular instrument or genre generating features matching one or more input features in the captured audio signals for that instrument or genre. The parameters of the trained probabilistic models may be stored in a database or in remote storage.

[0125] For each instrument or genre, at least one likelihood is evaluated that the respective probabilistic model could have generated the selected or transformed features from the input signal. The second classifications correspond to the models which have the largest likelihood of having generated the features of the input signal.

[0126] In this example, probabilities are generated for each instrument tag at step s15.8 and for each musical genre tag at step s15.9, as well as a probability whether the audio track is likely to be an instrumental track or a vocal track may also be generated. Also, for vocal tracks, another probability may be generated based on whether the vocals are provided by a male or female vocalist. In other embodiments, the method may generate only one or some of these second classifications and/or calculate additional second classifications at steps s15.8 and s15.9.

[0127] In this embodiment, in steps s15.8 and s15.9, probabilities p_{inst2} that the instrument tags will apply, or not apply, are produced by the second classifiers 134 using first and second Gaussian Mixture Models

(GMMs), based on the MFCCs and their first time derivatives calculated in step s15.2. Meanwhile, probabilities p_{gen2} that the audio track belongs to a particular musical genre are produced by the second classifiers 134 using third GMMs. However, the first and second GMMs used to compute the instrument-based probabilities p_{inst2} may be trained and used slightly differently from third GMMs used to compute the genre-based probabilities p_{gen2} , as will now be explained.

[0128] In the following, step s15.8 precedes step s15.9. However, in other embodiments, step s15.9 may be performed before, or in parallel with, step s15.8.

[0129] In this particular example, first and second GMMs are used to generate the instrument-based probabilities p_{inst2} (step s15.8), based on MFCC features 131 obtained in step s15.2.

[0130] In step s15.9, for each of the genres that may be tagged, a likelihood L is computed for the audio track belonging to that genre, based on the likelihood of each of the third GMMs being capable of outputting the MFCC feature vector of the audio track. For example, to determine which of the eighteen genres in the list hereinabove might apply to the audio track, eighteen likelihoods would be produced.

[0131] The genre likelihoods are then mapped to probabilities p_{gen2} , as follows:

$$p_{gen2}(i) = \frac{L(i)}{\sum_{j=1}^m L(j)}$$

where m is the number of genre tags to be considered.

[0132] The second classifications, which correspond to the probabilities p_{inst2} and p_{gen2} , are then output (step s15.10).

[0133] The first classifications p_{inst1} and p_{gen1} and the second classifications p_{inst2} and p_{gen2} for the audio track are normalized to have a mean of zero and a variance of 1 (step s15.11) and collected to form a feature vector for input to one or more second level classifiers 135 (step s15.12). In this particular example, the second level classifiers 135 include third classifiers 136. The third classifiers 136 may be non-probabilistic classifiers, such as SVM classifiers.

[0134] The third classifiers 136 may be trained in a similar manner to that described above in relation to the first classifiers 133. At the training stage, the first classifiers 133 and the second classifiers 134 may be used to output probabilities for the training sets of example audio tracks from the database. The outputs from the first and second classifiers 133, 134 are then used as input data to train the third classifier 135.

[0135] The third classifier 136 generates determine probabilities p_{inst3} for whether the audio track contains a particular instrument and/or probabilities p_{gen3} for whether the audio track belongs to a particular genre (step

s15.13). The probabilities p_{inst3} , p_{gen3} are then log normalized (step s15.14), as described above in relation to the first classifications, so that a threshold of 0.5 may be applied to generate the third classifications, which are output at step s15.15.

[0136] The method then determines whether each instrument tag and each genre tag 137 applies to the audio track based on the third classifications (step s15.16).

[0137] Where it is determined that an instrument or genre tag 137 applies to the audio track (step s15.16), the tag 137 is associated with the track (step s15.17). The process ends at s15.18.

[0138] Further details are provided in PCT/FI2014/051036. In the present case, the classification data, or tag 137 generated at s15.17 is used by the rules to identify the MIS. It will be appreciated however that the overview given with reference to Figures 14 to 16 is merely one of a number of possible methods used in identifying the MIS.

[0139] It will be appreciated that the above described embodiments are purely illustrative and are not limiting on the scope of the invention. Other variations and modifications will be apparent to persons skilled in the art upon reading the present application.

[0140] Moreover, the disclosure of the present application should be understood to include any novel features or any novel combination of features either explicitly or implicitly disclosed herein or any generalization thereof and during the prosecution of the present application or of any application derived therefrom, new claims may be formulated to cover any such features and/or combination of such features.

Claims

1. A method comprising:

receiving a plurality of audio signals representing captured audio from respective audio sources in a space;
processing the received audio signals to generate a spatial audio signal, in which the processing comprises:

detecting spatial positions of the audio sources; and
responsive to a detected change in spatial position of one or more of said audio sources, generating the spatial audio signal so that said change in spatial position in the generated signal is different from the detected change in spatial position for the respective audio sources.

2. The method of claim 1, further comprising:

defining a group of audio sources from the plu-

rality of audio signals; and

wherein the change in the spatial position in the generated spatial audio signal depends on the position of one or more of the other audio sources in said group of audio sources.

3. The method of claim 1 or claim 2, further comprising receiving a spatial audio signal associated with a microphone array configured to provide spatial audio capture, and wherein the processing comprises generating an updated version of the spatial audio signal associated with the microphone array.

4. The method of any preceding claim, wherein the processing further comprises identifying a subset of the audio sources, and the detecting and generating steps are performed only in relation to the subset.

5. The method of claim 4, wherein identifying the subset is performed manually through a user interface.

6. The method of claims 4, wherein identifying the subset is performed automatically based on attributes of the audio sources.

7. The method of any of claims 4 to 6, further wherein identifying the subset is based on the position of one or more sources in the space.

8. The method of claim 7, wherein identifying the subset is based on the position of the one or more sources relative to a primary source whose position is permitted to change in the signal.

9. The method of claim 7 or claim 8 when dependent on claim 3, wherein identifying the subset is based on the position of the one or more sources relative to the position of the microphone array.

10. The method of any of claims 4 to 6, wherein identifying the subset is determined based on the type of audio source.

11. The method of claim 10, wherein the type of audio source is determined by signal analysis of the audio signals from each audio source.

12. The method of claim 11, wherein the analysis of the audio signals identifies a type of vocal performance or instrument.

13. The method of any preceding claim, wherein generating the spatial audio signal comprises controlling the change in spatial position in accordance with one or more position modification rules.

14. Apparatus configured to perform the method of any preceding claim.

15. A computer program comprising instructions that when executed by a computer apparatus control it to perform the method of any of claims 1 to 13.

5

10

15

20

25

30

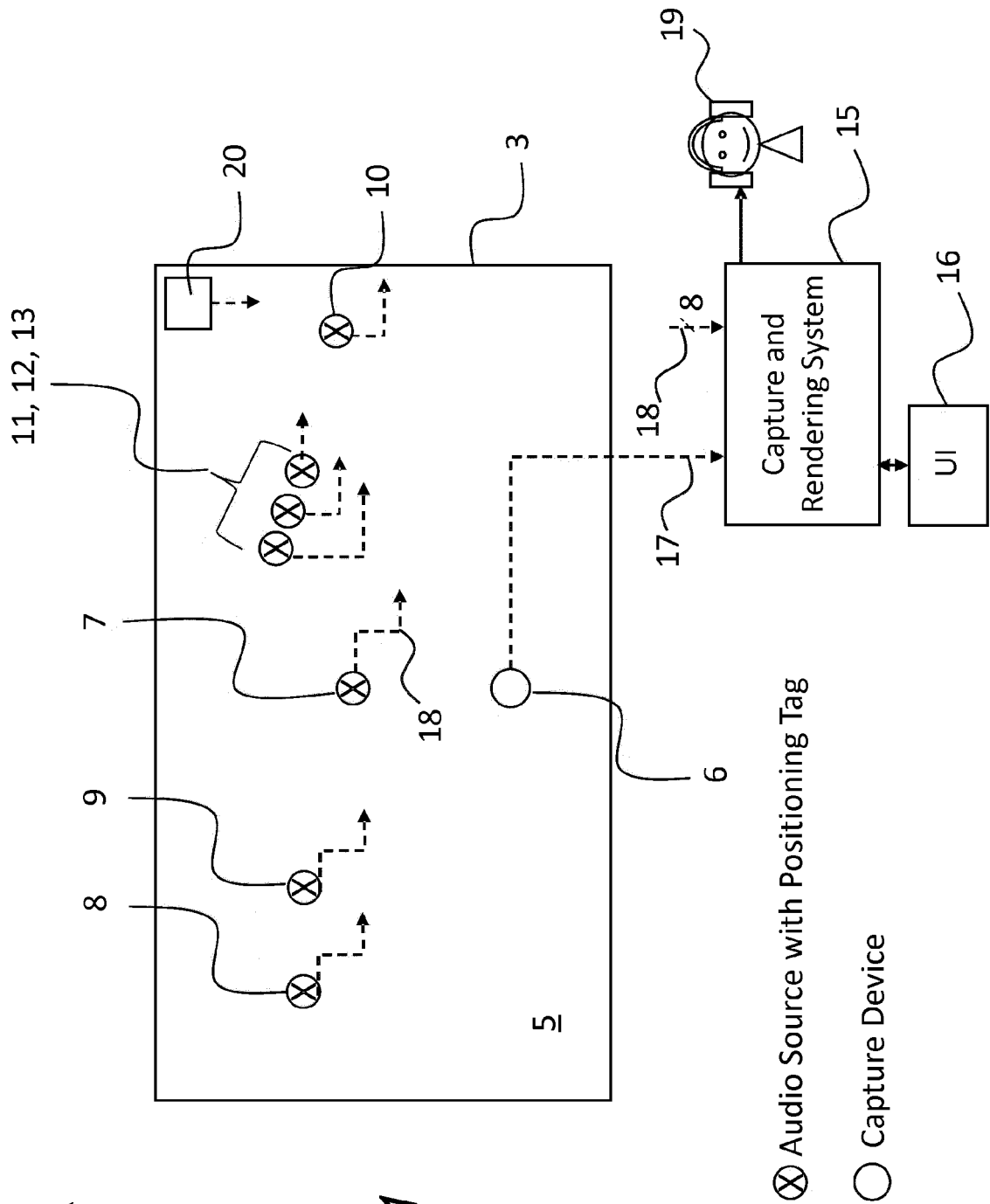
35

40

45

50

55



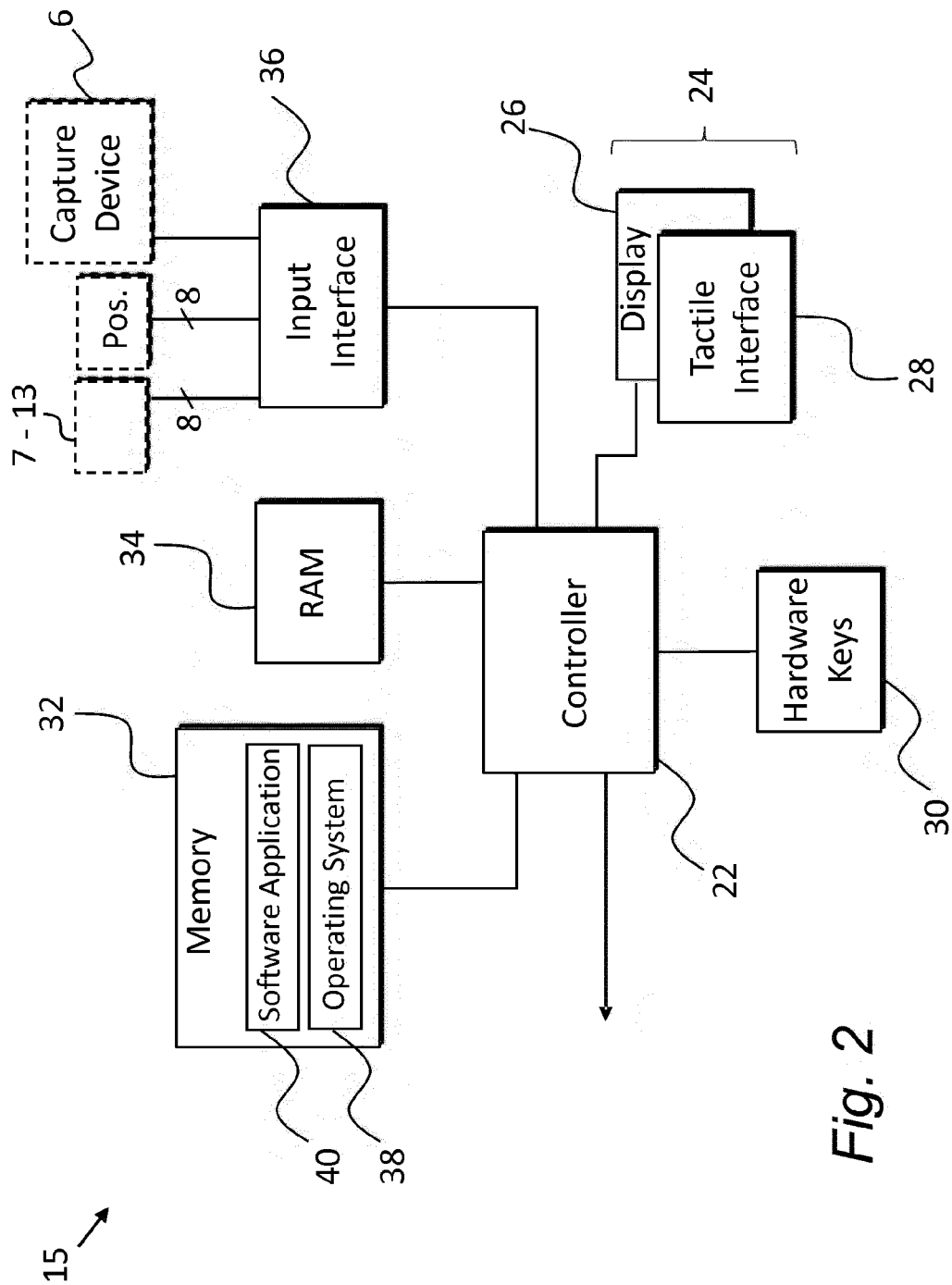


Fig. 2

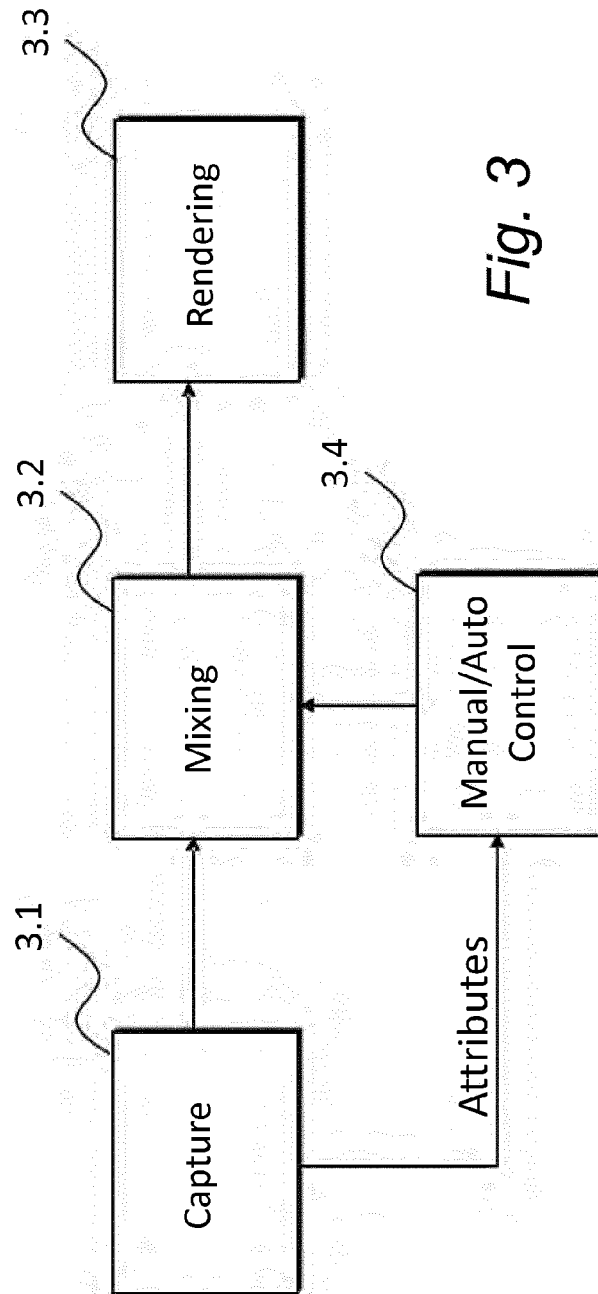


Fig. 3

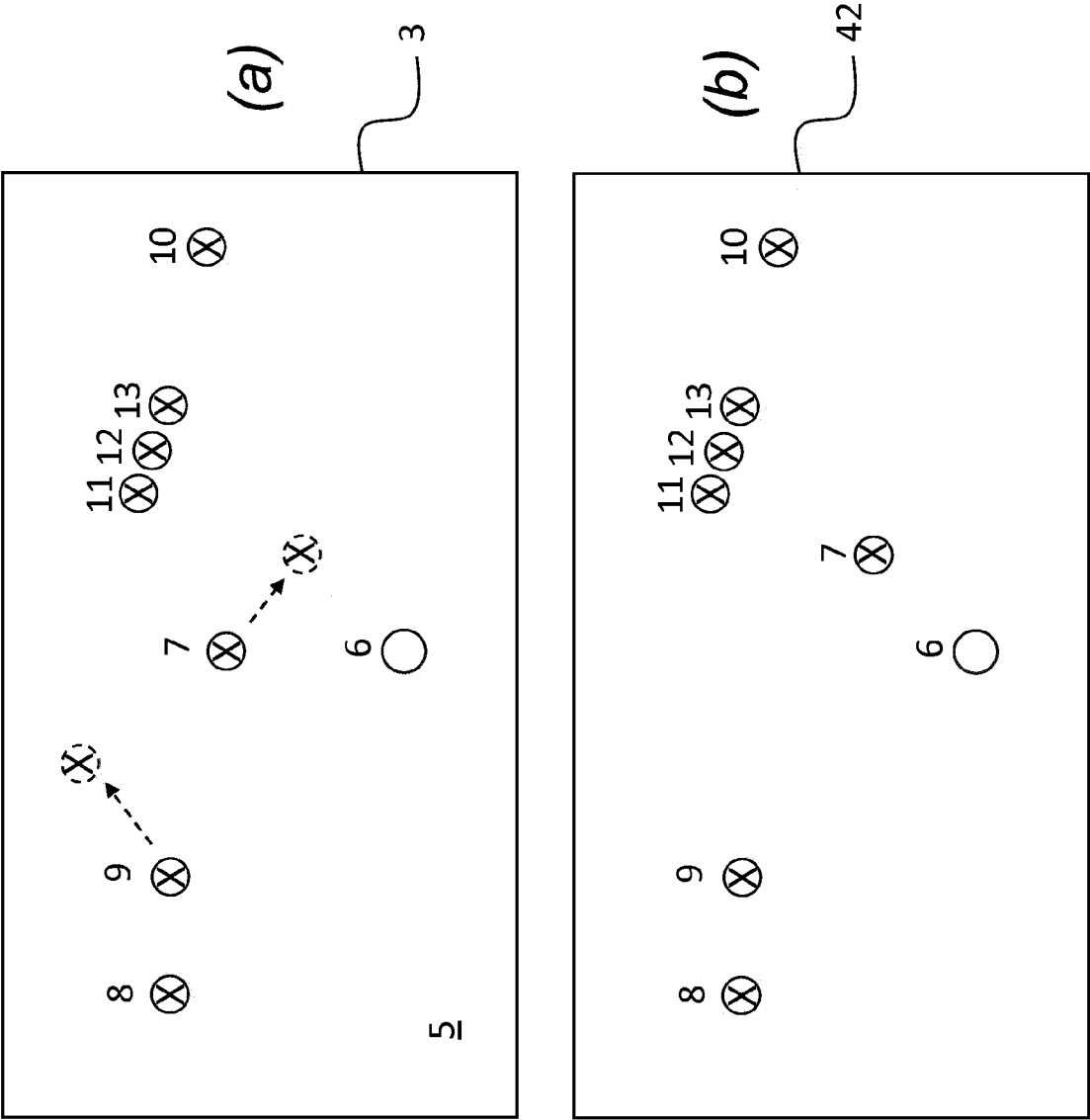


Fig. 4

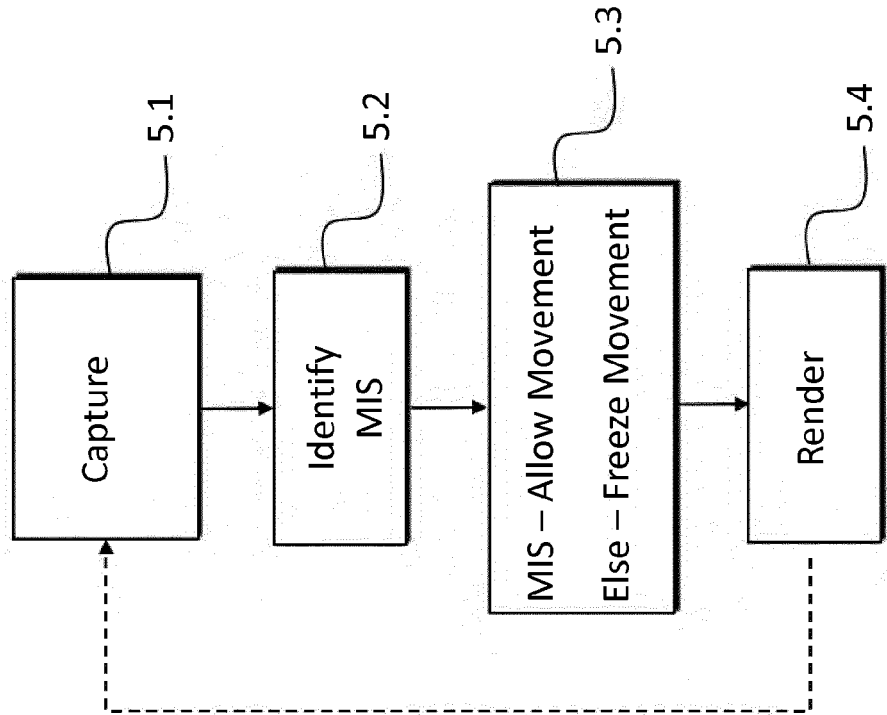


Fig. 5

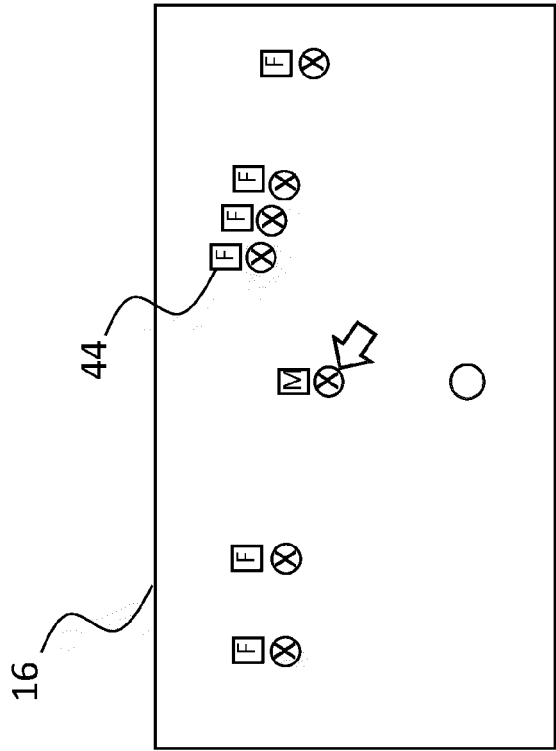
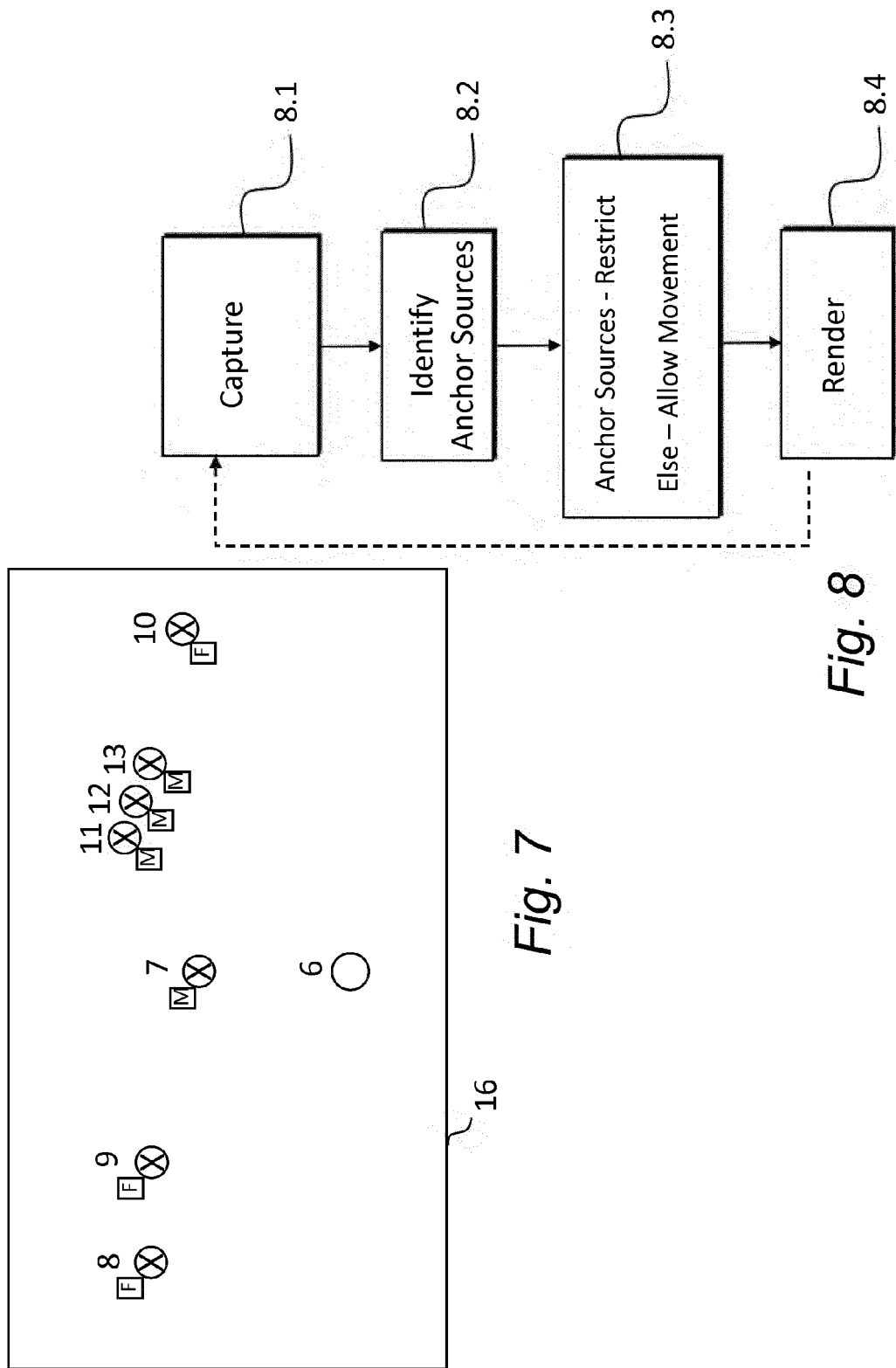
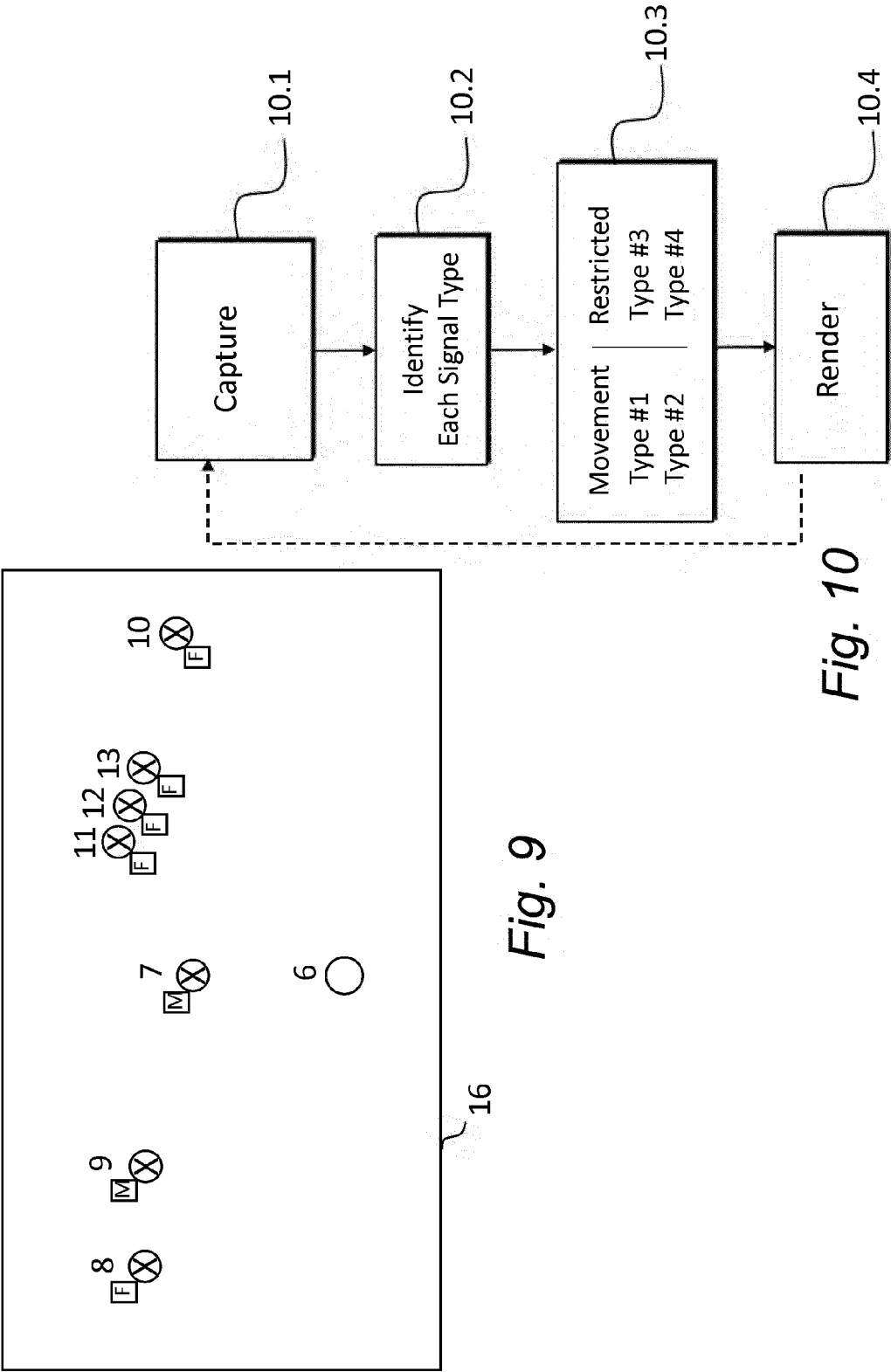


Fig. 6





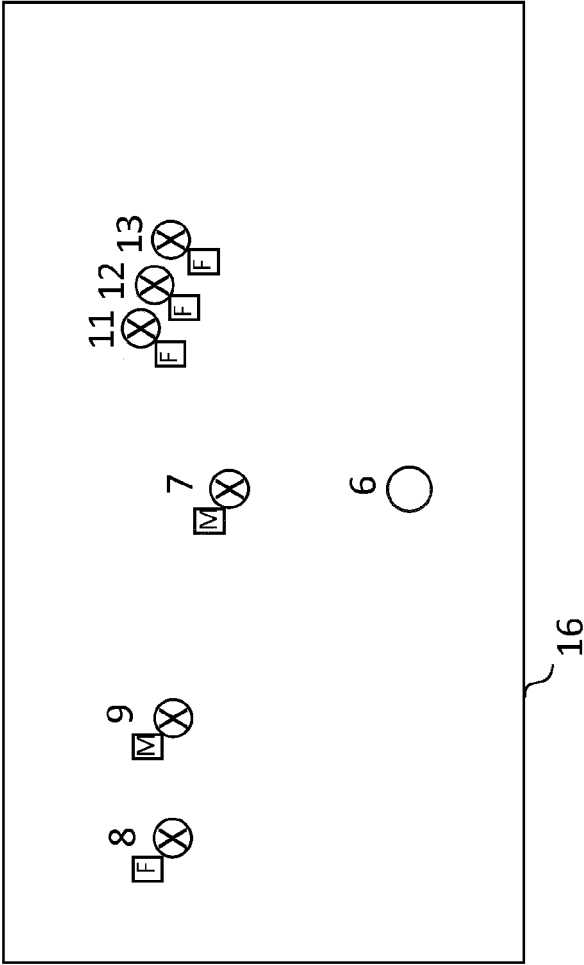


Fig. 11

Rendered Version

Captured Version

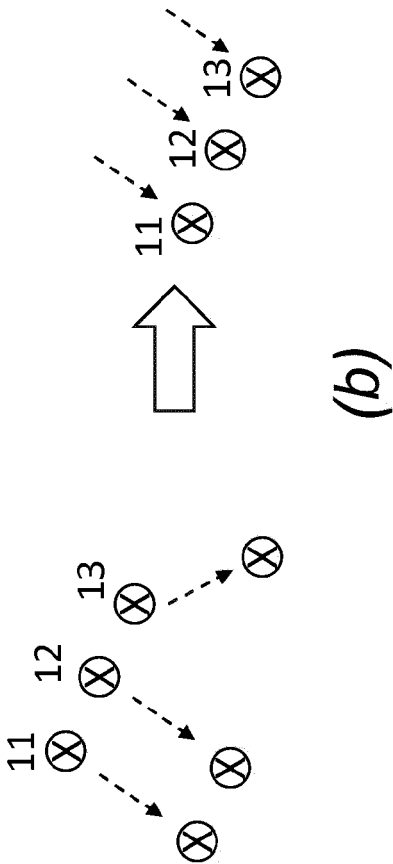
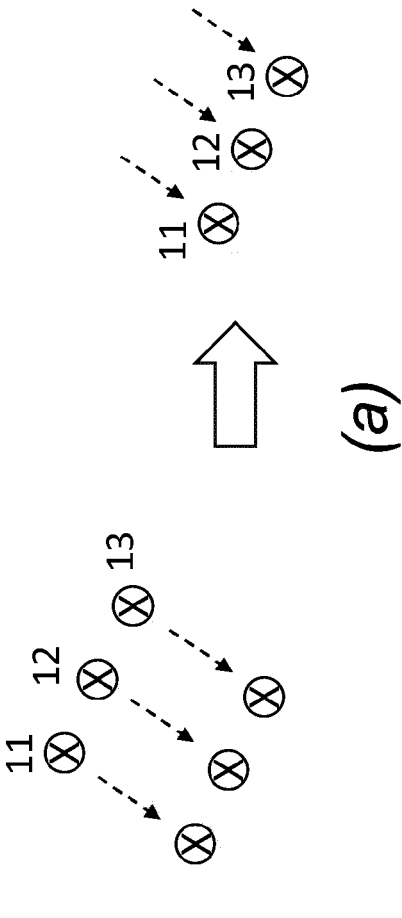


Fig. 12

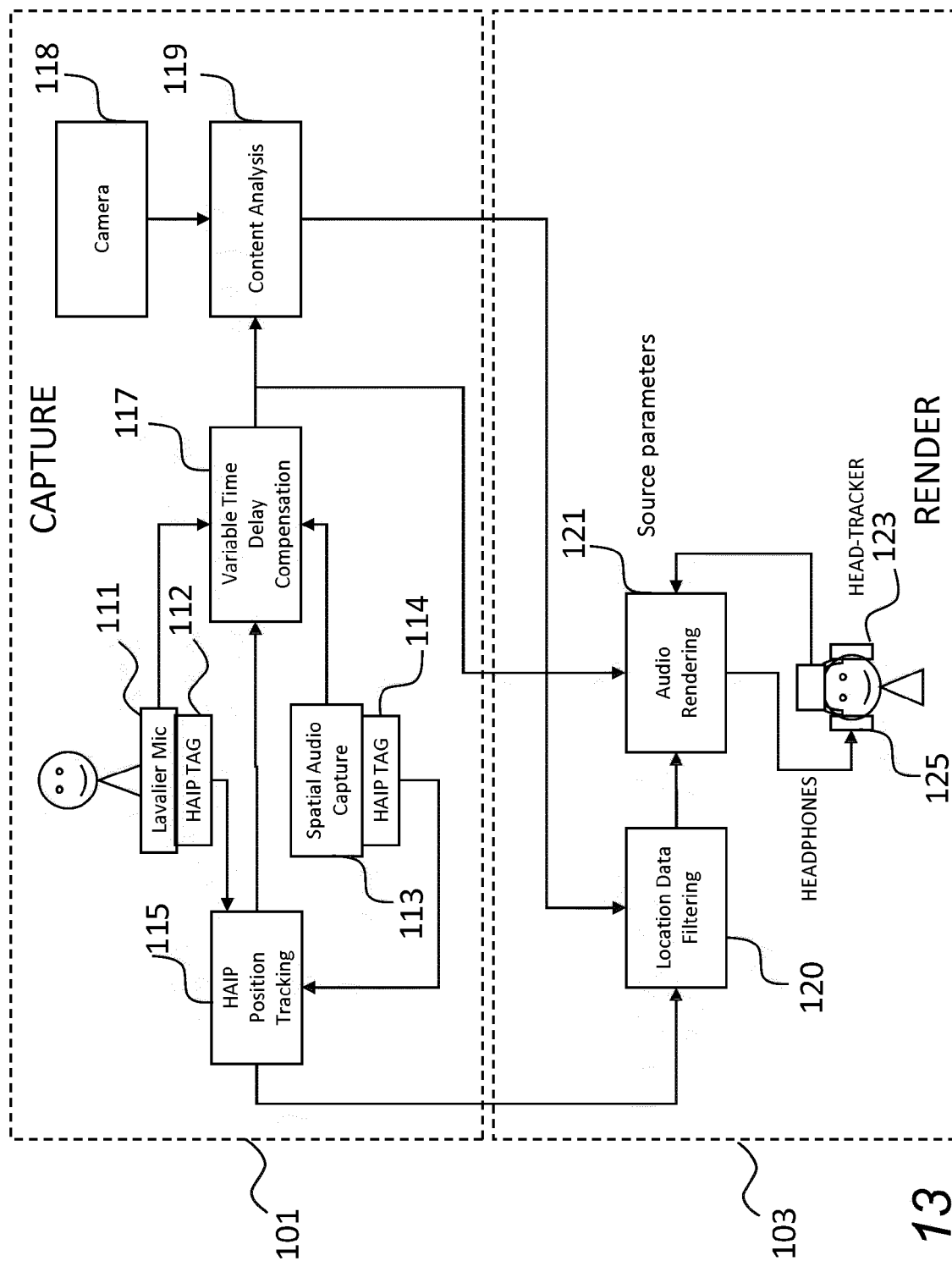


Fig. 13

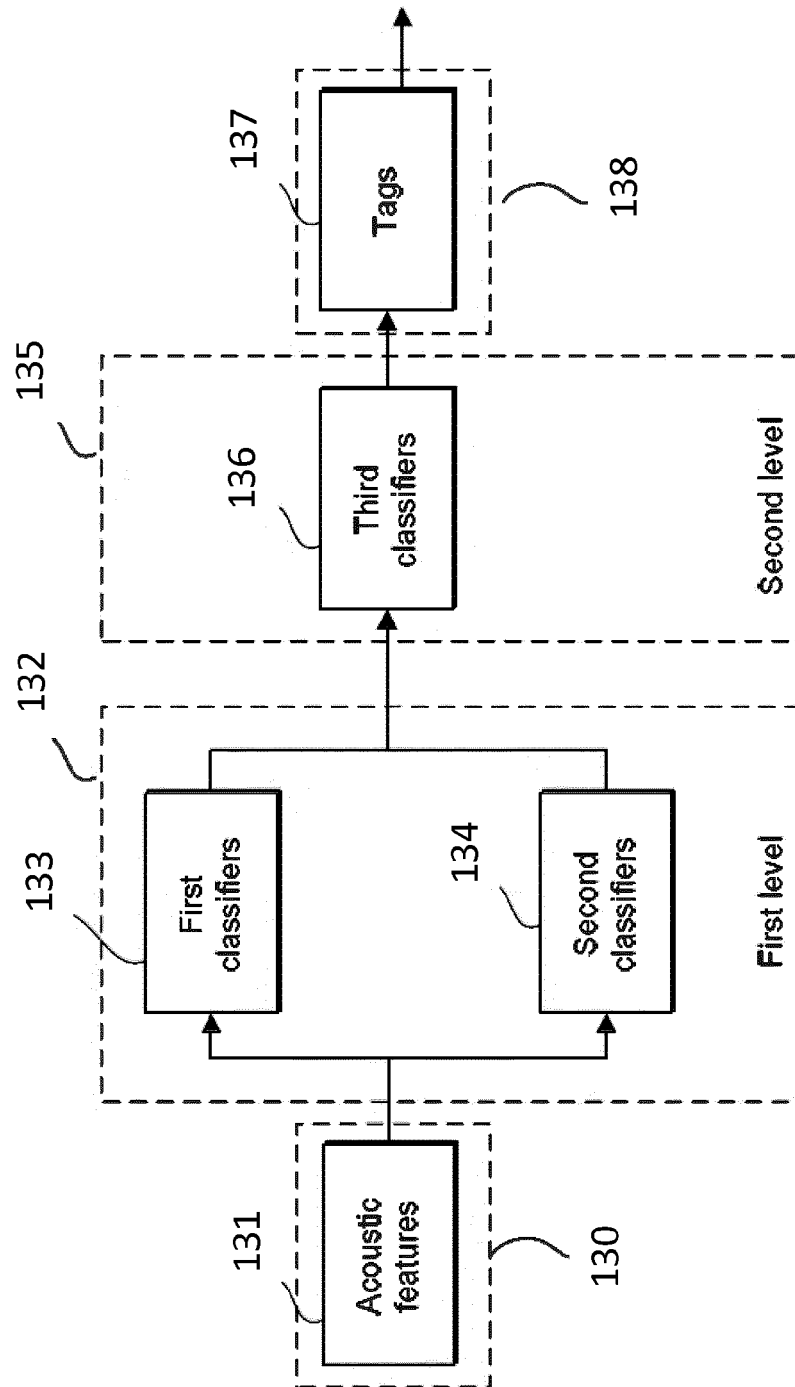


Fig. 14

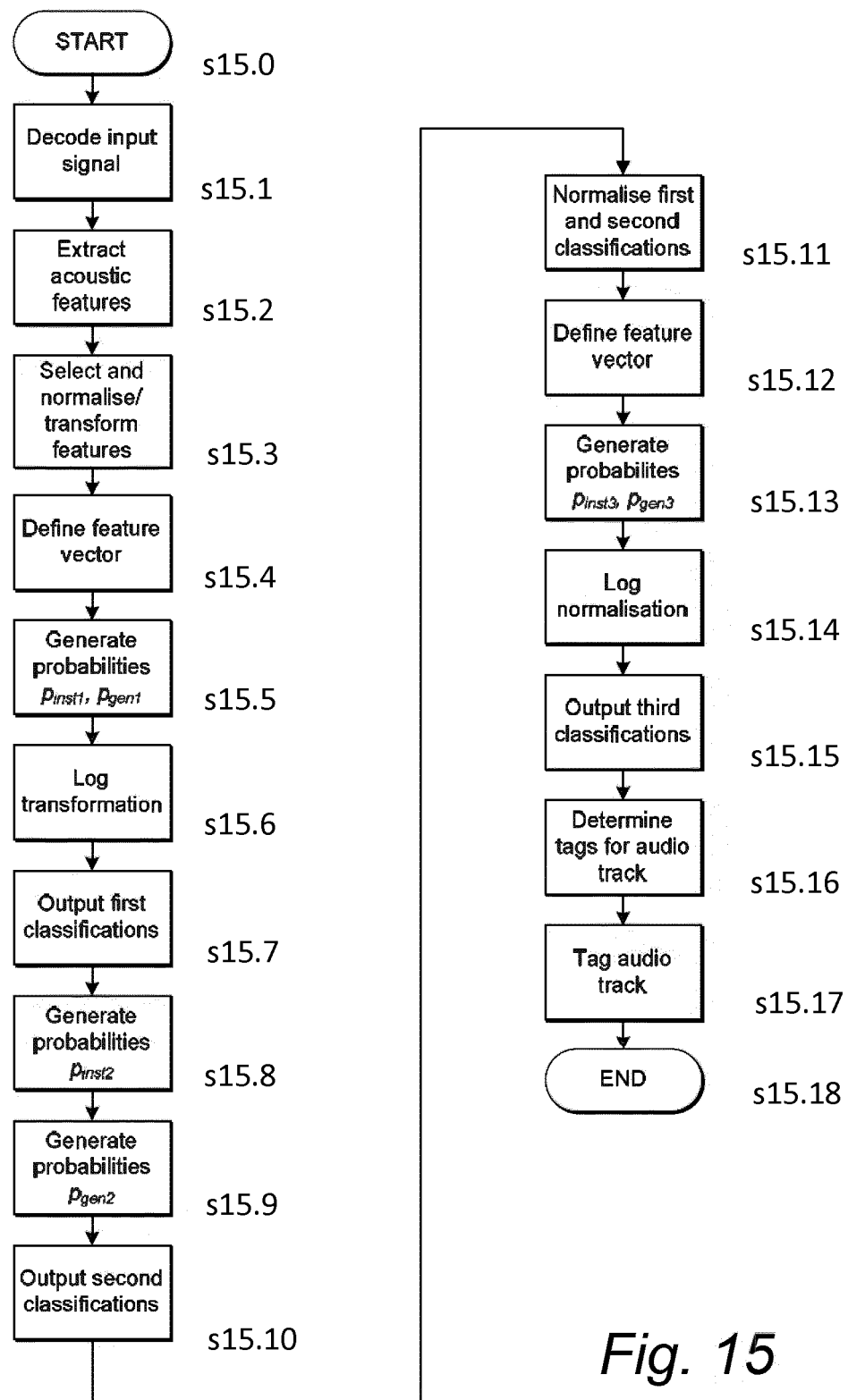


Fig. 15

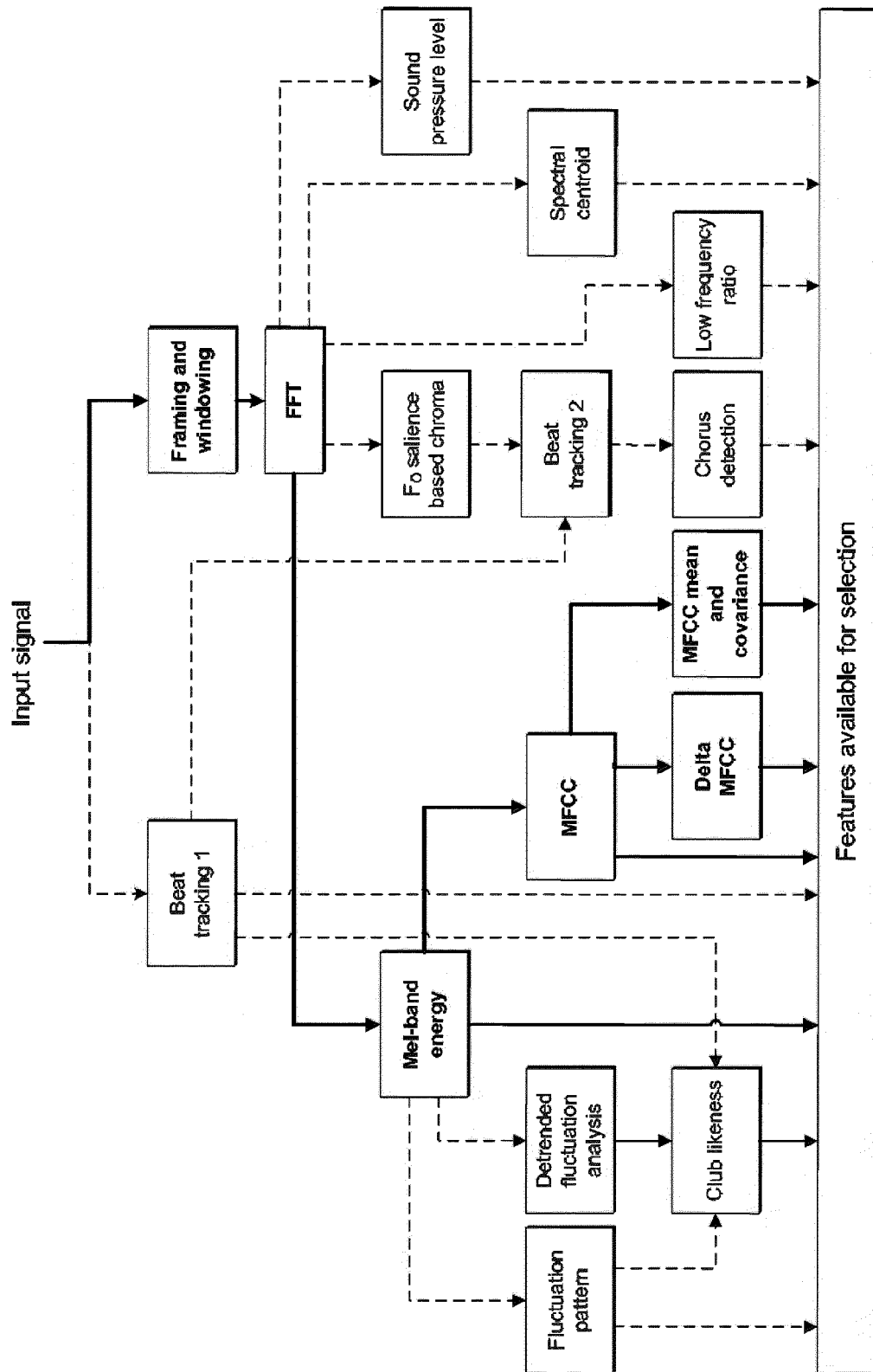


Fig. 16



EUROPEAN SEARCH REPORT

Application Number
EP 16 17 3264

5

10

15

20

25

30

35

40

45

50

55

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 2011/081024 A1 (SOULODRE GILBERT ARTHUR JOSEPH [CA]) 7 April 2011 (2011-04-07) * In particular paragraphs 0005-0008, 0036-0047, 0074, 0078, 0095, 0099, 0106 *	1-15	INV. H04S7/00
A	US 2010/223552 A1 (METCALF RANDALL B [US]) 2 September 2010 (2010-09-02) * In particular, paragraphs 0030-0033, 0039-0040, 0046, 0100, 0149 *	1-15	
A	US 2009/116652 A1 (KIRKEBY OLE [FI] ET AL) 7 May 2009 (2009-05-07) * paragraph [0034] *	3,9	
A	WO 2011/020065 A1 (SRS LABS INC [US]; TRACEY JAMES [US]; KATSIANOS THEMIS [US]; KRAEMER A) 17 February 2011 (2011-02-17) * In particular, paragraphs 0095-0096, 0100-0106 *	1-15	
The present search report has been drawn up for all claims			TECHNICAL FIELDS SEARCHED (IPC)
			H04S
Place of search		Date of completion of the search	Examiner
The Hague		24 November 2016	Thean, Andrew
CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document			

EPO FORM 1503 03/82 (P04C01)

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 16 17 3264

5

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

24-11-2016

10

15

20

25

30

35

40

45

50

55

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011081024 A1	07-04-2011	CA 2774415 A1	14-04-2011
		CN 102687536 A	19-09-2012
		EP 2486737 A1	15-08-2012
		JP 5400225 B2	29-01-2014
		JP 2013507048 A	28-02-2013
		KR 20120064104 A	18-06-2012
		KR 20140010468 A	24-01-2014
		US 2011081024 A1	07-04-2011
		WO 2011044064 A1	14-04-2011
US 2010223552 A1	02-09-2010	US 2010223552 A1	02-09-2010
		WO 2010101880 A1	10-09-2010
US 2009116652 A1	07-05-2009	CN 101843114 A	22-09-2010
		EP 2208363 A1	21-07-2010
		EP 2613564 A2	10-07-2013
		US 2009116652 A1	07-05-2009
		WO 2009056956 A1	07-05-2009
WO 2011020065 A1	17-02-2011	CN 102549655 A	04-07-2012
		CN 102576533 A	11-07-2012
		EP 2465114 A1	20-06-2012
		EP 2465259 A1	20-06-2012
		JP 5635097 B2	03-12-2014
		JP 5726874 B2	03-06-2015
		JP 2013502183 A	17-01-2013
		JP 2013502184 A	17-01-2013
		KR 20120061869 A	13-06-2012
		KR 20120062758 A	14-06-2012
		US 2011040395 A1	17-02-2011
		US 2011040396 A1	17-02-2011
		US 2011040397 A1	17-02-2011
		US 2013202129 A1	08-08-2013
		WO 2011020065 A1	17-02-2011
		WO 2011020067 A1	17-02-2011

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- FI 2014051036 W [0099] [0105] [0138]
- US 2007240558 A1 [0106]
- US 14302057 B [0106]
- WO 2013164661 A1 [0106]
- WO 2014001849 A1 [0106]
- GB 1401626 A [0106]
- US 2008236371 A1 [0108]

Non-patent literature cited in the description

- Locating Singing Voice Segments within Music Signals. **BERENZWEIG, A.L.** ; **ELLIS, D.P.W.** Applications of Signal Processing to Audio and Acoustics. IEEE Workshop, 2001, 119-122 [0062]
- **ERONEN ; K LAPURI.** Music Tempo Estimation With k -NN Regression. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, 50-57 [0106]