



(51) International Patent Classification:

H04R 5/027 (2006.01) *H04M 3/56* (2006.01)
H04R 29/00 (2006.01) *H04R 3/00* (2006.01)
G06K 9/62 (2006.01) *H04H 60/04* (2008.01)
H04N 5/265 (2006.01) *H04W 72/04* (2009.01)

(21) International Application Number:

PCT/FI2012/051033

(22) International Filing Date:

26 October 2012 (26.10.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: **NOKIA CORPORATION** [FI/FI]; Keilalahdentie 4, FI-02150 Espoo (FI).(72) Inventors: **MATE, Sujeet Shyamsundar**; Nosturinraitti 2 A 1, FI-33720 Tampere (FI). **OJANPERÄ, Juha**; Luuppikatu 3, FI-37120 Nokia (FI). **CURCIO, Igor Danilo Diego**; Hatanpään valtatie 12 C 57, FI-33100 Tampere (FI). **WILLNER, Kai**; Lohilantie 3 i 15, FI-37550 Lemppälä (FI).(74) Agents: **NOKIA CORPORATION** et al.; IPR Department, Jussi Jaatinen, Keilalahdentie 4, FI-02150 Espoo (FI).

(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, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, 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, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

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

(54) Title: MEDIA REMIXING SYSTEM

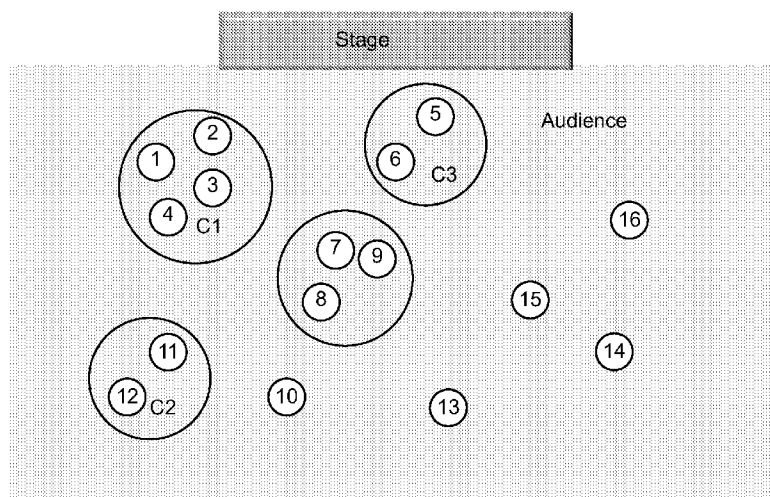


Fig. 4

(57) Abstract: A method comprising: receiving sensor data from a plurality of user devices attending an event; generating at least a spatial distribution of the user devices on the basis of the sensor data; dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

Media remixing system

Background

Multimedia capturing capabilities have become common features in portable devices. Thus, many people tend to record or capture an event, such as a music concert, a sport event or a private event such as a birthday or a wedding, they are attending. During many occasions, there are multiple attendants capturing content from an event, whereby variations in capturing location, view, equipment, etc. result in a plurality of captured versions of the event with a high amount of variety in both the quality and the content of the captured media.

Media remixing is an application where multiple media recordings are combined in order to obtain a media mix that contains some segments selected from the plurality of media recordings. Video remixing, as such, is one of the basic manual video editing applications, for which various software products and services are already available. Furthermore, there exist automatic video remixing or editing systems, which use multiple instances of user-generated or professional recordings to automatically generate a remix that combines content from the available source content. Some automatic video remixing systems depend only on the recorded content, while others are capable of utilizing environmental context data that is recorded together with the video content. The context data may be, for example, sensor data received from a compass, an accelerometer, or a gyroscope, or global positioning system (GPS) location data.

For a generic event, such as a music concert, sports event, etc., there may be multiple users in the audience capturing media content from the event, i.e. recording videos, audio clips and/or taking pictures. Some users capturing media content may be close to each other, whereas others may be further away. There is a higher likelihood of the users that are close to each other having higher commonality in the audio scene and thus in the captured media than users that are far apart. From the viewpoint of the media remix application, such commonality is redundant media content, which unnecessarily consumes network resources when uploaded or up-streamed and data storage resource when stored e.g. in the content management system.

Summary

Now there has been invented an improved method and technical equipment implementing the method for reducing the uploading or up-streaming of redundant

media content. Various aspects of the invention include a method, apparatuses and computer programs, which are characterized by what is stated in the independent claims. Various embodiments of the invention are disclosed in the dependent claims.

- 5 The aspects of the invention are based on the idea of selecting a subset of all the user devices capturing media content from the event as audio scene representatives based on their spatial distribution.

10 According to a first aspect, there is provided a method comprising: receiving sensor data from a plurality of user devices attending an event; generating at least a spatial distribution of the user devices on the basis of the sensor data; dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device
15 belongs.

According to an embodiment, the method further comprises determining a threshold distance for each cluster such that the user devices located within the threshold distance belong to the same cluster.

20 According to an embodiment, parameters affecting on the determination of the threshold distance for the clusters include one or more of the following:

- audio level of a common ambient audio scene in the event;
- regional variations of the audio level within the event;
- 25 - the number of users in the event;
- the size of the event venue;
- variation in density of the user devices.

30 According to an embodiment, the method further comprises generating a temporal distribution of the user devices on the basis of the sensor data.

According to an embodiment, the sensor data includes at least one of the following:

- position information indicating the position of the user device;
- orientation information indicating the orientation of the user device with relation to
35 the magnetic north;
- altitude information with relation to the horizontal;
- position information of the user device in 3-dimensional space in terms of the angles of rotation in three dimensions about the user device's center of mass;

- information whether the user device is recording or not recording media content from the event.

According to an embodiment, the method further comprises generating the one or more distributions of the user devices on the basis of the sensor data received from user devices recording content from the event.

According to an embodiment, the method further comprises receiving sensor data updates from the plurality of user devices attending the event; and updating the one or more distributions of the user devices on the basis of the updated sensor data.

According to an embodiment, selecting the representative user device for a cluster comprises determining the best quality audio from among the user devices in the cluster; and determining the best audio scene perspective among the user devices in the cluster by comparing location of interesting parts of the event and orientation of the user devices for a majority of interesting parts.

According to an embodiment, the method further comprises uploading or upstreaming, within a particular cluster, captured media content from the selected representative user device only.

According to an embodiment, the method further comprises synchronizing device clocks of the user devices recording media content from the event to each other.

According to an embodiment, audio quality of an audio track captured by a user device is analyzed by choosing, for a given audio track, a random sampling position for analyzing the quality of a small temporal segment of the audio; in response to the small segment being of good quality, choosing a subsequent sampling position for an analysis; and

repeating the choosing of the subsequent sampling position for the analysis for a predetermined number of times, in response to the previously analysed segment is of good quality; and determining the audio track to be of good quality.

According to an embodiment, choosing the subsequent sampling position is performed by using a half-interval search.

According to a second aspect, there is provided an apparatus comprising at least one processor, memory including computer program code, the memory and the computer program code configured to, with the at least one processor, cause the apparatus to

at least: receiving sensor data from a plurality of user devices attending an event; generating at least a spatial distribution of the user devices on the basis of the sensor data; dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

According to a third aspect, there is provided a computer program embodied on a non-transitory computer readable medium, the computer program comprising instructions causing, when executed on at least one processor, at least one apparatus to: receiving sensor data from a plurality of user devices attending an event; generating at least a spatial distribution of the user devices on the basis of the sensor data; dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

According to a fourth aspect, there is provided a computer readable storage medium stored with code thereon for use by an apparatus, which when executed by a processor, causes the apparatus to perform: receiving sensor data from a plurality of user devices attending an event; generating at least a spatial distribution of the user devices on the basis of the sensor data; dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

According to a fifth aspect, there is provided a system comprising at least one processor, memory including computer program code, the memory and the computer program code configured to, with the at least one processor, cause the system to at least: receive sensor data from a plurality of user devices attending an event; generate at least a spatial distribution of the user devices on the basis of the sensor data; divide the user devices into clusters of one or more user devices on the basis of the spatial distribution; and select a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

According to a sixth aspect, there is provided an apparatus comprising: means for receiving sensor data from a plurality of user devices attending an event; means for generating at least a spatial distribution of the user devices on the basis of the sensor

data; means for dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and means for selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

5

These and other aspects of the invention and the embodiments related thereto will become apparent in view of the detailed disclosure of the embodiments further below.

List of drawings

10

In the following, various embodiments of the invention will be described in more detail with reference to the appended drawings, in which

Figs. 1a and 1b show a system and devices suitable to be used in an automatic media remixing service according to an embodiment;

15 Fig. 2 shows an exemplified service architecture for creating a media remix;

Fig. 3 shows an exemplified implementation of a method according to some of the embodiments in a media remix application;

Fig. 4 shows, according to an embodiment, a process of clustering user devices in an event;

20 Fig. 5 shows, according to an embodiment, a method for selective temporal-segment sampling of an audio track performed on a user device to determine an approximate quality of the audio; and

Figs. 6a – 6c show an example of the temporal segment sampling process of Figure 5.

25

Description of embodiments

As is generally known, many contemporary portable devices, such as mobile phones, cameras, tablets, are provided with high quality cameras, which enable to capture high quality video files and still images. In addition to the above capabilities, such

30 handheld electronic devices are nowadays equipped with multiple sensors that can assist different applications and services in contextualizing how the devices are used. Sensor (context) data and streams of such data can be recorded together with the video or image or other modality of recording (e.g. speech).

35 Usually, at events attended by a lot of people, such as live concerts, sport games, political gatherings, and other social events, there are many who record still images and videos using their portable devices, thus creating user generated content (UGC).

A significant amount of this UGC will be uploaded to social media portals (SMP), such as Facebook, YouTube, Flickr®, and Picasa™, etc. These SMPs have become de facto storages of the generated social media content. The uploaded UGC recordings of the attendants from such events, possibly together with various sensor information, provide a suitable framework for the present invention and its embodiments.

The media content to be used in media remixing services may comprise at least video content including 3D video content, still images (i.e. pictures), and audio content including multi-channel audio content. The embodiments disclosed herein are mainly described from the viewpoint of creating an automatic media remix from video and audio content of source videos, but the embodiments are not limited to video and audio content of source videos, but they can be applied generally to any type of media content.

Figs. 1a and 1b show a system and devices suitable to be used in an automatic media remixing service according to an embodiment. In Fig. 1a, the different devices may be connected via a fixed network 210 such as the Internet or a local area network; or a mobile communication network 220 such as the Global System for Mobile communications (GSM) network, 3rd Generation (3G) network, 3.5th Generation (3.5G) network, 4th Generation (4G) network, Wireless Local Area Network (WLAN), Bluetooth®, or other contemporary and future networks. Different networks are connected to each other by means of a communication interface 280. The networks comprise network elements such as routers and switches to handle data (not shown), and communication interfaces such as the base stations 230 and 231 in order for providing access for the different devices to the network, and the base stations 230, 231 are themselves connected to the mobile network 220 via a fixed connection 276 or a wireless connection 277.

There may be a number of servers connected to the network, and in the example of Fig. 1a are shown servers 240, 241 and 242, each connected to the mobile network 220, which servers may be arranged to operate as computing nodes (i.e. to form a cluster of computing nodes or a so-called server farm) for the automatic media remixing service. Some of the above devices, for example the computers 240, 241, 242 may be such that they are arranged to make up a connection to the Internet with the communication elements residing in the fixed network 210.

There are also a number of end-user devices such as mobile phones and smart phones 251, Internet access devices (Internet tablets) 250, personal computers 260 of various sizes and formats, televisions and other viewing devices 261, video

decoders and players 262, as well as video cameras 263 and other encoders. These devices 250, 251, 260, 261, 262 and 263 can also be made of multiple parts. The various devices may be connected to the networks 210 and 220 via communication connections such as a fixed connection 270, 271, 272 and 280 to the internet, a
5 wireless connection 273 to the internet 210, a fixed connection 275 to the mobile network 220, and a wireless connection 278, 279 and 282 to the mobile network 220. The connections 271-282 are implemented by means of communication interfaces at the respective ends of the communication connection.

10 Fig. 1b shows devices for automatic media remixing according to an example embodiment. As shown in Fig. 1b, the server 240 contains memory 245, one or more processors 246, 247, and computer program code 248 residing in the memory 245 for implementing, for example, automatic media remixing. The different servers 241, 242, 290 may contain at least these elements for employing functionality relevant to each
15 server.

Similarly, the end-user device 251 contains memory 252, at least one processor 253 and 256, and computer program code 254 residing in the memory 252 for implementing, for example, gesture recognition. The end-user device may also have
20 one or more cameras 255 and 259 for capturing image data, stereo video, 3D video or alike. The end-user device may also contain one, two or more microphones 257 and 258 for capturing sound. The end-user device may also contain sensors for generating the depth information using any suitable technology. The different end-user devices 250, 260 may contain at least these same elements for employing
25 functionality relevant to each device. In another embodiment of this invention, the depth maps (i.e. depth information regarding the distance from the scene to a plane defined by the camera) obtained by interpreting video recordings from the stereo (or multiple) cameras may be utilized in the media remixing system. The end-user device may also have a time-of-flight camera, whereby the depth map may be obtained from
30 a time-of-flight camera or from a combination of stereo (or multiple) view depth map and a time-of-flight camera. The end-user device may generate depth map for the captured content using any available and suitable mechanism.

The end user devices may also comprise a screen for viewing single-view,
35 stereoscopic (2-view), or multiview (more-than-2-view) images. The end-user devices may also be connected to video glasses 290, e.g., by means of a communication block 293 able to receive and/or transmit information. The glasses may contain separate eye elements 291 and 292 for the left and right eye. These eye elements may either show a picture for viewing, or they may comprise a shutter functionality

e.g., to block every other picture in an alternating manner to provide the two views of three-dimensional picture to the eyes, or they may comprise an orthogonal polarization filter (compared to each other), which, when connected to similar polarization realized on the screen, provide the separate views to the eyes. Other arrangements for video glasses may also be used to provide stereoscopic viewing capability. Stereoscopic or multiview screens may also be autostereoscopic, i.e., the screen may comprise or may be overlaid by an optics arrangement, which results into a different view being perceived by each eye. Single-view, stereoscopic, and multiview screens may also be operationally connected to viewer tracking such a manner that the displayed views depend on viewer's position, distance, and/or direction of gaze relative to the screen.

It needs to be understood that different embodiments allow different parts to be carried out in different elements. For example, parallelized processes of the automatic media remixing may be carried out in one or more processing devices; i.e., entirely in one user device like 250, 251 or 260, or in one server device 240, 241, 242 or 290, or across multiple user devices 250, 251, 260 or across multiple network devices 240, 241, 242, 290, or across both user devices 250, 251, 260 and network devices 240, 241, 242, 290. The elements of the automatic media remixing process may be implemented as a software component residing on one device or distributed across several devices, as mentioned above, for example so that the devices form a so-called cloud.

One or more of the computers disclosed in Fig. 1a may be configured to operate a multimedia content remix service, which can be referred to as a media remix service. The media remix service is a service infrastructure that is capable of receiving user communication requests for inviting other users. The media remix service, together with the computer(s) running the service, further comprise networking capability to receive and process media content and corresponding context data from other data processing devices, such as servers operating social media portals (SMP). Herein, the term social media portal (SMP) refers to any commonly available portal that is used for storing and sharing user generated content (UGC). The UGC media content can be stored in various formats, for example, using the formats described in the Moving Picture Experts Group MPEG-4 standard. The context data may be stored in suitable fields in the media data container file formats, or in separate files with database entries or link files associating the media files and their timestamps with sensor information and their timestamps. Some examples of popular SMPs are YouTube, Flickr®, and Picasa™. It is apparent for a skilled person that the media

remix service and the social media portals SMP are implemented as network domains, wherein the operation may be distributed among a plurality of servers.

A media remix can be created according to the preferences of a user. The source content refers to all types of media that is captured by users, wherein the source content may involve any associated context data. For example, videos, images, audio captured by users may be provided with context data, such as information from various sensors, such as from a compass, an accelerometer, a gyroscope, or information indicating location, altitude, temperature, illumination, pressure, etc. A particular sub-type of source content is a source video, which refers to videos captured by the user, possibly provided with the above-mentioned context information.

A user can request the media remix service an automatically created media remix version from the material available for the service about an event, such as a concert. The service may be available to any user or it may be limited to registered users only. It is also possible to create a media remix version from private video material only. The service creates an automatic cut of the video clips of the users. The service may analyze the sensory data to determine which are interesting points at each point in time during the event, and then makes switches between different source media in the final cut. Audio alignment is used to find a common timeline for all the source videos, and, for example, dedicated sensor data (accelerometer, compass) analysis algorithms are used to detect when several users are pointing to the same location on the stage, most likely indicating an interesting event. Furthermore, music content analysis (beats, downbeats), is used to find a temporal grid of potential cut points in the event sound track.

Fig. 2 shows exemplified service architecture for creating an automatically created media remix. The service architecture may include components, known as such from contemporary video editing services, for example an interface 200 for the users contributing their recorded content from the event, which interface may annotate the contributed content for clustering the content related to the same event for generating the media remix, a content management system (CMS; 202) to store/tag/organize the content, and an interface 204 for delivering the media remix and its related source content to the users to consume.

The service architecture of Fig. 2 may further comprise a feedback module (FBM; 206) to capture the content consumption feedback about the content contributed by the users and the media remix versions that have been generated. The feedback

information may be provided to a synergistic intelligence module (SIM; 208), which contains the required intelligence or the logic required to analyze and create the information about the user contributed source content that is contributed to the service. The SIM is connected to a user apparatus 214 via a signalling interface 212,
5 which enables the user to request a media remix to be created according to user-defined parameters and also to provide new UGC content to be used in the media remix generation process.

In the analysis the SIM may utilize, in addition to the feedback information, also
10 information about the arrival distribution pattern of the source content. The SIM may use the UGC contribution data from past events in various locations and use it to generate a probabilistic model to predict user content contribution's arrival time (or upload time) to the service. The information provided by the SIM are received in a synergizing engine (SE; 210), which may be implemented as a separate module that
15 interacts with the CMS, the SIM and the FBM to generate the media remix versions that match the criteria signalled by the user requesting a media remix. The information provided by the SIM enables the SE to utilize the previous media remix versions and their consumption feedback as inputs, in addition to the newly provided source content and its consumption feedback, wherein the SE changes the weights of
20 different parameters which are used to combine the multitude of content.

For a generic event, such as a music concert, sports event, etc., there may be multiple users in the audience capturing media content from the event, i.e., recording
25 videos, audio clips and/or taking pictures. The presence of multiple users recording at the event implies that there may be some redundancy in the captured content in such a way that there may be multiple users recording at the same time during some time intervals, there may be just one user recording at some time intervals or there may be no users recording at some other time intervals during the event. Some users capturing media content may be close to each other, whereas others may be further
30 away. There is a higher likelihood of the users that are close to each other having higher commonality in the audio scene and thus in the captured media than users that are far apart.

From the viewpoint of the media remix application, such commonality is redundant
35 media content, which unnecessarily consumes network resources when uploaded or up-streamed and data storage resource when stored e.g., in the content management system. This is a serious problem in mobile domain given the fact that mobile network connectivity is always a bottleneck for uploading or up-streaming a large amount of data. As an example, a 720p resolution video with encoded a typical contemporary

video codec results in a file of 80MB size for each minute of video recording. The video resolution being used is increasing at a much faster pace than the corresponding increase in network bandwidth.

5 Now in order to reduce the uploading or up-streaming of such redundant media content, a spatial sampling method is now presented for selecting a subset of all the user devices capturing media content from the event as candidate audio scene representatives based on their spatial distribution. The user devices are clustered such that user devices that are closer than a predefined threshold distance to each
10 other are considered to belong to same cluster. The predefined threshold distance may be modulated based on the spatial expanse of the event. One user device is selected as an audio scene representative from each cluster.

A method according to some of the embodiments is illustrated in the flow chart of
15 Figure 3, wherein the operation is described from the perspective of a media remix application, typically executed on one or more servers in a network. The media remix application receives sensor data from a plurality of user devices attending an event (300). The sensor data may include at least one or more of the following:

- position information indicating the position of the user device. The position may
20 be determined, for example, using satellite positioning system such as GPS (Global Positioning System), using cell identification of a mobile communication network, using ad-hoc WLAN, Near Field Communication (NFC), Bluetooth or using any indoor positioning system. The position may also be determined as relative position to other user devices or a reference point.
- 25 - orientation information indicating the orientation of the user device with relation to the magnetic north. The orientation information may be determined, for example, using a compass.
- altitude information with relation to the horizontal, determined for example using an accelerometer.
- 30 - position information of the user device in 3-dimensional space in terms of the angles of rotation in three dimensions about the user device's center of mass, i.e., so-called roll, pitch and yaw. The 3d space position may be determined using a gyroscope.
- information whether the user device is recording or not recording media content
35 from the event.

On the basis of the sensor data collected from the plurality of user devices participating in the event, the media remix application generates at least a spatial and optionally a temporal distribution of the user devices (302). According to an

embodiment, only those user devices that are recording the media content from the event are considered, when generating update the spatial and the temporal distribution of the user devices. The user devices may continuously update their sensor data and the media remix application may consequently update the spatial and the temporal distribution of the users.

Information on the spatial and the temporal distribution of the users is used to divide the user devices into clusters of one or more user devices (304). The clusters may be formed such that the user devices located within a predefined threshold distance to each other are determined to form a cluster.

Parameters affecting on the determination of the threshold distance for the clusters may include one or more of the following:

- audio level of the common ambient audio scene in the event. This affects the distance up to which the audio can be heard.
- regional variations of the audio level within the event. For example, configuration of loudspeakers in the event may cause regional variations in the audio level.
- the number of users in the event. As the number of users increases, the closer they are likely to be to each other for a given event venue size.
- the size of the event venue. The bigger is the size of the venue, the larger threshold would likely be needed for cluster formation.
- variation in density of the users. Regions in the event venue with higher density of users would need smaller thresholds than the regions with less density.

The user devices that are recording media content from the event and that are located closer to each other than the threshold distance are clustered together. Recording user devices located outside the threshold distance from any other recording user are considered to have created a single user cluster.

For each cluster, a representative user device is selected (306) from all candidate user devices of the cluster to represent the audio scene for the area encompassed by the cluster to which the user device belongs. According to an embodiment, selecting the representative user device may consist of determination of the best quality audio from among the candidates in the cluster and determination of the best audio scene perspective by comparing the location of the interesting parts of the event and the candidates' orientation for the majority of the considered temporal interval.

Using the combination of the orientation of the candidate users and the quality of captured audio, the best candidate is chosen to represent the audio scene. From the

group of the user devices in a particular cluster, the captured media content is then preferably uploaded or up-streamed from the selected representative user device only. This results in avoidance of uploading or up-streaming and processing of content from other user devices in the cluster.

5

According to an embodiment, the device clocks of the user devices recording media content from the event are synchronized to each other. Thus, the sensor data received from the user devices may be timely synchronized, which enables the media remix application to update the spatial and the temporal distribution of the users, the confirmation of the clusters and the representative of the cluster flexibly and in timely manner.

10

The process of clustering user devices is illustrated in an example of Figure 4, wherein an audience is gathered in an event, for example a music concert. Within the audience, altogether 16 user devices recording media content from the event are detected. The recording user devices send their sensor data to the server comprising the media remix application. On the basis of the sensor data and the parameters derived therefrom, the media remix application determines the threshold distances for clustering the user devices.

15

20

In the example of Figure 4, four clusters including a plurality of user devices are determined: cluster 1 (C1) consisting of user devices 1, 2, 3 and 4, C2 consisting of user devices 11 and 12, C3 consisting of user devices 5 and 6, and C4 consisting of user devices 7, 8 and 9. User devices 10, 13, 14, 15 and 16 are determined as clusters of single user device. Figure 4 illustrates well how the threshold distance may vary between the clusters within the event venue. For example, due to parameter changes in audio level of the common ambient audio scene in the event, regional variations of the audio level within the event and/or variation in density of the users, the threshold distance for cluster 1 (C1) is larger than the threshold distance of cluster 3 (C3). For the single user clusters, the threshold distance is determined to be so small that no other user devices are located within the threshold distance.

25

30

An important aspect in the process of determining the representative user device with the best quality audio from among the candidate user devices in the cluster is a method for selective temporal-segment sampling of an audio track from the candidate user devices in the cluster is performed locally on each user device to determine an approximate quality of the audio. In addition, the segments of audio that are of not good enough quality may be determined.

35

The method is illustrated in the flow chart of Figure 5. For a given audio track, a random sampling position is chosen for analyzing the quality of a small temporal segment of the audio (500). If the analysis of the small segment results in (502) determining the segment of good quality, the analysis may be continued by choosing a subsequent sampling position (500). For choosing the subsequent sampling position, any well-known pattern, such as a half-interval search may be used. The steps (500, 502) are repeated N times such that a counter value is increased by one (504) for every good quality segment, and if the result of the first N evaluations is good quality (506), the audio track may be classified as a good quality audio track (508). On the other hand, if the analysis (500) of the small segment results in (502) determining the segment to of poor quality, a binary search is performed (510) before and after the bad quality segment. If the binary search reveals further bad quality segments (512), the bad quality segments may be subjected to a further binary search, until the width of the bad quality audio content in the audio track can be localized. The bad quality audio content in the audio track may then be discarded (514). This will remove the possibility of uploading or up-streaming bad quality audio track and consequently reduce the amount of data that needs to be uploaded or up-streamed.

An example of the temporal segment sampling process is illustrated in Figures 6a – 6c, wherein the audio quality from user device 1 in cluster C1 is determined. Within a time interval T, initially only a small central temporal segment (S1) is chosen for analysis, as shown in Figure 6a. If the segment S1 is determined to be of good quality, the audio for the whole time interval T is selected for further analysis. Next, two other temporal segments (S2, S3) may be chosen for analysis at the quarter and three-quarter intervals, as shown in Figure 6b. In this example, segment S3 turns out to be of bad quality, which results in further analysis in two temporal segments (S4, S5), one on both sides of segment S3, as shown in Figure 6c. Segment S4 is determined to be of good quality, whereas segment S5 is of bad quality. As a result, an audio segment starting from the beginning of the time interval T up to the segment S4 included is determined to be good quality content, whereas an audio segment of the rest of the time interval T, starting after segment S4, is determined as content to be skipped. For that period, an audio segment from another user device in cluster C1 may be selected to represent the audio scene of cluster C1.

According to an embodiment, the number of iteration rounds N; i.e. selecting further segments for analysis, may vary depending on the required level of accuracy in granularity. According to an embodiment, the number of iteration rounds N may

optimized in terms of the amount of analysis that needs to be performed on the user device and the savings in content upload or up-stream.

Thus, in a situation involving presence of multiple users recording content from an event, the spatial and temporal sampling of media can be used to generate an audio representation of the ambient audio scene at the event with minimal upload or up-stream of audio content.

The media remix service has been described above as implemented in a client-server-type media remix service. However, the implementation is not limited to client-server-type system, but according to an embodiment, the media remix service may be implemented as peer-to-peer-type system, where the processing steps described above are performed on one or more user devices.

In yet another embodiment, the system is client-server-type, but at least some of the steps described above are performed on the user device.

A skilled man appreciates that any of the embodiments described above may be implemented as a combination with one or more of the other embodiments, unless there is explicitly or implicitly stated that certain embodiments are only alternatives to each other.

The various embodiments may provide advantages over state of the art. For example, for creating a media remix from an event, audio content upload or up-stream requirements may be minimized while preserving audio scene information from the event. The minimized amount of the uploaded or up-streamed content from the event may enable faster collaborative media remixing. Since only a few user devices upload or up-stream the captured media content to the media remix application, significant power savings on average may be achieved in the user devices and bandwidth consumption savings on various network connections. The power savings may enable longer battery life in the user devices.

In general, the various embodiments of the invention may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. For example, some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device, although the invention is not limited thereto. While various aspects of the invention may be illustrated and described as block diagrams, flow charts, or using some other pictorial representation, it is well

understood that these blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

5 The embodiments of this invention may be implemented by computer software executable by a data processor of the mobile device, such as in the processor entity, or by hardware, or by a combination of software and hardware. Further in this regard it should be noted that any blocks of the logic flow as in the Figures may represent
10 program steps, or interconnected logic circuits, blocks and functions, or a combination of program steps and logic circuits, blocks and functions. The software may be stored on such physical media as memory chips, or memory blocks implemented within the processor, magnetic media such as hard disk or floppy disks, and optical media such as for example DVD and the data variants thereof, or CD.

15 The memory may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The data processors
20 may be of any type suitable to the local technical environment, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multi core processor architecture, as non-limiting examples.

25 Embodiments of the inventions may be practiced in various components such as integrated circuit modules. The design of integrated circuits is by and large a highly automated process. Complex and powerful software tools are available for converting a logic level design into a semiconductor circuit design ready to be etched and formed on a semiconductor substrate.

30 Programs, such as those provided by Synopsys, Inc. of Mountain View, California and Cadence Design, of San Jose, California automatically route conductors and locate components on a semiconductor chip using well established rules of design as well as libraries of pre stored design modules. Once the design for a semiconductor
35 circuit has been completed, the resultant design, in a standardized electronic format (e.g., Opus, GDSII, or the like) may be transmitted to a semiconductor fabrication facility or "fab" for fabrication.

The foregoing description has provided by way of exemplary and non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those skilled in the relevant arts in view of the foregoing description, when read in
5 conjunction with the accompanying drawings and the appended claims. However, all such and similar modifications of the teachings of this invention will still fall within the scope of this invention.

Claims:

1. A method comprising:
receiving sensor data from a plurality of user devices attending an event;
5 generating at least a spatial distribution of the user devices on the basis
of the sensor data;
dividing the user devices into clusters of one or more user devices on the
basis of the spatial distribution; and
selecting a representative user device for a cluster to represent the audio
10 scene for the area of the cluster to which the representative user device belongs.

2. A method according to claim 1, further comprising
determining a threshold distance for each cluster such that the user
15 devices located within the threshold distance belong to the same cluster.

3. A method according to claim 2, wherein
parameters affecting on the determination of the threshold distance for
the clusters include one or more of the following:

- audio level of a common ambient audio scene in the event;
- 20 - regional variations of the audio level within the event;
- the number of users in the event;
- the size of the event venue;
- variation in density of the user devices.

25 4. A method according to any preceding claim, further comprising
generating a temporal distribution of the user devices on the basis of the
sensor data.

30 5. A method according to any preceding claim, wherein
the sensor data includes at least one of the following:
- position information indicating the position of the user device;
- orientation information indicating the orientation of the user device with
relation to the magnetic north;
- altitude information with relation to the horizontal;
35 - position information of the user device in 3-dimensional space in terms
of the angles of rotation in three dimensions about the user device's center of mass;
- information whether the user device is recording or not recording media
content from the event.

6. A method according to any preceding claim, further comprising generating the one or more distributions of the user devices on the basis of the sensor data received from user devices recording content from the event.

5 7. A method according to any preceding claim, further comprising receiving sensor data updates from the plurality of user devices attending the event; and
 updating the one or more distributions of the user devices on the basis of the updated sensor data.

10 8. A method according to any preceding claim, wherein selecting the representative user device for a cluster comprises determining the best quality audio from among the user devices in the cluster; and
15 determining the best audio scene perspective among the user devices in the cluster by comparing location of interesting parts of the event and orientation of the user devices for a majority of interesting parts.

20 9. A method according to any preceding claim, further comprising uploading or up-streaming, within a particular cluster, captured media content from the selected representative user device only.

25 10. A method according to any preceding claim, further comprising synchronizing device clocks of the user devices recording media content from the event to each other.

 11. A method according to any preceding claim, wherein audio quality of an audio track captured by a user device is analyzed by
 choosing, for a given audio track, a random sampling position for
30 analyzing the quality of a small temporal segment of the audio;
 in response to the small segment being of good quality, choosing a subsequent sampling position for an analysis; and
 repeating the choosing of the subsequent sampling position for the analysis for a predetermined number of times, in response to the previously analysed
35 segment is of good quality; and
 determining the audio track to be of good quality.

 12. A method according to claim 11, wherein

choosing the subsequent sampling position is performed by using a half-interval search.

13. An apparatus comprising at least one processor, memory including computer program code, the memory and the computer program code configured to, with the at least one processor, cause the apparatus to at least:

receive sensor data from a plurality of user devices attending an event;

generate at least a spatial distribution of the user devices on the basis of the sensor data;

divide the user devices into clusters of one or more user devices on the basis of the spatial distribution; and

select a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

14. An apparatus according to claim 13, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

determine a threshold distance for each cluster such that the user devices located within the threshold distance belong to the same cluster.

15. An apparatus according to claim 14, wherein parameters affecting on the determination of the threshold distance for the clusters include one or more of the following:

- audio level of a common ambient audio scene in the event;
- regional variations of the audio level within the event;
- the number of users in the event;
- the size of the event venue;
- variation in density of the user devices.

16. An apparatus according to any of claims 13 - 15, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

generate a temporal distribution of the user devices on the basis of the sensor data.

17. An apparatus according to any of claims 13 - 16, wherein the sensor data includes at least one of the following:

- position information indicating the position of the user device;

- orientation information indicating the orientation of the user device with relation to the magnetic north;
- altitude information with relation to the horizontal;
- position information of the user device in 3-dimensional space in terms of the angles of rotation in three dimensions about the user device's center of mass;
- information whether the user device is recording or not recording media content from the event.

18. An apparatus according to any of claims 13 - 17, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

generate the one or more distributions of the user devices on the basis of the sensor data received from user devices recording content from the event.

19. An apparatus according to any of claims 13 - 18, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

receive sensor data updates from the plurality of user devices attending the event; and

update the one or more distributions of the user devices on the basis of the updated sensor data.

20. An apparatus according to any of claims 13 - 19, wherein for selecting the representative user device for a cluster the apparatus comprises computer program code configured to, with the at least one processor, cause the apparatus to at least:

determine the best quality audio from among the user devices in the cluster; and

determine the best audio scene perspective among the user devices in the cluster by comparing location of interesting parts of the event and orientation of the user devices for a majority of interesting parts.

21. An apparatus according to any of claim 13 - 20, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

upload or up-stream, within a particular cluster, captured media content from the selected representative user device only.

22. An apparatus according to any of claims 13 - 21, further comprising computer program code configured to, with the at least one processor, cause the apparatus to at least:

synchronize device clocks of the user devices recording media content from the event to each other.

23. An apparatus according to any of claim 13 - 22, wherein for analyzing audio quality of an audio track captured by a user device the apparatus comprises computer program code configured to, with the at least one processor, cause the apparatus to at least:

choose, for a given audio track, a random sampling position for analyzing the quality of a small temporal segment of the audio;

choose, in response to the small segment being of good quality, a subsequent sampling position for an analysis;

repeat the choosing of the subsequent sampling position for the analysis for a predetermined number of times, in response to the previously analysed segment is of good quality; and

determine the audio track to be of good quality.

24. An apparatus according to claim 23, wherein choosing the subsequent sampling position is arranged to be performed by using a half-interval search.

25. A computer program embodied on a non-transitory computer readable medium, the computer program comprising instructions causing, when executed on at least one processor, at least one apparatus to:

receive sensor data from a plurality of user devices attending an event;

generate at least a spatial distribution of the user devices on the basis of the sensor data;

divide the user devices into clusters of one or more user devices on the basis of the spatial distribution; and

select a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

26. A computer readable storage medium stored with code thereon for use by an apparatus, which when executed by a processor, causes the apparatus to perform:

receiving sensor data from a plurality of user devices attending an event;

generating at least a spatial distribution of the user devices on the basis of the sensor data;

dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and

5 selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

27. A system comprising at least one processor, memory including computer program code, the memory and the computer program code configured to, with the at least one processor, cause the system to at least:

receive sensor data from a plurality of user devices attending an event;

generate at least a spatial distribution of the user devices on the basis of the sensor data;

15 divide the user devices into clusters of one or more user devices on the basis of the spatial distribution; and

select a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

28. An apparatus comprising:

20 means for receiving sensor data from a plurality of user devices attending an event;

means for generating at least a spatial distribution of the user devices on the basis of the sensor data;

25 means for dividing the user devices into clusters of one or more user devices on the basis of the spatial distribution; and

means for selecting a representative user device for a cluster to represent the audio scene for the area of the cluster to which the representative user device belongs.

30 29. An apparatus according to claim 28, further comprising

means for determining a threshold distance for each cluster such that the user devices located within the threshold distance belong to the same cluster.

30. An apparatus according to claim 29, wherein

35 parameters affecting on the determination of the threshold distance for the clusters include one or more of the following:

- audio level of a common ambient audio scene in the event;
- regional variations of the audio level within the event;
- the number of users in the event;

- the size of the event venue;
- variation in density of the user devices.

5 31. An apparatus according to any of claims 28 - 30, further comprising means for generating a temporal distribution of the user devices on the basis of the sensor data.

10 32. An apparatus according to any of claims 28 - 31, wherein the sensor data includes at least one of the following:

- position information indicating the position of the user device;
- orientation information indicating the orientation of the user device with relation to the magnetic north;
- altitude information with relation to the horizontal;
- position information of the user device in 3-dimensional space in terms of the angles of rotation in three dimensions about the user device's center of mass;
- information whether the user device is recording or not recording media content from the event.

20 33. An apparatus according to any of claims 28 - 32, further comprising means for generating the one or more distributions of the user devices on the basis of the sensor data received from user devices recording content from the event.

25 34. An apparatus according to any of claims 28 - 33, further comprising means for receiving sensor data updates from the plurality of user devices attending the event; and means for updating the one or more distributions of the user devices on the basis of the updated sensor data.

30 35. An apparatus according to any of claims 28 - 34, further comprising means for determining the best quality audio from among the user devices in the cluster; and means for determining the best audio scene perspective among the user devices in the cluster by comparing location of interesting parts of the event and orientation of the user devices for a majority of interesting parts.

35

 36. An apparatus according to any of claims 28 - 35, further comprising means for uploading or up-streaming, within a particular cluster, captured media content from the selected representative user device only.

37. An apparatus according to any of claims 28 - 36, further comprising means for synchronizing device clocks of the user devices recording media content from the event to each other.

5 38. An apparatus according to any of claims 28 - 37, further comprising means for choosing, for a given audio track, a random sampling position for analyzing the quality of a small temporal segment of the audio;

 means for choosing, in response to the small segment being of good quality, a subsequent sampling position for an analysis; and

10 means for repeating the choosing of the subsequent sampling position for the analysis for a predetermined number of times, in response to the previously analysed segment is of good quality; and

 means for determining the audio track to be of good quality.

15 39. An apparatus according to claim 38, wherein said means for choosing the subsequent sampling position is arranged to performed a half-interval search.

1/5

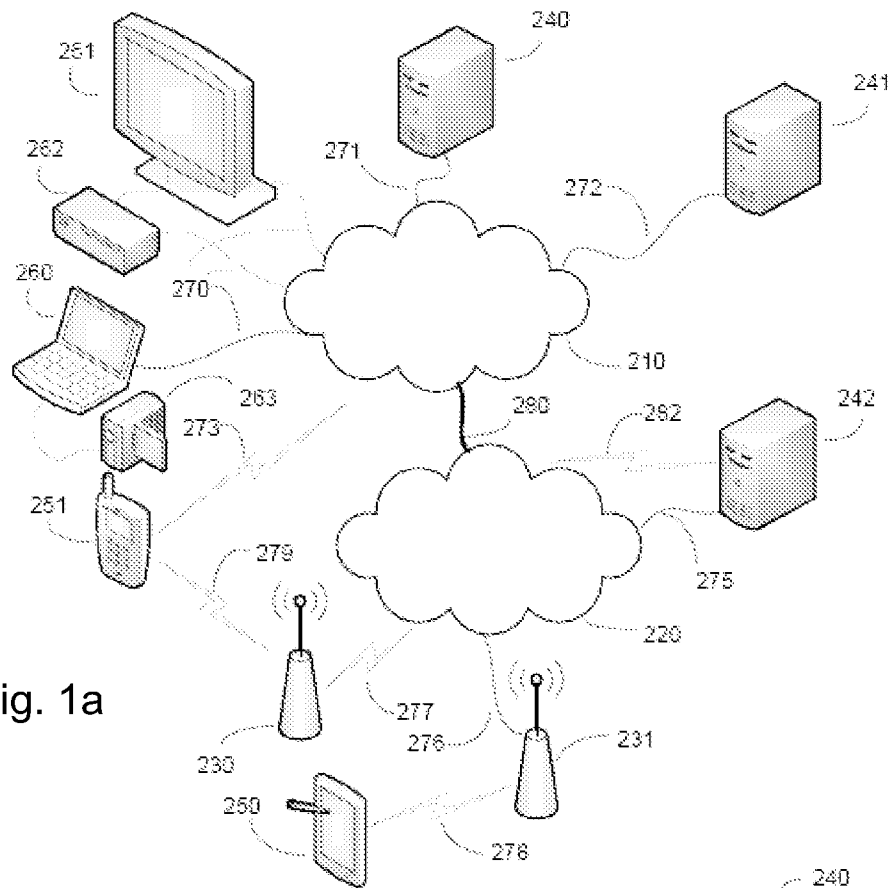


Fig. 1a

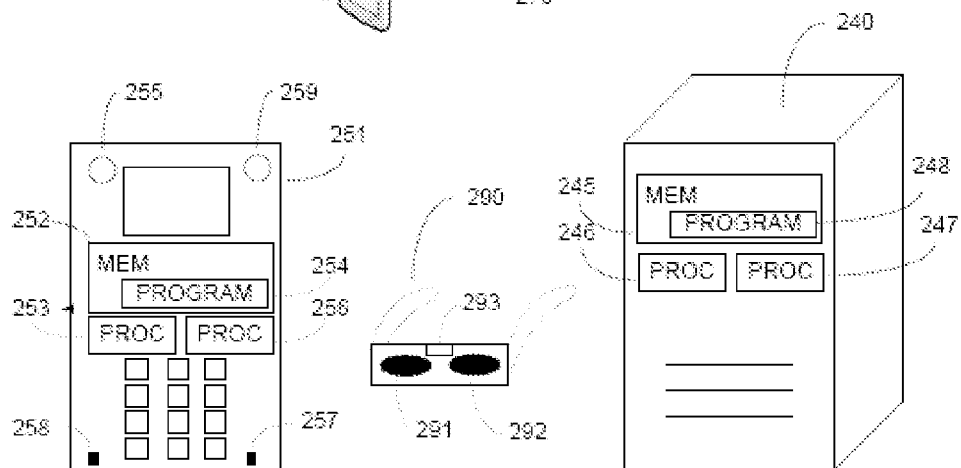


Fig. 1b

2/5

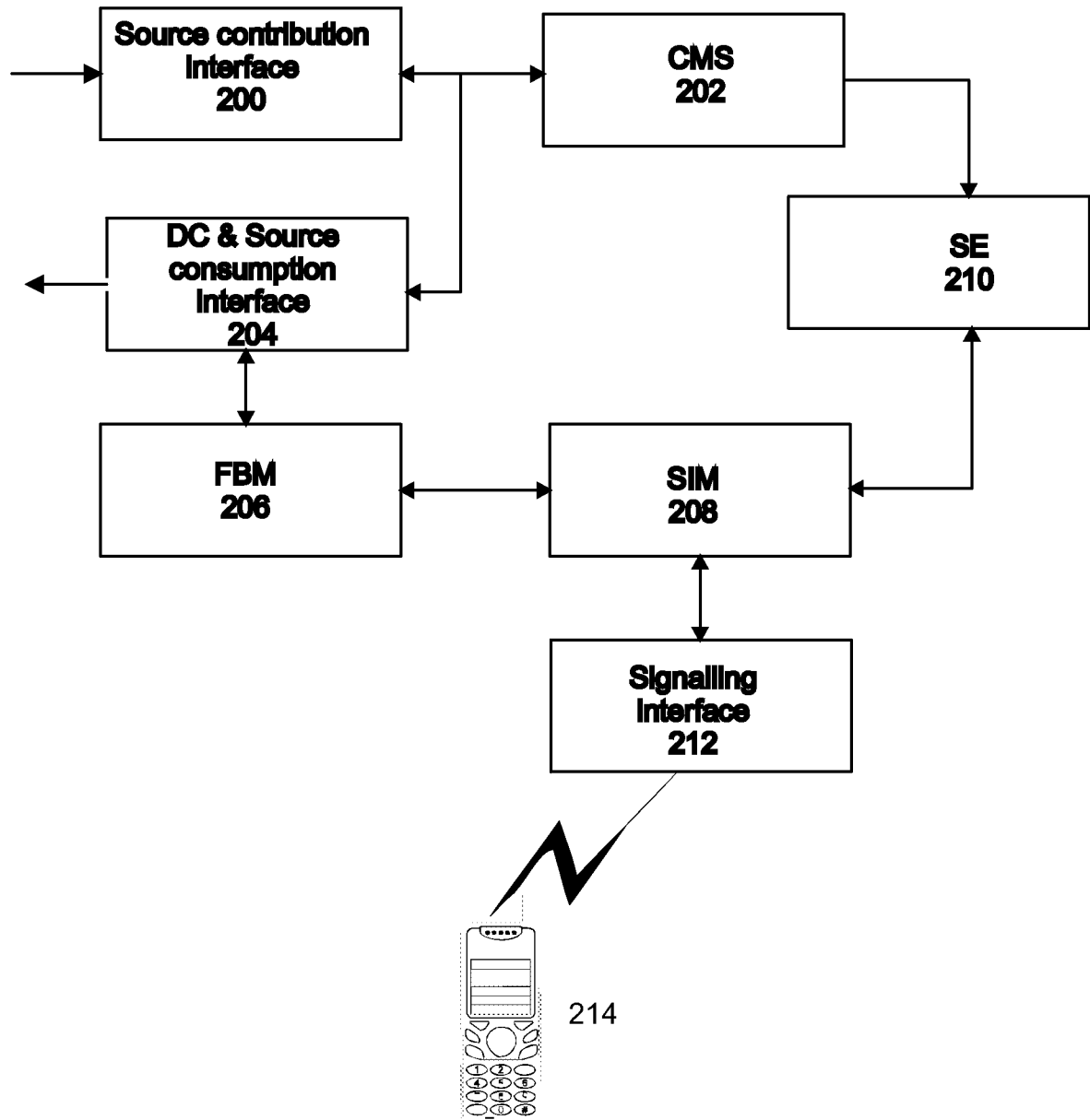


Fig. 2

3/5

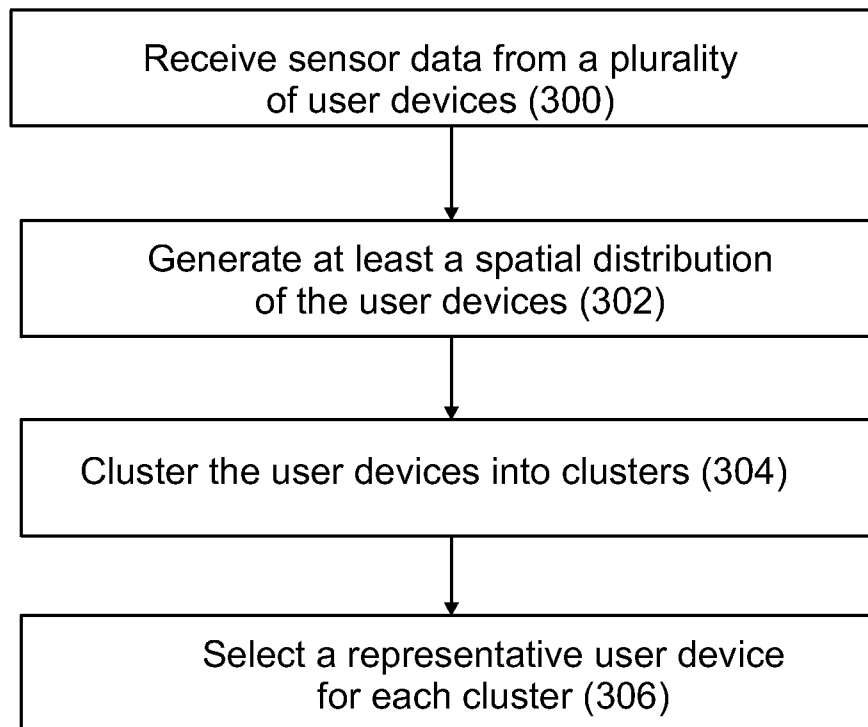


Fig. 3

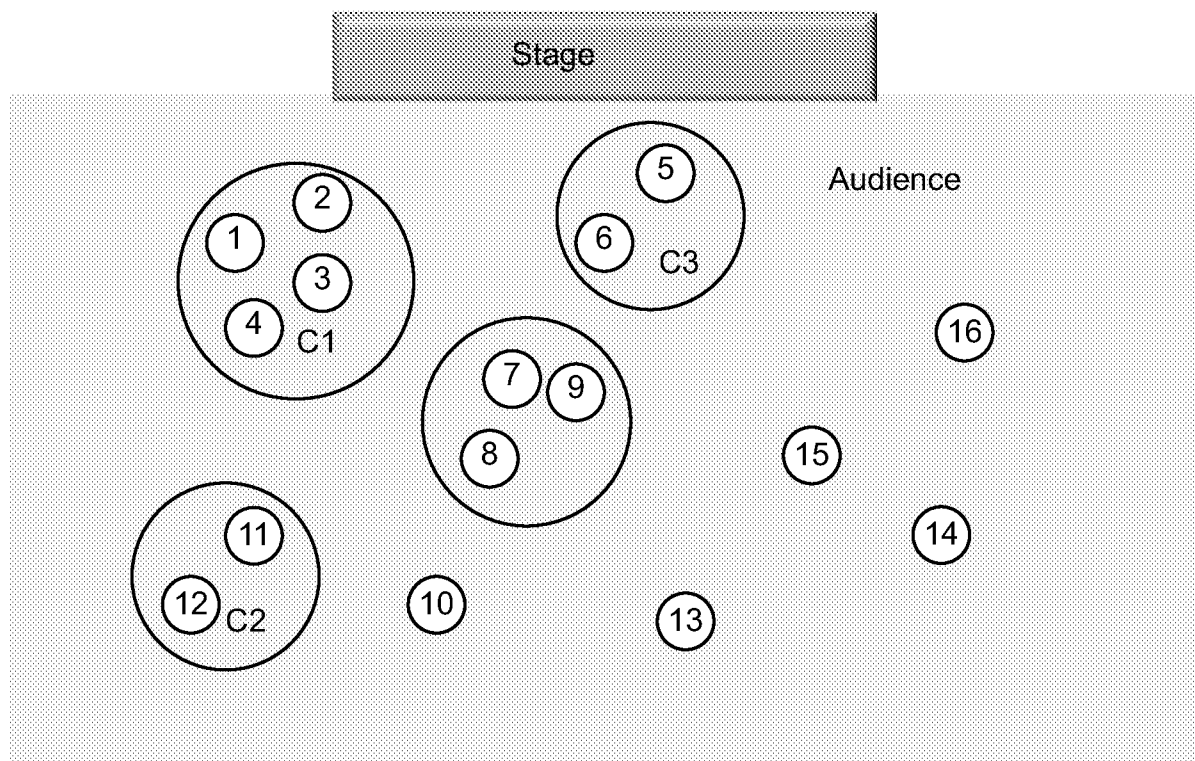


Fig. 4

4/5

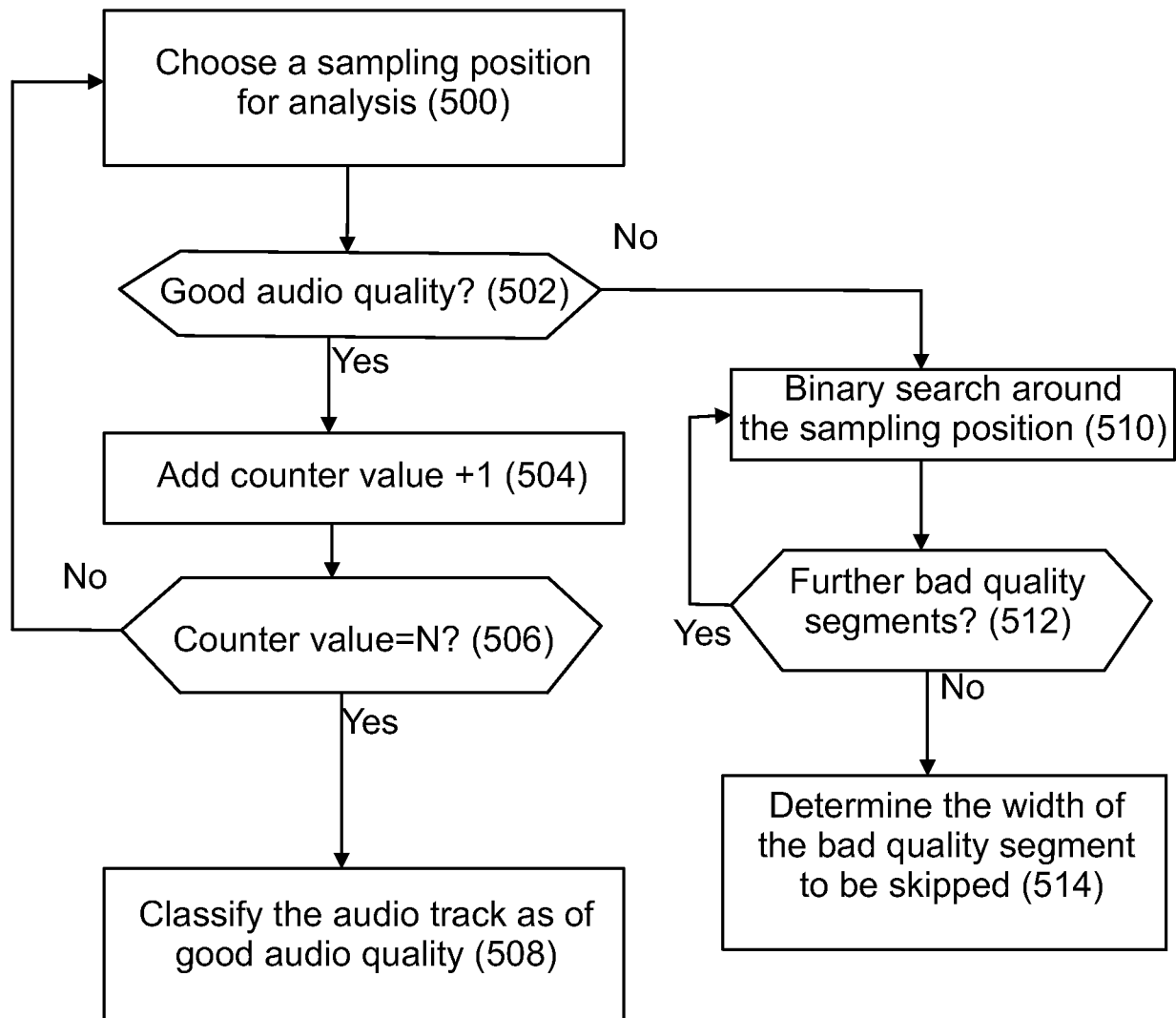
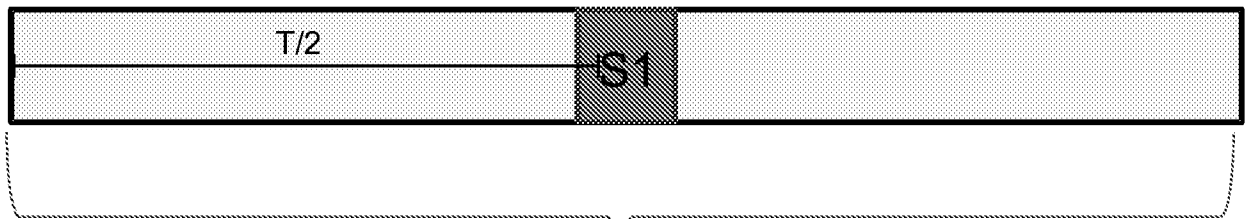


Fig. 5

5/5



Time interval T Fig. 6a

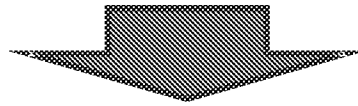
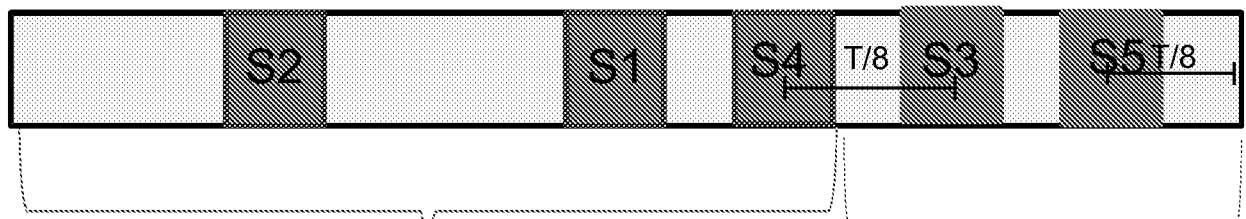
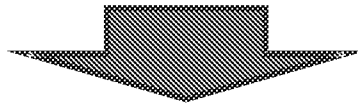


Fig. 6b



Good quality content

Content to be skipped

Fig. 6c

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI2012/051033

A. CLASSIFICATION OF SUBJECT MATTER

See extra sheet

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)

IPC: H04N, G10L, G06K, H04R, H04S, H04W, G11B, H04H, H04L, H04M, H04Q

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

FI, SE, NO, DK

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

EPO-Internal, WPI, BIOSIS, COMPDX, EMBASE, INSPEC, MEDLINE, PUBCOMP, PUBSUBS, TDB, NPL, XP3GPP, XPAIP, XPESP, XPESP2, XPETSI, XPI3E, XPIEE, XPIETF, XPIOP, XPIPCOM, XPJPEG, XPMISC, XPOAC, XPRD, XPTK, Google, Google Patents

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2012028902 A1 (NOKIA CORP [FI]) 08 March 2012 (08.03.2012) Figures 1-7; page 1, line 17 – page 2, line 12; page 12, lines 1-2, 7-20; page 13, lines 17-26; page 14, lines 4-22; page 18, lines 21-29; page 19, line 20 - page 22, line 4; page 22, lines 13-15; page 23, lines 16-24; page 24, lines 4-13; page 25, lines 21-28; page 27, lines 4-11	1-39
X	HIMAWAN, Ivan; McCOWAN, Iain; SRIDHARAN, Sridha; 'Clustering of ad hoc microphone arrays for robust blind beamforming', In: Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pages 2814-2817, ISBN 978-1-4244-4296-6. the whole document, especially Figs. 1-2; abstract; Secs. 2-3	1-3, 5, 13-15, 17, 25-30, 32
A	EP 1613124 A2 (POLYCOM INC [US]) 04 January 2006 (04.01.2006) the whole document, especially Figs. 1-3; abstract	



Further documents are listed in the continuation of Box C.



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

14 August 2013 (14.08.2013)

Date of mailing of the international search report

16 August 2013 (16.08.2013)

Name and mailing address of the ISA/FI
National Board of Patents and Registration of Finland
P.O. Box 1160, FI-00101 HELSINKI, Finland

Facsimile No. +358 9 6939 5328

Authorized officer

Janne Viljas

Telephone No. +358 9 6939 500

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI2012/051033

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 6850496 B1 (KNAPPE MICHAEL E [US] et al.) 01 February 2005 (01.02.2005) the whole document, especially Figs. 1-3, 8; abstract	
A	US 2005193421 A1 (CRAGUN BRIAN J [US]) 01 September 2005 (01.09.2005) the whole document, especially Figs. 1-6; abstract; paragraphs [0023], [0062]	
A	US 2009087161 A1 (ROBERTS DALE T [US] et al.) 02 April 2009 (02.04.2009) the whole document, especially Figs. 1-4, 7; abstract; paragraphs [0014], [0020]	
A	US 2010183280 A1 (BEAUREGARD GERALD THOMAS [SG] et al.) 22 July 2010 (22.07.2010) the whole document, especially Figs. 1-2, 6; abstract; paragraph [0008]	
A	WO 2009026347 A1 (UNIV SYRACUSE [US]) 26 February 2009 (26.02.2009) the whole document, especially Figs. 1, 3; abstract; paragraph [00025]	

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2012/051033

CLASSIFICATION OF SUBJECT MATTER

Int.Cl.

H04R 5/027 (2006.01)

H04R 29/00 (2006.01)

G06K 9/62 (2006.01)

H04N 5/265 (2006.01)

H04M 3/56 (2006.01)

H04R 3/00 (2006.01)

H04H 60/04 (2008.01)

H04W 72/04 (2009.01)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/FI2012/051033

Patent document cited in search report	Publication date	Patent family members(s)	Publication date
WO 2012028902 A1	08/03/2012	CN 103180907 A EP 2612324 A1	26/06/2013 10/07/2013
EP 1613124 A2	04/01/2006	CN 1716986 A CN 100484177 C JP 2006020314 A JP 4255461 B2 TW I264934 B US 2006013416 A1	04/01/2006 29/04/2009 19/01/2006 15/04/2009 21/10/2006 19/01/2006
US 6850496 B1	01/02/2005	None	
US 2005193421 A1	01/09/2005	None	
US 2009087161 A1	02/04/2009	EP 2206114 A1 JP 2010541415 A WO 2009042858 A1	14/07/2010 24/12/2010 02/04/2009
US 2010183280 A1	22/07/2010	WO 2010068175 A2	17/06/2010
WO 2009026347 A1	26/02/2009	AU 2008288928 A1 CA 2697233 A1 EP 2181507 A1 US 2009068943 A1 US 8301076 B2 US 2013039496 A1	26/02/2009 26/02/2009 05/05/2010 12/03/2009 30/10/2012 14/02/2013