US 20120198339A1

(19) United States

(12) Patent Application Publication

Williams et al.

(10) Pub. No.: US 2012/0198339 A1

(43) Pub. Date: Aug. 2, 2012

(54) **AUDIO-BASED APPLICATION ARCHITECTURE**

(76) Inventors: **Hunter Williams**, Seattle, WA (US); **William D. Carr**, Seattle, WA (US); **Kevin J. Gasper**, Seattle, WA (US); **Cameron Janes**, Seattle, WA (US); **Andrew J. Watts**, Seattle, WA (US); **James H. Wood**, Seattle, WA (US)

(21) Appl. No.: **13/016,009**

(22) Filed: **Jan. 28, 2011**

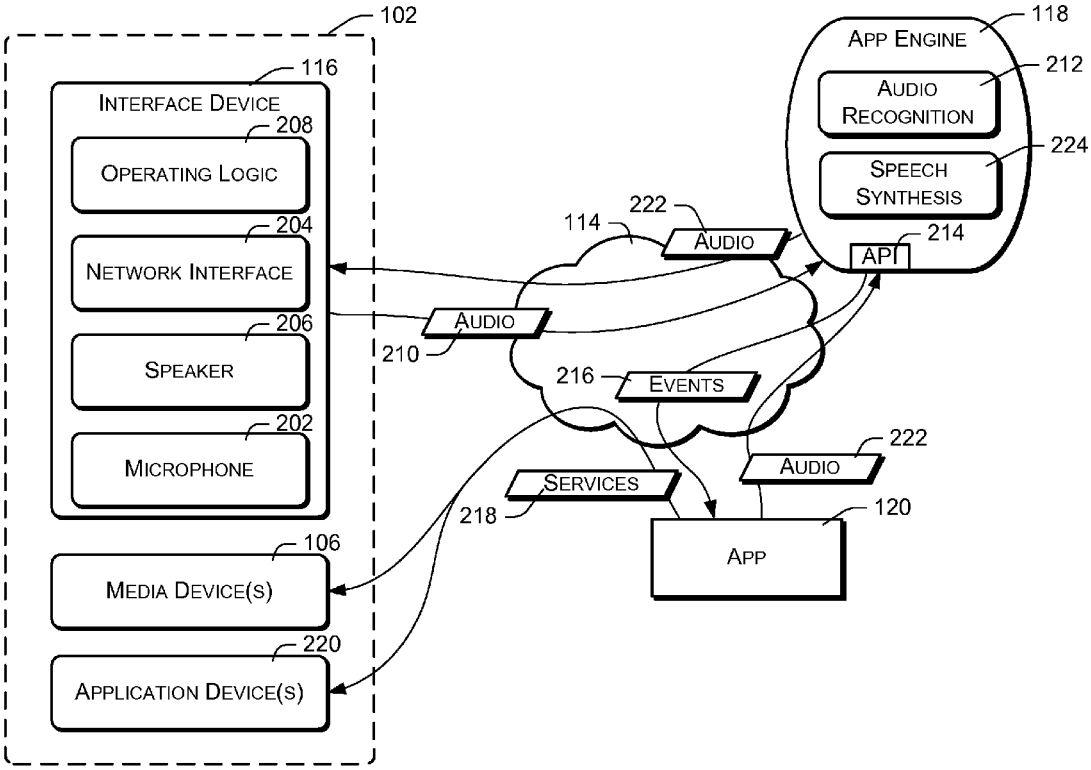
Publication Classification

(51) **Int. Cl.**
G06F 3/16 (2006.01)
G06F 15/16 (2006.01)

(52) **U.S. Cl.** **715/727**; 709/223

(57) **ABSTRACT**

An application architecture comprises one or more audio interfaces placed within the premises of users. A cloud-based application engine receives audio information from the interfaces and provides information to cloud-based applications based on the audio within the user premises. The other applications utilize the information to provide or enhance services to the users.



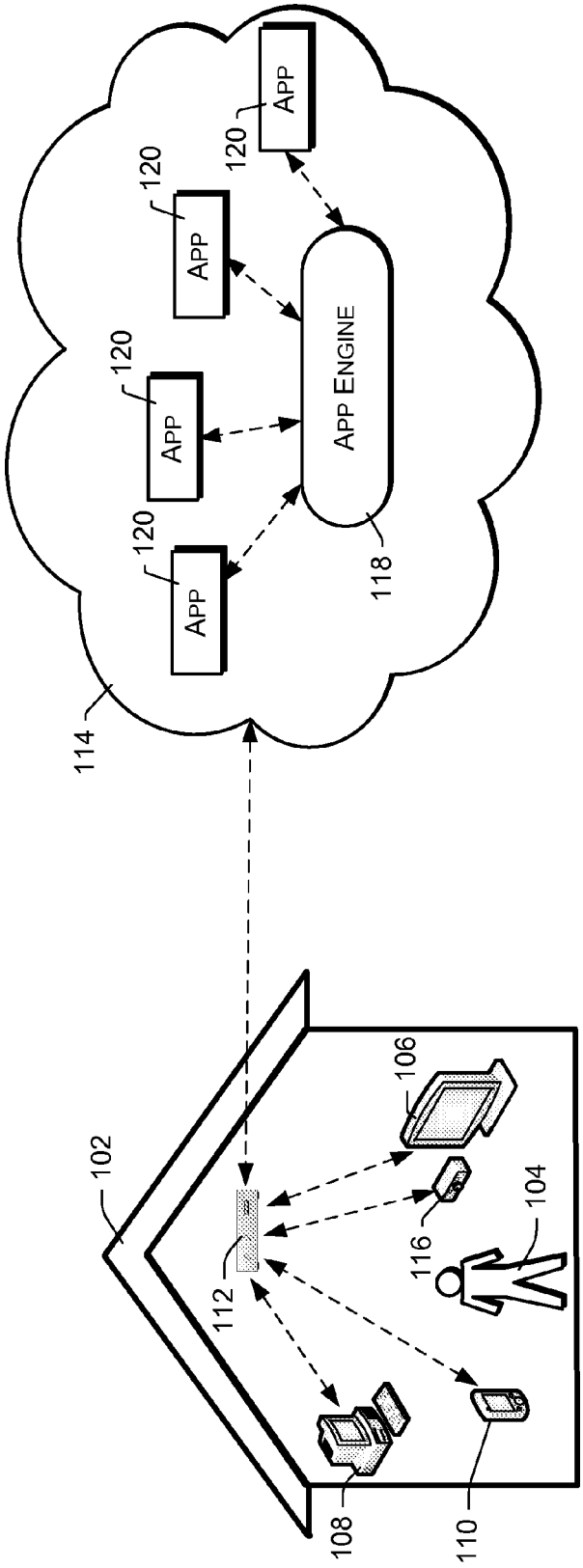

114
APP
120
APP
120
APP
120
APP
120
APP ENGINE
118
102
112
106
116
104
108
110

FIG. 1

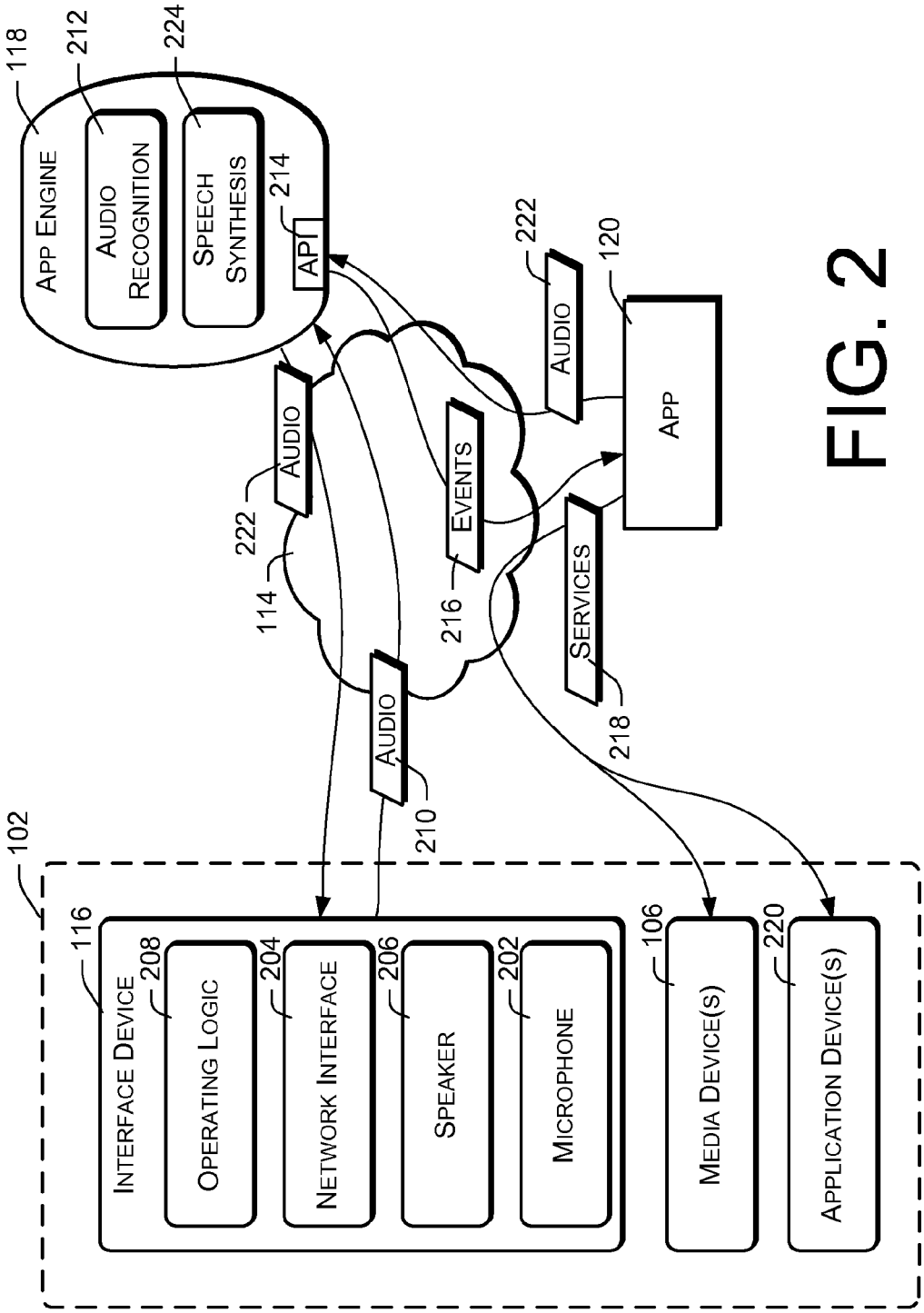

118
212
224
App Engine
Audio Recognition
Speech Synthesis
214
API
222
Audio
120
App
Audio
222
114
Events
216
Services
218
Audio
210
102
116
Interface Device
208
Operating Logic
204
Network Interface
206
Speaker
202
Microphone
106
Media Device(s)
220
Application Device(s)

FIG. 2

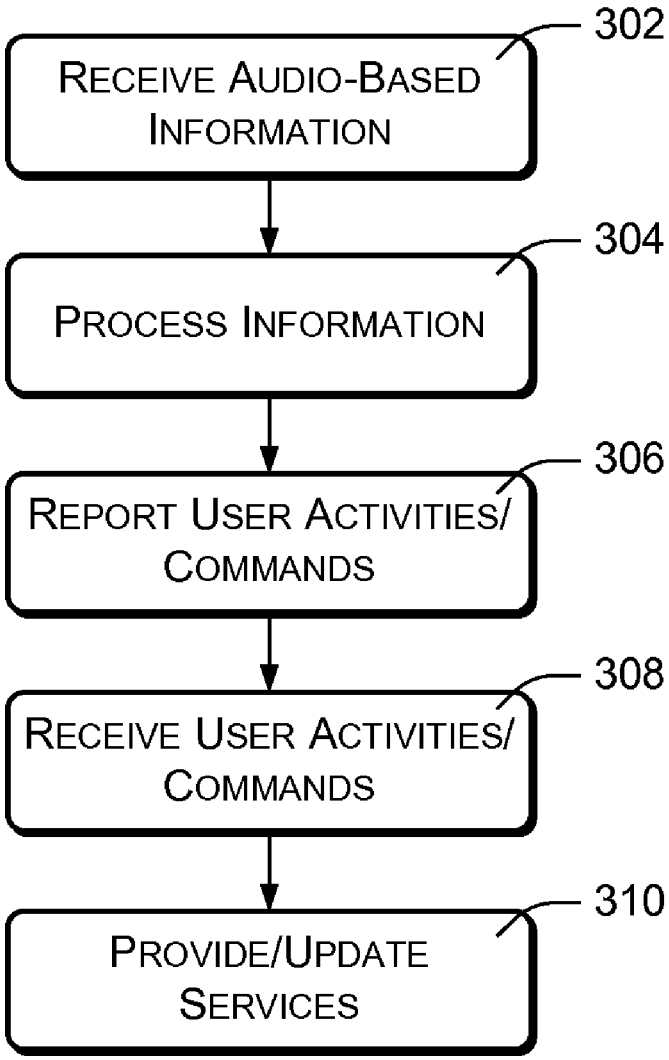

302
RECEIVE AUDIO-BASED INFORMATION
304
PROCESS INFORMATION
306
REPORT USER ACTIVITIES/ COMMANDS
308
RECEIVE USER ACTIVITIES/ COMMANDS
310
PROVIDE/UPDATE SERVICES

FIG. 3

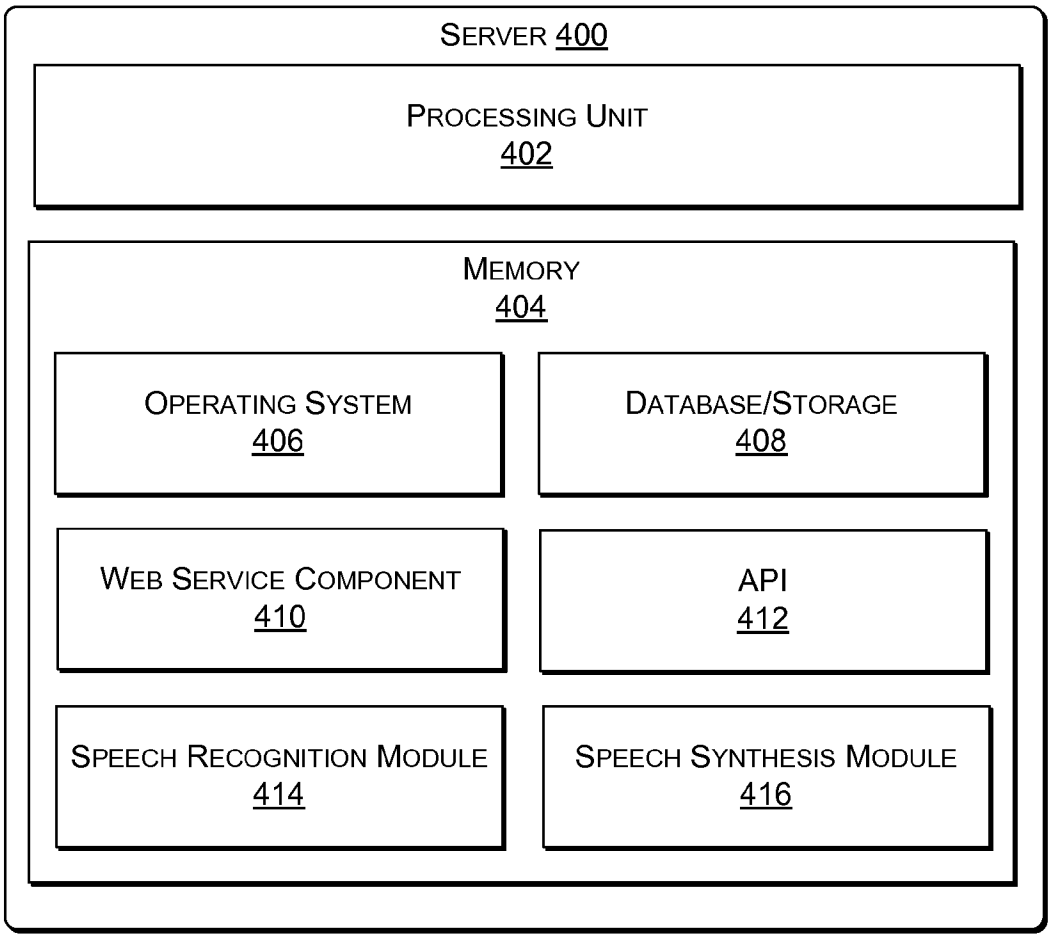

SERVER 400
PROCESSING UNIT
402
MEMORY
404
OPERATING SYSTEM
406
DATABASE/STORAGE
408
WEB SERVICE COMPONENT
410
API
412
SPEECH RECOGNITION MODULE
414
SPEECH SYNTHESIS MODULE
416

FIG. 4

AUDIO-BASED APPLICATION ARCHITECTURE

BACKGROUND

[0001] Homes and other user premises are increasingly equipped with always-on Internet or "cloud" connectivity. In many cases, even mobile users have constant or nearly constant data connectivity. The common availability of network communications has created a number of new possibilities for services and other functionality, using the variety of connected devices accessible to users.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] The detailed description is set forth with reference to the accompanying figures. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical items.

[0003] FIG. **1** is a block diagram illustrating an operational environment in which applications may utilize an application engine, and may respond to audio from user premises to provide various services to users.

[0004] FIG. **2** is a block diagram illustrating additional details regarding the system of FIG. **1**.

[0005] FIG. **3** is a flow diagram illustrating an example process that may be implemented within the environment of FIGS. **1** and **2**.

[0006] FIG. **4** is a block diagram illustrating components of a server that may be used to implement many of the techniques described herein.

DETAILED DESCRIPTION

[0007] Described herein are systems and techniques for utilizing audio information that may already exist within a home or other location. To collect the audio information, network-enabled microphones can be placed at different activity centers within a home or user premises. The microphones may be incorporated in small, self-contained units, with wireless networking capabilities, and configured to transmit audio-based information to a cloud-based application engine or service. The cloud-based application engine may monitor the audio-based information and process it to identify events, status, or other information about current activities within the home. The application engine may also identify commands that are spoken by users within the home.

[0008] In the described embodiment, the application engine exposes an application interface that other cloud-based entities can access. The application interface allows the other entities to receive data regarding in-home activities, status, commands, and so forth. Authorization and authentication safeguards are employed to ensure that only entities authorized by users are able to access this information.

[0009] Authorized cloud-based applications may utilize the services of the application engine to enhance their own services to users. For example, a cloud-based application might communicate with the application engine to receive notifications of commands spoken by users, and may act on those commands. As a more specific example, a user may speak a command such as "purchase movie tickets for Rambo tonight." The application engine, upon receiving this audio, might convert it to a textual stream and provide it to an application that purchases movie tickets for users using payment information that the user has previously provided and authorized for use in response to the engine receiving certain voice-based commands. The application, in response to receiving the textual command, may purchase the requested movie tickets and transmit them electronically to a device associated with the user, such as the user's computer or smartphone.

System Architecture Example

[0010] FIG. **1** illustrates an environment **100** in which these techniques might be practiced. The environment includes a home or other user premises **102**. User premises may include houses, offices, automobiles, and other spaces. Although a single home **102** is illustrated in FIG. **1**, most implementations will include multiple premises and users.

[0011] Within the home **102** is a user **104** and several devices associated with the user **104**. The illustrated devices include a media device **106**, representing any one or more of various types of devices such as televisions, audio players, video players, and so forth. The illustrated devices also include a computer **108**, which represents one or more of various types of devices such as laptop computers, desktop computers, tablet computers, netbooks, other network-enabled devices, and so forth. A personal computing device **110** may also be associated with the user, such as a smartphone, pager, PDA (personal digital assistant), book reader device, or other type of portable device, including various types of electronic communications devices.

[0012] Note that the devices shown within the home **102** are merely examples of a wide variety of devices that may be present within a user premises or associated with a user. Many such devices may have some sort of network connectivity. In this case, the in-home devices connect to a router or access point **112**, which is in turn connected to a public network **114**. Connection between in-home devices and the router **112** may be through wired or wireless technologies.

[0013] Although the devices within the home **102** are illustrated as connecting to the network **114** via the in-home router **112**, network connectivity might be implemented in various other ways. For example, some in-home devices may communicate with the network **114** using cellular data technologies, or through other types of wired and wireless technologies that do not involve an in-home router or access point. Thus, the particular configuration shown in FIG. **1** is presented merely as an illustration of one possible way in which network connectivity may be achieved.

[0014] The network **114** represents what has come to be known as the "cloud," which may include a vast array of network-based entities, servers, clients, applications, and services. Blocks within the network **114** represent cloud-based applications and/or services that are relevant to this discussion.

[0015] The network **114** may be a conglomeration of various discrete networks and systems, interconnected in various ways so that for functional purposes they appear more or less as a unified network. The network **114** may thus include various types of communication networks, including local-area networks, wide-area networks, the Internet, wireless networks, cable television networks, telephone networks, cellular communications networks, etc. The techniques discussed herein may also be practiced in smaller networks, including localized and/or private networks.

[0016] The user's home **102** may be equipped with one or more on-premises audio monitoring devices **116**, referred to

herein as in-home audio interface devices. An audio interface device **116** may in some embodiments comprise a device having a microphone and a network interface. In certain embodiments, the interface device may also have a speaker or other form of audio output.

[0017] The interface device **116** may be relatively small, so that several such devices can be unobtrusively placed at various locations within the home **102**. As an example, interface devices can be implemented in a small, cylindrical enclosures, resembling hockey pucks.

[0018] Each interface device **116** may have a power connection to receive operating power. In the described embodiment, other connections are not needed. In some embodiments, however, audio may be provided to one or more audio interface devices **116** from an external source; connectors, receptacles, or ports may be available on the interface devices **116** for this purpose.

[0019] The audio interface devices **116** can be placed within the home **102** so that their microphones detect ambient sound or noise within the home **102**. It may be convenient in many situations to place the audio interface devices near media devices within the home **102**, such as the media device **106**, so that audio produced by such media devices will be detected and clearly received by the audio interface devices **116**. It may also be convenient to place the interface devices near activity centers, such as in the kitchen, on desks, and so forth, so that user speech can be clearly received.

[0020] In the embodiment shown in FIG. **1**, the interface devices **116** communicate with and transmit environmental information to a cloud-based application service or engine **118**, via the router **112** and the network **114**. As mentioned above, such cloud-based communication might be implemented in many ways, and is not limited to the particular router-based configuration shown in FIG. **1**. As an example, the interface devices might use cellular or other wireless data communications technologies to communicate with the application engine **118**.

[0021] FIG. **1** also illustrates a plurality of cloud-based applications or services **120** that interact with the application engine **118** and that base their services at least in part on information obtained or derived from the in-home audio interface devices **116**. The applications **120** may be web-based applications that are accessible to users via a network browser, operating in conjunction with the computer **108** or another Internet terminal or access device. Various mobile devices and communications devices can also be used to access such applications. Some of the applications **120** may serve as resources for programs that execute on local devices such as the computer **108** or the personal device **110**. Such local programs may operate in conjunction with the cloud-based applications **120** to provide various different services and functionality, ranging from email services to games. Cloud-based or network-based applications such as the applications **120** may communicate and operate in conjunction with a wide variety of network-connected devices, including entertainment devices, game boxes, media devices, home automation and security devices, communications devices, and so forth.

[0022] In operation, the application engine **118** receives audio-based information from the interface devices **116**. The audio information may include audio streams, user commands or notifications derived from vocal utterances, environmental information derived from on-premises audio, and so forth. The application engine **118** processes the audio-based information to determine various data about user activities, status, environment, commands, etc. This data can then be obtained by authorized applications **120**, which can act on the information to provide or enhance services to the user **104**.

[0023] FIG. **2** illustrates additional details regarding interactions between the described devices and services. As mentioned above, one or more in-home audio interface devices **116** may be located within the home **102**. An individual interface device **116** may include a microphone **202** that is configured to detect ambient noise, sounds, and speech. The interface device **102** may also include a network interface **204** that facilitates network communications with various entities, including the application engine **118** and one or more applications **120**, via the network **114**. The interface device may also include a speaker **206**, and operating logic **208**. The operating logic **208** may be implemented as one or more programmable processing units, associated memory, and appropriate instructions stored in the memory and executed by the processing unit. Other types of logic may also be used to configure the interface device **116** to perform the functions described here.

[0024] In one embodiment, the interface device **116** is configured to send audio-based information **210** to the application engine **118**, based on ambient audio detected by the microphone **202** of the interface device **116**. The audio-based information **210** may comprise a continuous audio stream, or may comprise separate audio streams corresponding to different periods of activity or noise within the home **102** and nearby the interface device **116**. For example, an audio stream may be transmitted whenever ambient noise levels exceed a minimum threshold.

[0025] In other embodiments, the interface device **116** may pre-process the audio from the microphone **202**, and may transmit higher level audio-based information to the application engine **118**. For example, the operating logic **208** of the interface device **116** might perform speech recognition on the ambient audio, and transmit interpreted commands or text streams to the application engine **118**. As another example, the operating logic **208** might be configured to recognize and identify music that is being played by one of the media devices **106**, and to transmit data to the application engine **118** indicating or identifying the music currently being played. In some embodiments, the interface device **116** may utilize online or cloud-based services, such as music databases, to perform its functions.

[0026] Note that individual interface devices **116** can be associated with respective premises through some type of initialization or registration procedure, and communications with interface devices can be authenticated using cryptographic techniques.

[0027] The application engine **118** is configured to receive the audio-based information **210**. Depending on the type or format of the received information, the application engine **118** processes it to identify and create event, status, and/or command data relating to in-home activities. In some embodiments, the application engine **118** may keep a record of data objects or event descriptors, each representing and identifying a discrete event or item of information. As examples, event descriptor objects may be created to identify or indicate the following types of information:

[0028] spoken commands uttered by users;

[0029] songs that are being played within the home;

[0030] television shows that are being viewed within the home;

[0031] movies or other media items that are being presented or rendered;

[0032] the percentage of completion of currently rendered media items.

[0033] Information such as this can be identified by means of sound, voice, and/or music recognition technologies. For example, voice recognition technologies can be used to convert user speech into text, which can in turn be parsed to identify particular commands or instructions. Music recognition technologies can compare songs to known databases to identify the songs, and possibly the performers and other data regarding the songs. Similar technologies can be used to compare audio against databases of other know media such as television shows, podcasts, movies, and so forth. In addition to identifying particular media items being played, the application engine may identify or monitor the particular point within a media item that is currently being played.

[0034] As mentioned above, the audio may be processed within the interface device **116**, and the application engine may receive high-level notifications of events, media, etc. However, the audio-based information **210** received by the application engine **118** may comprise a raw audio stream, which may be processed by the application engine **118** to derive event descriptors. Thus, the application engine may include audio recognition logic **212**, such as speech-to-text conversion logic and other sound recognition functionality to recognize various types of sounds and media.

[0035] In the described embodiment, the application engine **118** exposes an application interface **214** (also referred to as an application programming interface or API) to the cloud-based applications **120**. The API **214** allows an application **120** to obtain event descriptors that have been created by the application engine **118**. With appropriate authentication, the application **120** may query the application engine **118** through the API **214** to obtain and receive event descriptors or notifications **216** regarding events and conditions within the home **102**. Authorization for an application **120** to receive information regarding a particular home or installation will typically be granted by a user associated with the home—only those applications authorized by the user are allowed to access the in-home information.

[0036] The application **120** may use the information from the event notifications **216** to initiate or customize services **218** on behalf of users associated with the home **102**. Such services may be provided through any of the media devices **106** located in the home, as well as through other application devices **220**, including computers **108**, personal devices **110**, and other devices available to users. As mentioned above, the services **218** may be provided as web-based services that are accessible through web or Internet browsers, in conjunction with any suitable hardware. Services may also be provided in conjunction with other types of systems, including telephone systems, media players, automation systems and devices, and so forth. Examples of applications **120** and associated services **218** will be described below.

[0037] The API **214** of the application engine **118** may also allow the application **120** to provide audio **222** for eventual transmission to the interface device **116**, to be played on the speaker **206**. The audio may be in the form of text, in which speech synthesis or text-to-speech conversion may be performed by the application engine **118** or by the operating logic **108** of the interface device **116**. In FIG. **2**, the application engine **118** is illustrated as having a speech synthesis module **224**. Alternatively, the audio **222** may be provided from the application **120** as a digital audio stream.

[0038] The various network communications between the components of FIG. **2** may be performed using existing networking protocols and techniques. Event descriptors **216** can be implemented in various ways, such as by the use of existing database technologies, and may be communicated using database query protocols or other protocols. Furthermore, some implementations may use various push technologies to actively transmit descriptors **216** to the applications **120** as they are identified. Audio and audio-based information may be formatted as files, audio streams, data objects, notifications, text streams, and so forth.

[0039] FIG. **3** shows actions that may be performed in accordance with the embodiments described herein. An action **302** comprises receiving audio-based information from premises associated with one or more users, such as a home, office, automobile, etc. The audio-based information may comprise an audio stream or data derived from an audio stream. Audio may be received by an in-home or on-premises interface, by a cloud-based service or application engine, by other cloud-based services and applications, or combinations of these entities.

[0040] At **304**, the audio-based information is processed to create data structures identifying current user-specific activities, events, status, environment, commands, and media within the user premises. Data structures may comprise database records or other data objects. Processing **304** may include speech recognition, speech-to-text conversion, music recognition, and other types of audio recognition, and may result in text-based or code-based notifications or commands.

[0041] At **306**, the identified events, commands, and so forth are reported to applications that have been authorized by the user. The events and commands may be reported using database communication protocols or other protocols, using either push or pull models.

[0042] At **308**, applications receive reported event and command notifications. At **310**, the applications provide and/or enhance their services based on the notifications. For example, applications may act in response to received commands, or may provide certain information to a user that relates to the detected environment within the user premises.

Example Services

[0043] The architecture above can be utilized to provide a wide variety of services and functionality, of which the following are simple examples.

[0044] A cloud-based application may be configured to received vocal commands from users and to perform operations based on those commands. For example, a user may speak a grocery order, and the spoken order may be interpreted using the voice recognition functionality described above. Once interpreted, the order may be submitted to a grocery store for subsequent delivery or simply recorded as a shopping list that the user can view on their smartphone when shopping at the grocery store. An application such as this may apply personalized settings for each order, such as preferred delivery schedules and historical portion sizes.

[0045] Similarly, a cloud-based application may allow users to order tickets without needing a computer. A user may tell the application what movie they want to see. The appli-

cation may use pre-configured information about the user (such as location) to determine which theater is near the user. The application may respond to a spoken command to order movie tickets by using the speech synthesis functionality of the system to query the user regarding movie selections and times.

[0046] An application may monitor ambient noise from the user's environment to determine what media (music, movie, television, etc.) is currently being rendered. Once the media is identified, the application may trigger an extension application on the user's smartphone or computer, which may provide instant access to rich metadata, recommendations, links, user communities, special interest groups, discussion boards, and so forth.

[0047] A similar type of application may identify the song the user is currently listening to, and in response may provide information about related articles, offer samples, suggest related music and playlists, or present information allowing the user to purchase individual music tracks.

[0048] An application may provide general information in response to users' uttered questions. A user may ask a question, such as "What's the recipe for deviled eggs?" The application may respond by emailing or otherwise providing a recipe for deviled eggs. In some cases, the application may use the speech synthesis functionality of the system to provide audio answers to spoken questions.

[0049] Similarly, the user may ask for directions to a certain location. The application may send a map, textual directions, and computerized routes to the user via the user's personal communications device, or may send information to the user's navigation device regarding the desired location.

[0050] As another example, such an application may allow the user to inquire about a flight or other event. The system might respond by sending information to the user via a device associated with the user, or by audibly responding using speech synthesis.

[0051] An application might also be designed to allow a user to control various on-premises devices and equipment, such as media equipment and home-automation devices. For example, such an application may accept spoken commands from a user regarding on-premises devices, and may respond by interacting with such devices through available network connections.

Server Example

[0052] FIG. **4** illustrates relevant components of a server **400** that may be used to implement the functionality of the application engine **118** and/or any of the applications **120**. Generally, any of these entities may be implemented by one or more servers, with the various functionality described above distributed in various ways across the different servers. Servers may be located together or separately, and organized as virtual servers, server banks, and/or server farms. The described functionality may be provided by the servers of a single entity or enterprise, or may utilize the servers and/or services of multiple entities or enterprises.

[0053] In a very basic configuration, an example server **400** may comprise a processing unit **402** composed of one or more processors, and memory **404**. Depending on the configuration of the server **400**, the memory **404** may be a type of computer storage media and may include volatile and nonvolatile memory. Thus, the memory **404** may include, but is not limited to, RAM, ROM, EEPROM, flash memory, or other memory technology.

[0054] The memory **404** may be used to store any number of functional components that are executable by the processing unit **402**. In many embodiments, these functional components comprise instructions or programs that are executable by the processing unit **402**, and that when executed implement operational logic for performing the actions attributed above to the content service **108** and the ingestion service **110**. In addition, the memory **404** may store various types of data that are referenced by executable programs, including content items that are supplied to consuming devices such as electronic reader **104**.

[0055] Functional components stored in the memory **404** may include an operating system **406** and a database **408** to store various items related to functionality provided by the server **400**. Functional components of the server **400** may also comprise a web service component **410** that interacts with remote devices such as computers, media consumption devices, and so forth.

[0056] The memory **404** may also have instructions implementing one or more APIs **412**, which may comprise the API **214** described with reference to FIG. **2**, configured to interface with applications **120** as described. The memory may also include a speech recognition module **414** and a speech synthesis module **416**, providing functionality as also described above.

[0057] The server **400** may of course include many other logical, programmatic, and physical components that are not shown in FIG. **4**.

CONCLUSION

[0058] The various techniques described above are assumed in the given examples to be implemented in the general context of computer-executable instructions or software, such as program modules, executed by one or more computers or other devices. Generally, program modules include routines, programs, objects, components, data structures, etc. for performing particular tasks or implement particular abstract data types.

[0059] Other architectures may be used to implement the described functionality, and are intended to be within the scope of this disclosure. Furthermore, although specific distributions of responsibilities are defined above for purposes of discussion, the various functions and responsibilities might be distributed and divided in different ways, depending on particular circumstances.

[0060] Similarly, software may be stored and distributed in various ways and using different means, and the particular software storage and execution configurations described above may be varied in many different ways. Thus, software implementing the techniques described above may be distributed on various types of computer-readable media, not limited to the forms of memory that are specifically described.

[0061] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described. Rather, the specific features and acts are disclosed as illustrative forms of implementing the claims. For example, the methodological acts need not be performed in the order or combinations described herein, and may be performed in any combination of one or more acts.

1. A system comprising:
one or more network interfaces;
a cloud-based service that is accessible by on-premises audio monitoring devices to receive audio-based information via the one or more network interfaces from multiple user premises, the user premises being associated respectively with users; and
wherein the cloud-based service exposes one or more application interfaces to multiple cloud-based applications that provide services to the users based at least in part on the audio-based information.

2. The system of claim **1**, wherein the audio-based information comprises audio streams.

3. The system of claim **1**, wherein the audio-based information comprises user commands derived from vocal utterances of the users.

4. The system of claim **1**, wherein the audio-based information comprises environmental information derived at least in part from on-premises audio.

5. The system of claim **1**, wherein the one or more application interfaces provide information regarding current user-specific activity based at least in part on the audio-based information.

6. The system of claim **1**, wherein the one or more application interfaces provide information regarding current user media consumption based at least in part on the audio-based information.

7. The system of claim **1**, wherein the one or more application interfaces provide information regarding current user video consumption based at least in part on the audio-based information.

8. The system of claim **1**, wherein the one or more application interfaces provide information regarding current user audio consumption based at least in part on the audio-based information.

9. The system of claim **1**, wherein the cloud-based service is responsive via the one or more application interfaces to provide audio to the on-premises audio monitoring devices at least partly in response to the cloud-based applications.

10. A method comprising:
under control of one or more computing systems configured with executable instructions,
receiving audio-based information from on-premises audio monitoring devices located on premises of multiple users;
processing the audio-based information to determine current user-specific information; and
reporting the current user-specific information to multiple network-based applications that provide services to the users based at least in part on the user-specific information.

11. The method of claim **10**, wherein the audio-based information comprises continuously-received audio streams from the multiple users within their respective premises.

12. The method of claim **10**, wherein the audio-based information comprises audio streams, the method further comprising processing the audio streams to identify the current user-specific information.

13. The method of claim **10**, wherein the audio-based information comprises audio streams, the method further comprising processing the audio streams to identify current user media consumption.

14. The method of claim **10**, wherein the audio-based information comprises audio streams, the method further comprising performing voice recognition on the audio streams to identify user commands.

15. The method of claim **10**, wherein the audio-based information specifies current user media consumption.

16. The method of claim **10**, further comprising providing audio to the on-premises audio monitoring devices at least partly in response to the network-based applications.

17. A network-based system, comprising:
one or more processors;
memory containing instructions that are executable by the one or more processors to perform actions comprising:
receiving information from a network-based service regarding current activities of multiple users in multiple user premises, wherein the information is based at least in part on audio monitoring from within the user premises; and
providing network-based services to the users based at least in part on the received information regarding the activities in their respective user premises.

18. The system of claim **17**, wherein the received information indicates current media consumption of individual users.

19. The system of claim **17**, wherein:
the received information indicates current media consumption of individual users; and
the network-based services supplement the current media consumption.

20. The system of claim **17**, wherein:
the received information indicates spoken commands from individual users; and
the network-based services act in response to the spoken commands.

21. An in-home interface device, comprising:
a microphone configured to receive audio from a nearby user and from nearby media devices;
a network interface configured to communicate with a network-based application service;
wherein the in-home interface device is configured to continuously transmit environmental information to one or more network-based applications via the network interface and the network-based application service, and wherein the environmental information is based at least in part on the received audio.

22. The device of claim **21**, wherein the environmental information comprises an audio stream.

23. The device of claim **21**, further comprising a speech recognizer, and wherein the environmental information comprises notifications derived at least in part from user utterances using the speech recognizer.

24. The device of claim **21**, further comprising a speech recognizer, and wherein the environmental information comprises media identifications.

* * * * *