April 16, 2018

Systems and Methods for Identifying a Speaker and/or Their Attention with Cameras and/or Microphones

Seth Raphael

SYSTEMS AND METHODS FOR IDENTIFYING A SPEAKER AND/OR THEIR ATTENTION WITH CAMERAS AND/OR MICROPHONES

Introduction:

The present disclosure provides systems and methods for improving an automated response to an audible query. Automated assistant devices (e.g., "smart speakers") that can listen and respond to an audible query from a user are becoming increasingly common. Typically, an automated assistant device will listen for an audible query that begins with a "hotword". The device can detect when a user speaks the hotword, as a signifier that a query will follow, and the device can listen for the rest of the query. In many cases, an automated assistant device can be configured to control one or more "smart devices" (e.g., lights, speakers, televisions, refrigerators, microwave ovens, dishwashers, coffee machines, or other devices/appliances) that can be connected to the assistant device via a communication network (e.g., LAN, PAN, WWW, etc.). The device can use various processing techniques to comprehend the query and determine an appropriate response, such as, for example, turning on/off a light, start/stop a microwave oven, play music from a speaker, etc. The present disclosure enables a computing system (e.g., an automated assistant device), and methods for controlling the same, to determine a source location of an audible query, and determine a context of the audible query based on the source location. The computing system can use the context to better comprehend the query in order to determine a response that is relevant to the context.

Summary:

According to aspects of the present disclosure, the computing system can include two or more microphone devices, and/or one or more camera devices. The computing system can use the microphone and/or camera devices to determine a source location of an audible query. The computing system can determine a context of the audible query based in part on the source

location. The computing system can also use the one or more camera devices and/or microphones to identify a source of the audible query. The computing system can determine a context of the audible query based in part on the identified source. The computing system can also use the one or more camera devices and/or microphones to identify an object that is the subject of the attention of the user that uttered the audible query. For example, the computing system can identify an object that such user is looking at. The computing system can determine an appropriate or contextual response based at least in part on the identity of the object that is the subject of the attention of the user.

Detailed Description

According to aspects of the present disclosure, the computing system can include two or more microphone devices. The computing system can use the microphone devices to detect an audible query spoken by a user. The microphone devices can be associated with different locations within an environment (e.g., building, house, room, backyard, or other defined space).

In some implementations, the computing system can use the two or more microphones to determine a source location of an audible query. The computing system can determine a delay between when each of the two or more microphones detects the audible query, and use the delay to determine a source location of the hotword, relative to the two or more microphones. As an example, the computing system can include first and second microphone devices. If the computing system detects an audible query via first microphone device, and subsequently via the second microphone device, then the computing system can determine that a source location of the query is closer to the first microphone device than the second microphone device.

As another example, the computing system can include first, second, and third microphone devices. The computing system can detect a delay between each of the first, second,

and third microphone devices detects an audible query to triangulate a source location of the query.

In some implementations, the computing system can use a location associated with each of the two or more microphones to determine a source location of an audible query. As an example, the computing system can include a first microphone device associated with a bedroom location. If the computing system detects an audible query via the bedroom microphone device, then the computing system can determine that a source location of the audible query is the bedroom.

As another example, the computing system can include a first microphone device associated with a bedroom location and a second microphone device associated with a hallway location. If the computing system detects an audible query via the hallway microphone device, and subsequently via the bedroom microphone device, then the computing device can determine that a source location of the audible query is closer to the hallway than the bedroom. The computing system can also use the location associated with the hallway microphone device to determine that the source location is the hallway.

In some implementations, the computing system can determine a context of an audible query based on a source location of the query, and determine a response to the query based on the context. As an example, the computing system can detect an audible query including a hotword followed by “play music”. The computing system can determine a source location of the audible query as a living room location. The computing system can determine if there are any speaker devices associated with the living room location. If there is a speaker device associated with the living room location, then the computing system can control the speaker device to play music. If there is more than one speaker device associated with the living room location, then the

computing device can control all the speaker devices to play music, or request a user to specify one of the speaker devices.

As another example, the computing system can detect a hotword spoken by a first user, and determine a source location of the hotword. The computing system can then listen for a query following the hotword that originates from the source location. In this way, the computing system can distinguish the audible query from a background noise, such as, for example, one or more other users/people that are speaking simultaneously with the first user. The computing system can ignore speech/sounds originating from a location other than the determined source location of the hotword, so that the computing system does not confuse the speech/sounds from the other users/people as the query of the first user. For example, the computing system can use beam-forming techniques to isolate audio originating from the source location and, conversely, ignore or filter audio generated at other locations.

As another example, the computing system can detect an audible query to “turn on lights”. If the computing system determines a source location of the query as a bedroom, then the computing system can use the source location to determine that the query is referring to lights in the bedroom. The computing system can then turn on the bedroom lights.

As another example, the computing system can detect an audible query to “cook on for 5 minutes”. If the computing system determines a source location of the query as a kitchen, then the computing system can use the source location to determine that the query is referring to a microwave oven in the kitchen. The computing system can then activate the microwave oven for 5 minutes.

According to aspects of the present disclosure, the computing system can alternatively or additionally include one or more camera devices. The one or more camera devices can be

associated with different locations within an environment (e.g., building, house, room, backyard, or other defined space).

In some implementations, the computing system can use a location associated with each of the one or more camera devices to determine a source location of an audible query. As an example, the computing system can detect an audible query that is spoken by a user. If the computing system identifies the user via a first camera device associated with a living room location, then the computing system can determine a source location of the query as the living room. If the computing system identifies the user via a second camera device associated with a bedroom location, then the computing system can determine a source location of the query as the bedroom.

In some implementations, the computing system can use the one or more camera devices to identify a source of an audible query. As another example, the computing system can detect a hotword that is spoken by a user. The computing system can determine a source location of the hotword, and use one or more cameras associated with the source location to identify the user at the source location. The computing system can use the one or more cameras to track a location of the user as the user moves around. The computing system can use the tracked location of the user to only listen for a query following the hotword that is originating from the tracked location, and ignore sounds from other locations and/or other users at the other locations.

In some implementations, the computing system can determine a context of an audible query based on a source of the query, and determine a response to the query based on the context. As an example, the computing system can identify a user as a source of an audible query. The computing system can use the one or more camera devices to determine a direction that the user is facing and/or an object that is the subject of the user's attention (e.g., the user is

looking at the object, the user is pointing at the object, and/or the like), in order to comprehend the query. In particular, the computing system can create a three-dimensional model of the user's face and/or body (using the one or more camera devices), and determine the direction that the user is facing and/or the object of the user's attention based on a position and/or orientation of the model and its components (e.g., head, arm, hand, etc.). If the query is to "turn off" and the user is looking at a television in a living room, then the computing system can turn off the television in the living room. If the query is to "turn off" and the user is looking and/or pointing at a lamp in a bedroom, then the computing system can turn off the lamp in the bedroom.

As another example, the computing system can detect an audible query to "play music". The computing system can determine a source location of the audible query as a living room location. If the computing system determines that there is more than one speaker device associated with the living room location, then the computing device can determine which one of the speaker devices that the user is facing toward, and control the determined speaker device to play music.

As another example, the computing system can identify first and second users in an environment. The first user can be looking at a television in a living room location, and the second user can be looking at a microwave oven in a kitchen location. If the computing system identifies the first user as a source of an audible query to "turn off", then the computing system can determine that the query refers to the television in the living room location. In this way, even if the computing system identifies the second user looking at the microwave oven in the kitchen location, the computing system can determine that the audible query is not referring to the microwave oven. Thus, the systems described herein can disambiguate which of multiple people in a room made the command.

In some implementations, the computing system can use the one or more camera devices to detect an audible query without a hotword. As an example, the computing system can use the one or more cameras to identify a first user. If the computing system determines that the first user is looking at a speaker device, then the computing device can use this as a signifier that a query will follow that relates to the speaker device. The first user can speak a query without a hotword while looking at the speaker device. The computing system can process the query and play music on the speaker device even if the query did not start with the hotword. If the computing system determines that the first user is instead looking at a television when the computing system detects a query to "turn off" from the first user, then the computing system can process the query and turn off the television even if the query did not start with the hotword.

Figure 1 depicts a block diagram of an example computing system according to example embodiments of the present disclosure. Portions of the example computing system can be physically located within a physical structure 10. Example structures 10 include a building, house, vehicle, or other structures. A structure can also refer to a specific portion or division of a building such as different floors of the building; different office spaces within a building; or similar divisions of building space. For example, Company A's office space within a building can be a first structure while Company B's office space within the building can be a second structure.

As examples, an automated assistant device 102 and a plurality of smart devices 150a-c can be physically located within the structure 10. For example, the assistant device 102 can itself be a smart device and/or one of the smart devices 150a-c can operate as an automated assistant device. Example smart devices can include network-connected computing devices such as sound speakers, home alarms, door locks, microphones, cameras, lighting systems, treadmills, weight

scales, smart beds, irrigation systems, garage door openers, appliances (e.g., refrigerator, HVAC, dishwasher, stove, etc.), baby monitors, fire alarms, or other smart computing devices. These devices can offer or provide various services or operations. For example, the services or operations can be performed by computer application(s) executed by the device(s). Smart devices are not required to be network connected.

The assistant device 102 can communicate with the smart devices using a variety of different communications protocols, methods, hardware, etc., and combinations thereof. As one example, the assistant device 102 can communicate with the smart device 150a using short range wireless communications techniques such as Bluetooth, ZigBee, Bluetooth Low Energy, infrared signals, optical signals, etc. As another example, the control computing device can communicate with the smart devices 150b-c over a local area network 181. For example, the local area network 181 can be a WiFi network associated with the structure 10. The assistant device 102 can also communicate with smart devices using wired connections such as, for example, Ethernet connections.

The assistant device 102 and/or the plurality of smart devices 150a-c can also communicate with one or more computing devices external to the structure 10. For example, such computing devices external to the structure 10 can include one or more web servers 20, an additional smart device 160, and/or a registration server 30. For example, communications between devices located within the structure 10 and devices external to the structure can occur over a wide area network 182. For example, the wide area network 182 can include the Internet, cellular networks, or the like. Communications between devices located within the structure 10 and devices external to the structure can flow through the local area network 181 but are not required to do so.

Thus, the assistant device 102 can itself be one of the smart devices described above or can be a separate computing device with a primary purpose other than controlling the smart devices. As one example, the assistant device 102 can be a smart speaker that includes and implements an intelligent personal assistant. As another example, the assistant device 102 can be a smartphone.

A user can interact with the assistant device 102 to access or otherwise control one or more of the smart devices. A user can also interact with the assistant device 102 to control management of the one or more smart devices. To provide an example, the user may issue a voice command (e.g., an audible query) to the assistant device 102 that requests access to or control of the smart device(s) (e.g., "Turn the thermostat down two degrees."). The assistant device 102 can process the voice command, determine whether the user is authorized to control such device(s) and, if so, communicate with the device(s) (e.g., wirelessly via a network) to effectuate the actions requested by the voice command. As an alternative example, the user may directly interact with the smart device (e.g., the smart thermostat) and, in such instance, the smart device can be considered to be the assistant device 102.

Likewise, the assistant device 102 and/or one or more additional smart computing devices may operate to enable the user to engage with, manage, or otherwise control one or more applications and/or web services. As one example, a user may request that music be streamed and played via a music streaming application/web service executed by the assistant device 102 and/or one or more additional smart computing devices. As another example, a user may request to add a new event to a calendar managed by a calendar application executed by the assistant device 102 and/or one or more additional smart computing devices.

Drawings:

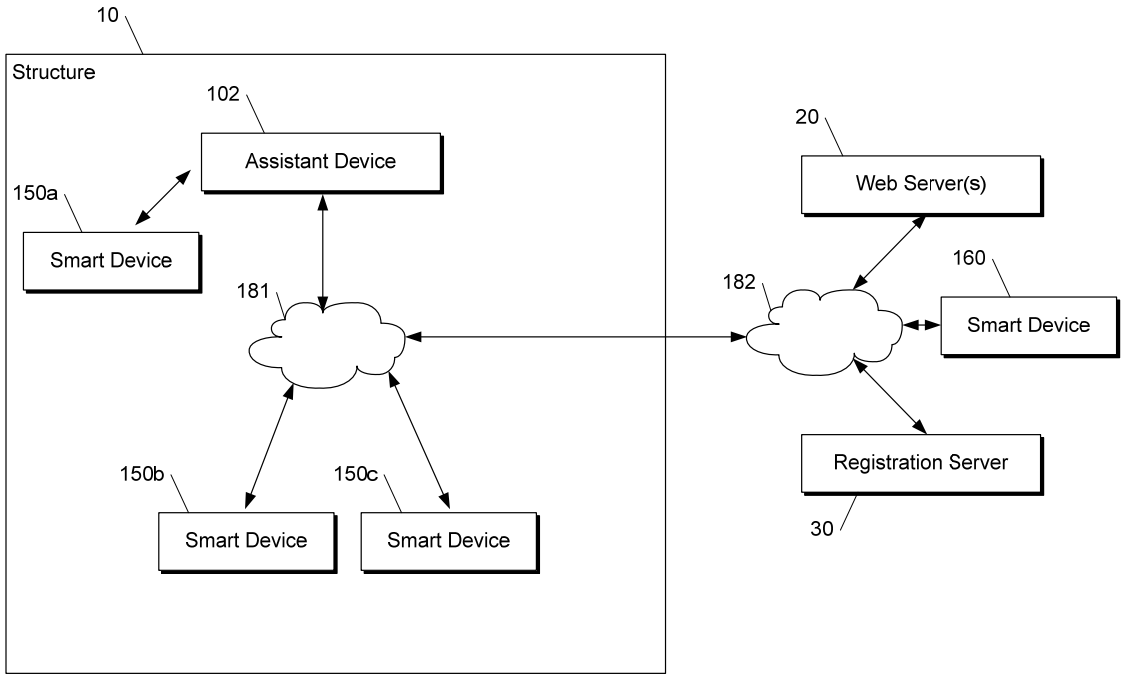

Figure 1

Abstract:

The present disclosure describes systems and methods that can determine a context of an audible query based at least in part on a source and/or a source location of the query. More particularly, the present disclosure enables a computing system, and methods for controlling the same, to determine a source location of an audible query, identify a source (e.g., a user/speaker) of the audible query, and determine a context for the audible query based in part on a subject of an attention of the source. The computing system can user the determined context to improve a response to the audible query. Keywords associated with the present disclosure include: computing systems (e.g., smartphone, smartwatch, mobile phone, automated assistant, home assistant, personal assistant); user; multiple users; hotword; query; context; context-aware; localization; triangulation; facial recognition; object tracking.