



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
17.09.2025 Bulletin 2025/38

(51) International Patent Classification (IPC):
G10L 15/22 ^(2006.01) **G10L 15/28** ^(2013.01)
H04R 1/08 ^(2006.01)

(21) Application number: **24382264.0**

(52) Cooperative Patent Classification (CPC):
G10L 15/28; G10L 15/22; H04R 1/08

(22) Date of filing: **12.03.2024**

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

(72) Inventors:
• **Pérez García, Guillermo**
41005 Sevilla (ES)
• **de Amores Carredano, José Gabriel**
41005 Sevilla (ES)
• **Cano Montes, Antonio**
41005 Sevilla (ES)
• **Castro Malet, Manuel**
41005 Sevilla (ES)

(71) Applicant: **4i Intelligent Insights S.L.**
41005 Sevilla (ES)

(74) Representative: **Clarke, Modet y Cia., S.L.**
C/ Suero de Quiñones 34-36
28002 Madrid (ES)

(54) **VOICE-BASED INTERACTION SYSTEM AND METHOD**

(57) Voice-based interaction system and method. The system (100) comprises a person location unit (8) configured to detect people in a zone of interest (4) and obtain a location (9a, 9b) of each person (11a, 11b) detected in the zone of interest (4); a plurality of directional microphones (5a, 5b); a speech recognition module (12) configured to receive an input audio signal (19a, 19b) from each one of the directional microphones (5a, 5b); a driving module (6) configured to move each direc-

tional microphone (5a, 5b) independently; and a control unit (10) configured to receive the location (9a, 9b) of each person (11a, 11b) detected in the zone of interest (4), assign each directional microphone (5a, 5b) to a different person (11a, 11b), and actuate the driving module (6) so that each directional microphone (5a, 5b) is directed towards the location (9a, 9b) of the assigned person (11a, 11b).

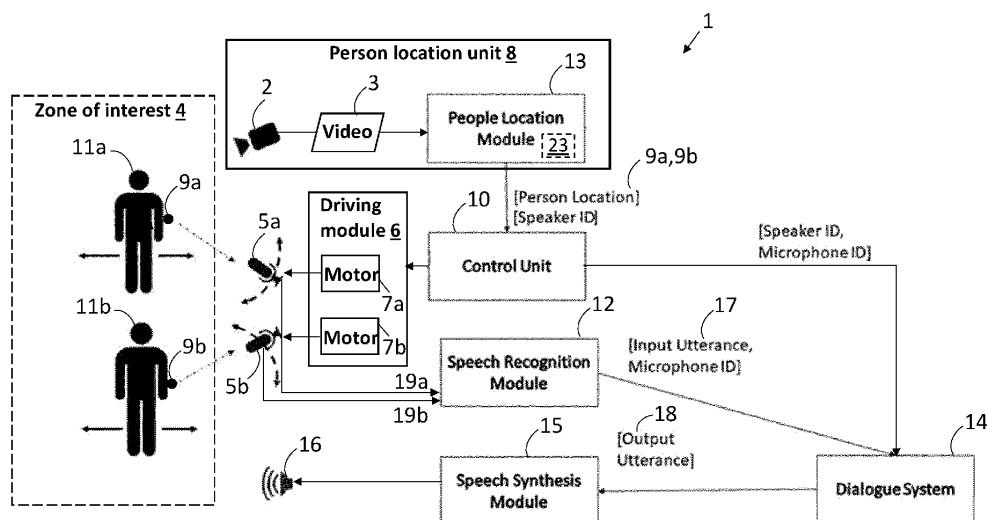


FIG. 1

Description

FIELD OF THE INVENTION

[0001] The present invention relates to voice-based interaction systems with speech recognition for human-machine interaction (HMI). The HMI may include sending voice commands to perform some actions (e.g. instruct a robot to move, order an intelligent loudspeaker to turn on the lights) or having a conversation with a machine, among other interactions.

BACKGROUND OF THE INVENTION

[0002] Existing voice-based interaction systems include speech recognition solutions with human-like performance. These systems range from smart speakers (like Amazon's Alexa) to full-fledged robots (like Hanson's Sophia).

[0003] However, all these solutions struggle when deployed in scenarios with multiple people speaking. They are often unable to distinguish which utterances are directed to the voice-based interaction system from other utterances from unrelated background conversations. Also, in multiparty conversations, they don't know who among the speakers said the last utterance and whether this utterance was directed to the voice-based interaction system or to another person in the group.

[0004] The present invention solves these problems, providing a voice-based interaction system that receives much cleaner audio signals from multiple speakers and thus has an improved speech recognition performance.

SUMMARY OF THE INVENTION

[0005] The present invention refers to a voice-based interaction system and method that solves the aforementioned problems.

[0006] The voice-based interaction system comprises a control unit; a person location unit configured to detect people in a zone of interest and obtain a location of each person detected in the zone of interest; a plurality of directional microphones; a speech recognition module configured to receive an input audio signal from each one of the directional microphones; a driving module configured to move each directional microphone independently. The control unit is configured to receive the location of each person detected in the zone of interest, assign each directional microphone to a different person, and actuate the driving module so that each directional microphone is directed towards the location of the assigned person.

[0007] The voice-based interaction method comprises detecting one or more persons in a zone of interest; obtaining a location of each person detected in the zone of interest; assigning each one of a plurality of directional microphones to a different detected person; aiming each directional microphone at the location of the correspond-

ing assigned person; and applying speech recognition to an input audio signal acquired by each one of the directional microphones to obtain an input utterance. The directional microphones employed by the system are a plurality of independent microphones (not a monolithic microphone array), each one with its own acquired audio signal.

[0008] The present invention improves the speech recognition since each speaker has a dedicated directional microphone pointing at him/her and, therefore, the speech recognition module receives much cleaner audio signals.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] A series of drawings which aid in better understanding the invention and which are expressly related with an embodiment of said invention, presented as a non-limiting example thereof, are very briefly described below.

Figure 1 depicts a voice-based interaction system according to an embodiment.

Figure 2 depicts a voice-based interaction system according to another embodiment.

Figure 3 depicts a flow diagram of a voice-based interaction method according to an embodiment.

DETAILED DESCRIPTION OF THE INVENTION

[0010] Figure 1 shows a voice-based interaction system 1 according to an embodiment. The voice-based interaction system 1 comprises:

- A person location unit 8 configured to detect people in a scene or zone of interest 4 and obtain a location (9a, 9b) of each person (11a, 11b) detected in the zone of interest 4.
- A plurality of directional microphones (5a, 5b).
- A speech recognition module 12, wired or wirelessly connected to the directional microphones (5a, 5b), and configured to receive an input audio signal (19a, 19b) from each one of the directional microphones (5a, 5b).
- Driving means, implemented in a driving module 6, configured to move each directional microphone (5a, 5b) independently, the directional microphones (5a, 5b) being aimed at configurable positions of the zone of interest 4.
- A control unit 10 configured to receive the location (9a, 9b) of each person (11a, 11b) detected in the zone of interest 4, assign each directional micro-

phone (5a, 5b) to a different person (11a, 11b), and actuate the driving module 6 so that each directional microphone (5a, 5b) is directed or aimed towards the location (9a, 9b) of the assigned person (11a, 11b).

[0011] In the example depicted in Figure 1, two persons (a first person 9a and a second person 9b) are detected by the person location unit 8 in the zone of interest 4, and the voice-based interaction system 1 includes a first directional microphone 5a and a second directional microphone 5b. The control unit 10 assigns each one of the directional microphones (5a, 5b) to a different person (11a, 11b); for example, as shown in Figure 1, the first directional microphone 5a is assigned to the first person 9a, and the second directional microphone 5b is assigned to the second person 9b.

[0012] Should the number of microphones be greater than the number of persons detected in the zone of interest 4, the exceeding microphones can become temporarily unassigned to a person or can be assigned to persons who already have an assigned microphone (in the latter case, a person may have more than one assigned microphone). For instance, when only one person is detected in the system of Figure 1, the first directional microphone 9a is assigned to said person, whereas the second directional microphone 9b remains unassigned. Alternatively, both microphones could be assigned to the detected person, and the speech recognition module may use the input audio signals (19a, 19b) coming from both directional microphones (5a, 5b) to improve the speech recognition.

[0013] If the number of microphones is lower than the number of persons detected in the zone of interest 4 (e.g. two microphones and three persons), the exceeding person(s) remain unassigned. In an embodiment, to determine the unassigned persons the person location unit 8 may include a vision module with one or more cameras 2 that acquire video images showing the face of the persons in the scene and a speak detector unit 23 that analyzes the video and detects whether each person detected in the video is speaking or is silent. A microphone is not assigned to the persons who are not speaking. When a person who has no microphone assigned starts speaking, the system preferably automatically assigns him/her a microphone applying a predefined policy. This policy could be, for instance, LRU (Least Recently Used, assigning the microphone of the speaker who has not spoken for a longer time), LFU (Least Frequently Used, assigning the microphone of the speaker that is speaking less in average) or any combination or alternative policy.

[0014] The person location unit 8 could be implemented in several ways. People detection performed by the person location unit 8 could be based on computer vision techniques, processing the video input from one or more cameras and applying people detection and person tracking algorithms, like current deep learning models based on large neural networks. Optionally, the cameras

could have depth estimation, infrared detection, or other advanced capabilities for better performance.

[0015] Alternatively, the person location strategy could be based on sound source estimation or even in indirect cues, like smartphone signal location (assuming that each person is carrying on such devices). The person location unit 8 may also combine any of these different strategies (computer vision techniques, sound source estimation techniques, etc.) to optimize the precision of the location module.

[0016] In the embodiment shown in Figure 1, the person location unit 8 comprises a camera 2 covering the area where the speakers may be located (i.e. covered area or zone of interest 4), although more than one camera 2 may be employed. The camera 2 is configured to capture video 3 from the zone of interest 4, and a people location module 13 is configured to detect and track people in the video 3 using any known detection and tracking algorithm.

[0017] The driving module 6 preferably comprises an actuator (7a, 7b) per directional microphone (5a, 5b). As depicted in Figure 1, the actuators (7a, 7b) may be, for instance, motors (e.g. stepper motors). The actuators (7a, 7b) can move the corresponding directional microphone (5a, 5b) independently from the other directional microphones, according to at least one degree of freedom, e.g. by rotating the microphone around a rotation axis or by moving the microphone between two or more linearly arranged positions. The actuators (7a, 7b) attached to the directional microphones (5a, 5b) could be of any type provided they have at least one degree of freedom, represented in Figure 1 with dashed lines. When activated, the actuators (7a, 7b) may change the position and/or orientation of the directional microphones (5a, 5b) such that they are aimed at configurable positions of the zone of interest 4.

[0018] In an embodiment, each actuator (7a, 7b) is configured to move the corresponding microphone (5a, 5b) according to at least two degrees of freedom including a pitch angle (i.e. rotation around the side-to-side axis of the microphone) and a yaw angle (i.e. rotation around a vertical axis). Three degrees of freedom (pitch, yaw, and roll) could also be used. Actuators with low resource consumption would be preferred for unplugged deployments like stand-alone robots. The chosen actuators preferably move smoothly, generating low background noise to minimize its impact on the speech recognition module 12.

[0019] The person location unit 8 may continuously detect and track people in the zone of interest 4, and the control unit 10 may thus also correspondingly update the position and/or orientation of the directional microphones (5a, 5b) such that they are continuously directed to the different detected people, following their moves within the zone of interest 4.

[0020] The speech recognition module 12, which is configured to convert the input audio signals (19a, 19b) coming from the directional microphones (5a, 5b)

into text (input utterance 17), could be any one in the market and employ any known speech recognition technique, like current solutions based on large language models.

[0021] Since each person (speaker, when talking) on the scene will have a dedicated directional microphone pointing at him/her, the voice-based interaction system 1 will receive much cleaner audio signals, improving the speech recognition performance of the speech recognition module 12. When receiving the input audio signals (19a, 19b) from the loudspeakers (5a, 5b), the speech recognition module 12 will use the audio signal with which the text conversion is better, i.e. the cleanest audio signal acquired by the microphone that is aimed at the speaker. In an embodiment, the people location module 13 may include a speak detector unit 23 that analyzes the video 3 and detects which person is speaking. In this case, the people detection module 13 outputs the speaker identification ("Speaker ID" in Figure 1) together with the location (9a, 9b) of each person detected in the video 3. For instance, the output may be: {(person 1, location 1), (person 2, location 2); speaker=person 2}. In this case, the speech recognition module 12 would use the input audio signal captured by the directional microphone assigned to the speaker (e.g. second input audio signal 19b).

[0022] In an embodiment, the people location module 13 comprises the speak detector unit 23 configured to repeatedly detect the person who is speaking (the speaker) in the video 3 and, if one or more speakers are detected, repeatedly send at least one speaker identification to the control unit 10. The control unit 10 is preferably configured to assign the directional microphones (5a, 5b) to the different persons in the scene based on the information received regarding the speaker identification, according to a predefined policy (e.g. considering the speaker identifications received in a predetermined time interval). The directional microphones can be assigned, for instance, giving preference to the persons who spoke the latest in the video (e.g. the two directional microphones in Figure 1 are assigned to the two persons in the zone of interest 4 who spoke the latest, the rest of the persons in the zone of interest 4 remaining temporarily unassigned). This configuration is useful when there are more persons (11a, 11b) in the zone of interest 4 than directional microphones (5a, 5b). This way, currently speaking persons can always have an assigned microphone, independently on the number of persons in the zone of interest 4 and the number of microphones present in the system 1.

[0023] The input utterance 17 may be a command sent by a user to the voice-based interaction system 1 for execution (e.g. a command to set a certain temperature in the room). When used for conversational purposes, the voice-based interaction system 1 may further comprise a dialogue system 14, a speech synthesis module 15 and at least one loudspeaker 16 (Figure 1 depicts only one loudspeaker). The dialogue system 14 may be config-

ured to receive the input utterance 17 (a text string) from the speech recognition module 12, generate an output utterance 18 (a text string) in response to the input utterance 17, and send the output utterance 18 to the speech synthesis module 15, which generates an output speech that is played in the loudspeaker 16. The dialogue system 14 could also be any one in the market and employ any known technique. A dialogue system 14 designed for multiparty conversations could benefit from the additional information provided by the control unit 12, which can assign a microphone channel to an individual speaker. The present invention provides a robust dialogue interaction by assigning moving microphones to speakers.

[0024] Figure 2 depicts another embodiment with an advanced feature including a person identification solution and a user profile stored as persistent memory. This additional feature would allow the voice-based interaction system 1 to remember users from previous conversations and use all the information previously gathered from that particular user in the dialogue.

[0025] In this embodiment, the voice-based interaction system 1 further comprises a user identification module 20, configured to identify each person (11a, 11b) detected in the zone of interest 4, and a user profile database 21. The dialogue system 14 is configured to retrieve information 22 of each identified person stored in the user profile database 21 and generate the output utterance 18 using the retrieved information 22.

[0026] The user identification could be based in biometric techniques like face identification or voice print. The user profile could be stored as any type of persistent data storage like a database of RDF triplets.

[0027] Figure 3 depicts a flow diagram of a voice-based interaction method 100 according to an embodiment of the present invention. The method 100 comprises detecting 110 one or more persons (19a, 19b) in a zone of interest 4; obtaining 120 a location (9a, 9b) of each person (11a, 11b) detected in the zone of interest 4; assigning 130 each one of a plurality of directional microphones (5a, 5b) to a different detected person (11a, 11b); directing or aiming 140 each directional microphone (5a, 5b) towards the location (9a, 9b) of the corresponding assigned person (11a, 11b); and applying 150 speech recognition to an input audio signal (19a, 19b) acquired by each one of the directional microphones (5a, 5b) to obtain an input utterance 17.

[0028] The step of detecting 110 one or more persons in the zone of interest 4 may comprises acquiring, by at least one camera 2, video 3 from the zone of interest 4; detecting persons in the video 3; and tracking the detected persons in the video 3. The method 100 may further comprise repeatedly detecting a speaker in the video 3, wherein the directional microphones (5a, 5b) are assigned 130 to the persons in the zone of interest 4 based on the detection of speaker(s) in the video 3 (e.g. taking into account the last X detected speakers, or the speakers detected in the last Y seconds, or the frequency

of each speaker in the last Z seconds).

[0029] The step of aiming 140 each directional microphone (5a, 5b) may comprise moving and/or rotating the directional microphone according to at least one degree of freedom, preferably according to at least two degrees of freedom including a pitch angle and a yaw angle.

[0030] The method 100 may further comprise generating an output utterance 18 in response to the input utterance 17; applying speech synthesis to the output utterance 18 to generate an output speech; and playing the output speech in at least one loudspeaker 16. The step of generating the output utterance 18 may comprise identifying each person (11a, 11b) detected in the zone of interest 4; retrieving information 22 of each identified person stored in a user profile database 21; and generating the output utterance 18 using the retrieved information 22.

Claims

1. A voice-based interaction system, comprising:

a person location unit (8) configured to detect people in a zone of interest (4) and obtain a location (9a, 9b) of each person (11a, 11b) detected in the zone of interest (4);
a plurality of directional microphones (5a, 5b);
a speech recognition module (12) configured to receive an input audio signal (19a, 19b) from each one of the directional microphones (5a, 5b);
a driving module (6) configured to move each directional microphone (5a, 5b) independently; and
a control unit (10) configured to:

receive the location (9a, 9b) of each person (11a, 11b) detected in the zone of interest (4);
assign each directional microphone (5a, 5b) to a different person (11a, 11b); and
actuate the driving module (6) so that each directional microphone (5a, 5b) is directed towards the location (9a, 9b) of the assigned person (11a, 11b).

2. The voice-based interaction system of claim 1, wherein the person location unit (8) comprises at least one camera (2) configured to capture video (3) from the zone of interest (4), and a people location module (13) configured to detect and track people in the video (3).

3. The voice-based interaction system of claim 2, wherein the people location module (13) comprises a speak detector unit (23) configured to repeatedly detect a speaker in the video (3) and send a speaker

identification to the control unit (10); and wherein the control unit (10) is configured to assign the directional microphones (5a, 5b) to the persons in the zone of interest (4) based on the information received regarding the speaker identification.

4. The voice-based interaction system of any preceding claim, wherein the driving module (6) comprises an actuator (7a, 7b) per directional microphone (5a, 5b), wherein each actuator (7a, 7b) is configured to move the corresponding directional microphone (5a, 5b) according to at least one degree of freedom.

5. The voice-based interaction system of claim 4, wherein each actuator (7a, 7b) is configured to move the corresponding directional microphone (5a, 5b) according to at least two degrees of freedom including a pitch angle and a yaw angle.

6. The voice-based interaction system of any preceding claim, further comprising a dialogue system (14), a speech synthesis module (15) and at least one loudspeaker (16); and wherein the dialogue system (14) is configured to:

receive from the speech recognition module (12) an input utterance (17);
generate an output utterance (18) in response to the input utterance (17); and
send the output utterance (18) to the speech synthesis module (15) for reproduction in the at least one loudspeaker (16).

7. The voice-based interaction system of claim 6, further comprising a user identification module (20) configured to identify each person (11a, 11b) detected in the zone of interest (4); and wherein the dialogue system (14) is configured to retrieve information (22) of each identified person stored in a user profile database (21) and generate the output utterance (18) using the retrieved information (22).

8. A voice-based interaction method, comprising:

detecting (110) one or more persons (11a, 11b) in a zone of interest (4);
obtaining (120) a location (9a, 9b) of each person (11a, 11b) detected in the zone of interest (4);
assigning (130) each one of a plurality of directional microphones (5a, 5b) to a different detected person (11a, 11b);
aiming (140) each directional microphone (5a, 5b) at the location (9a, 9b) of the corresponding assigned person (11a, 11b); and
applying (150) speech recognition to an input audio signal (19a, 19b) acquired by each one of

the directional microphones (5a, 5b) to obtain an input utterance (17).

9. The voice-based interaction method of claim 8, wherein detecting one or more persons in the zone of interest (4) comprises:

acquiring, by at least one camera (2), video (3) from the zone of interest (4);
detecting persons in the video (3); and
tracking the detected persons in the video (3).

10. The voice-based interaction method of claim 9, further comprising repeatedly detecting a speaker in the video (3); and wherein the directional microphones (5a, 5b) are assigned (130) to the persons in the zone of interest (4) based on the speaker detection in the video (3).

11. The voice-based interaction method of any of claims 8 to 10, wherein aiming each directional microphone (5a, 5b) comprises moving and/or rotating the directional microphone according to at least one degree of freedom.

12. The voice-based interaction method of claim 11, wherein aiming each directional microphone (5a, 5b) comprises moving and/or rotating the directional microphone according to at least two degrees of freedom including a pitch angle and a yaw angle.

13. The voice-based interaction method of any of claims 8 to 12, further comprising:

generating an output utterance (18) in response to the input utterance (17);
applying speech synthesis to the output utterance (18) to generate an output speech; and
playing the output speech in at least one loudspeaker (16).

14. The voice-based interaction method of claim 13, wherein generating the output utterance (18) comprises:

identifying each person (11a, 11b) detected in the zone of interest (4);
retrieving information (22) of each identified person stored in a user profile database (21);
and
generating the output utterance (18) using the retrieved information (22).

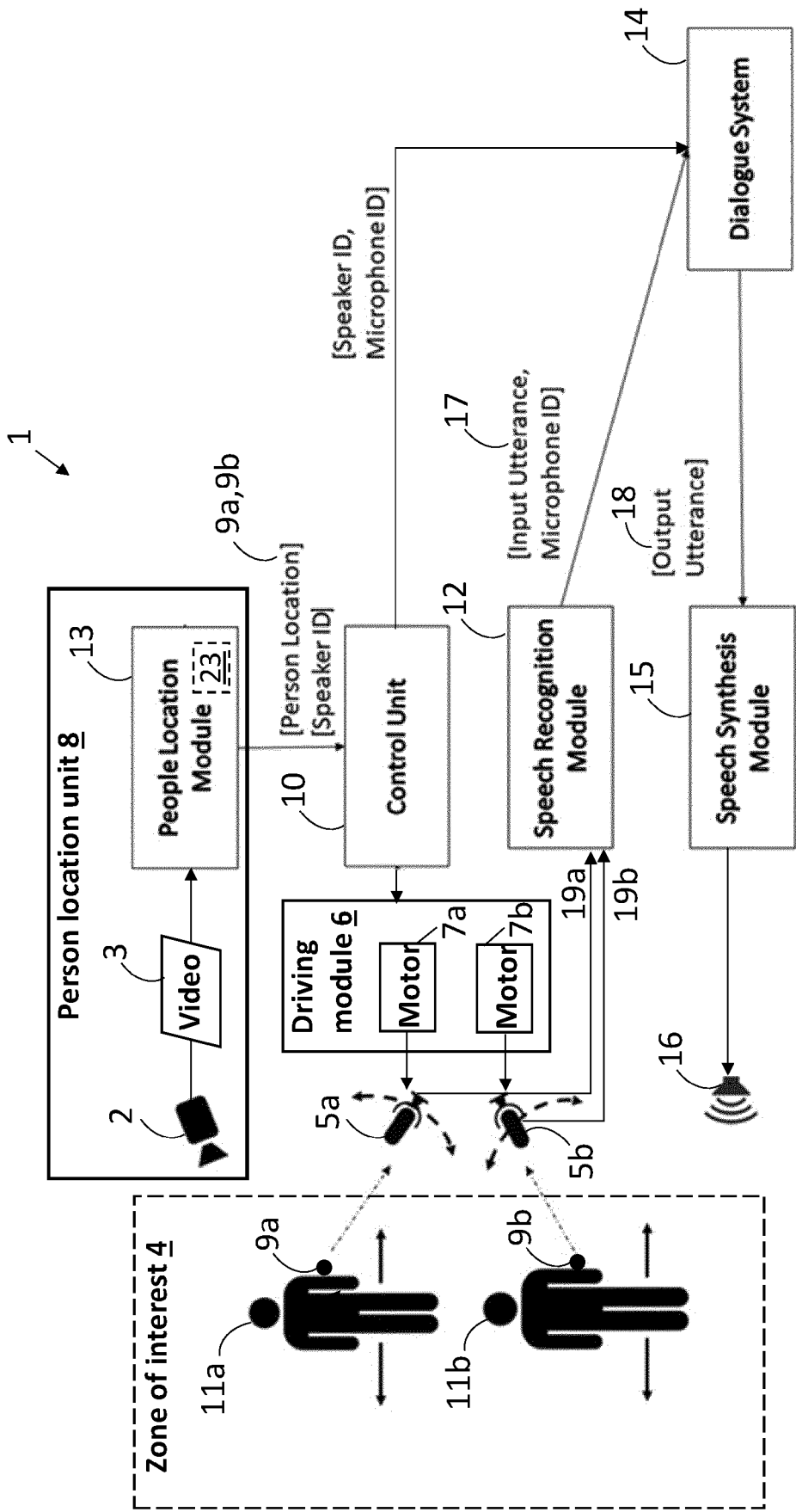


FIG. 1

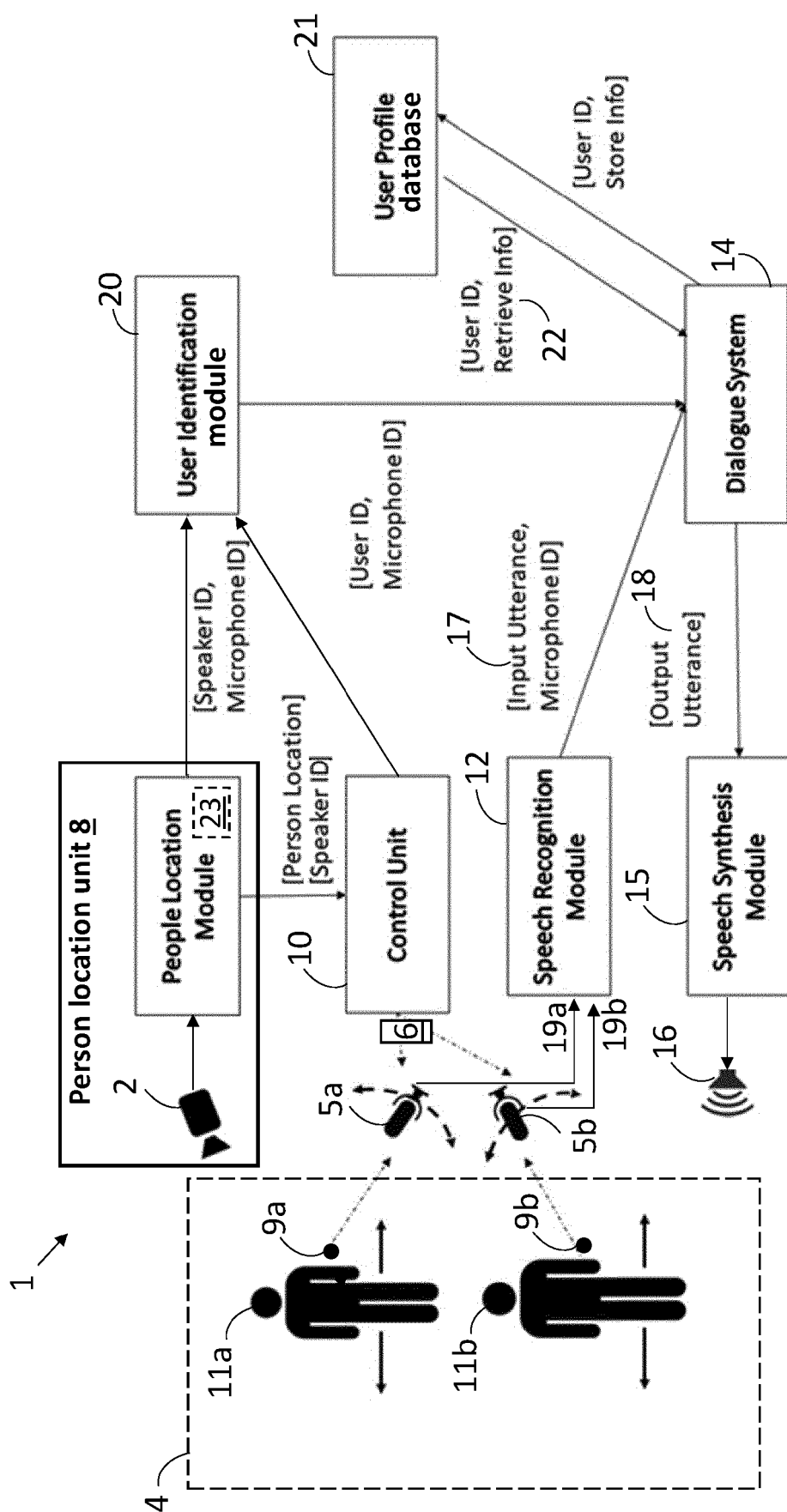
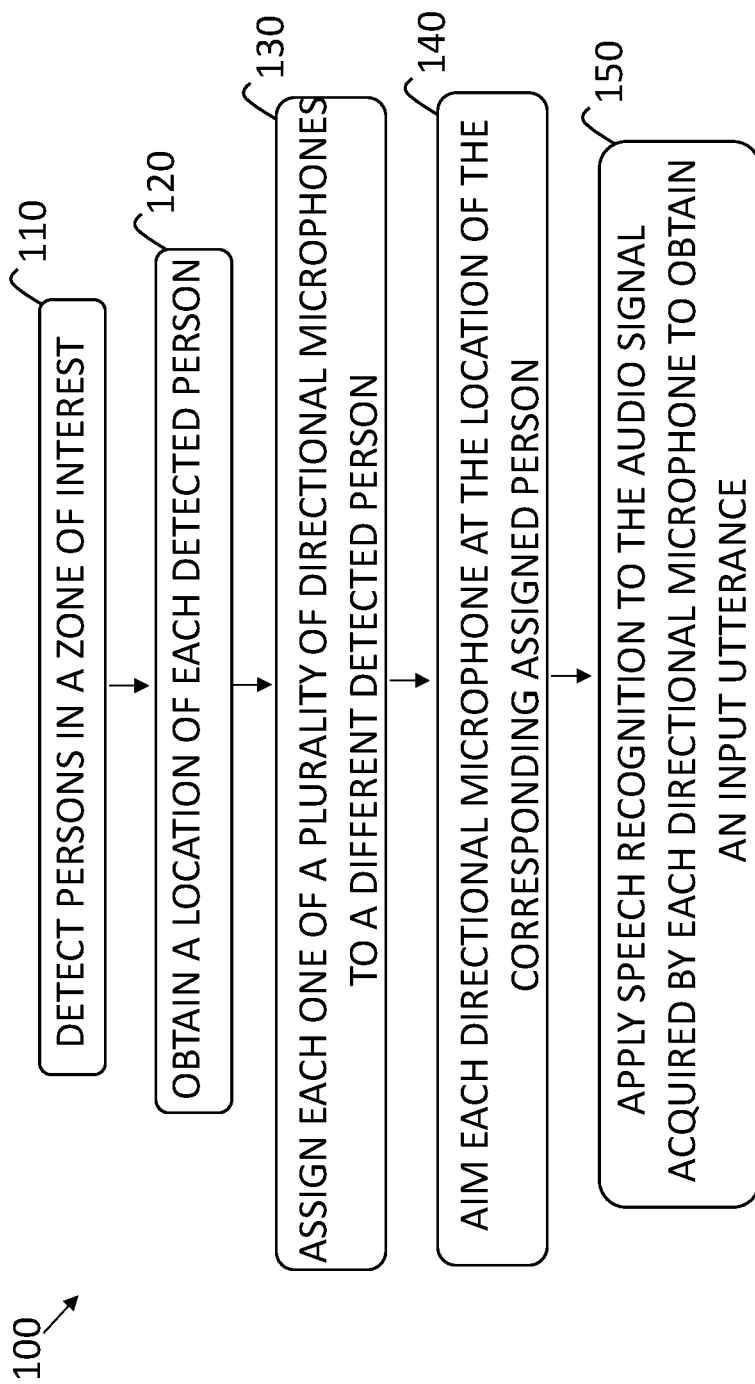


FIG. 2

FIG. 3



EUROPEAN SEARCH REPORT

Application Number

EP 24 38 2264

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	US 2023/319416 A1 (JORDAN ROBERT MICHAEL [US] ET AL) 5 October 2023 (2023-10-05) * figures 6,7 * * paragraphs [0054] - [0055] * -----	1-14	INV. G10L15/22 G10L15/28 H04R1/08
A	US 9 966 059 B1 (AYRAPETIAN ROBERT [US] ET AL) 8 May 2018 (2018-05-08) * column 12, lines 11-17 * * column 30, lines 24-32 * -----	1-14	
			TECHNICAL FIELDS SEARCHED (IPC)
			G10L H04S H04R
The present search report has been drawn up for all claims			
Place of search		Date of completion of the search	Examiner
Munich		20 June 2024	Chétry, Nicolas
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document			
T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document			

EPO FORM 1503 03.82 (P04C01)

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

EP 24 38 2264

5

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

20-06-2024

10

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2023319416 A1	05-10-2023	US 2023319416 A1	05-10-2023
		WO 2023192597 A1	05-10-2023
US 9966059 B1	08-05-2018	NONE	

15

20

25

30

35

40

45

50

55

EPO FORM P0459

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