(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 3 772 735 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
10.02.2021 Bulletin 2021/06

(51) Int Cl.:
G10L 21/02 *(2013.01)* *H04R 25/00* *(2006.01)*

(21) Application number: **19194153.3**

(22) Date of filing: **28.08.2019**

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

(30) Priority: **09.08.2019 EP 19191075**

(71) Applicant: **Honda Research Institute Europe GmbH**
63073 Offenbach/Main (DE)

(72) Inventors:
- **Heckmann, Martin**
  63073 Offenbach/Main (DE)
- **Richter, Andreas**
  63073 Offenbach/Main (DE)

(74) Representative: **Beder, Jens**
Mitscherlich PartmbB
Patent- und Rechtsanwälte
Sonnenstraße 33
80331 München (DE)

(54) ASSISTANCE SYSTEM AND METHOD FOR PROVIDING INFORMATION TO A USER USING SPEECH OUTPUT

(57) An assistance system performs a method for providing information to an assisted person using speech output. The system comprises at least one sensor for acoustically sensing an environment, in which the assistance system and the assisted person are located. The system further comprises a processor that is configured to analyze a sensor output from the at least one sensor. The processor estimates a potential interference of an intended speech output with the sensed acoustic environment in the common environment on the basis of the analysis result. Additionally, the processor obtains information on an assisted person's hearing capacity and determines an expected intelligibility of speech output on the basis of the estimated interference and the obtained information on the assisted person's hearing capacity.

The system further comprises a speech output signal generation unit. The speech presentation signal generation unit generates a speech presentation signal in accordance with the determined modality for intended speech presentation. The speech output signal is then supplied at least to a loudspeaker or hearing aid for outputting the speech output including the information to be provided to the user.

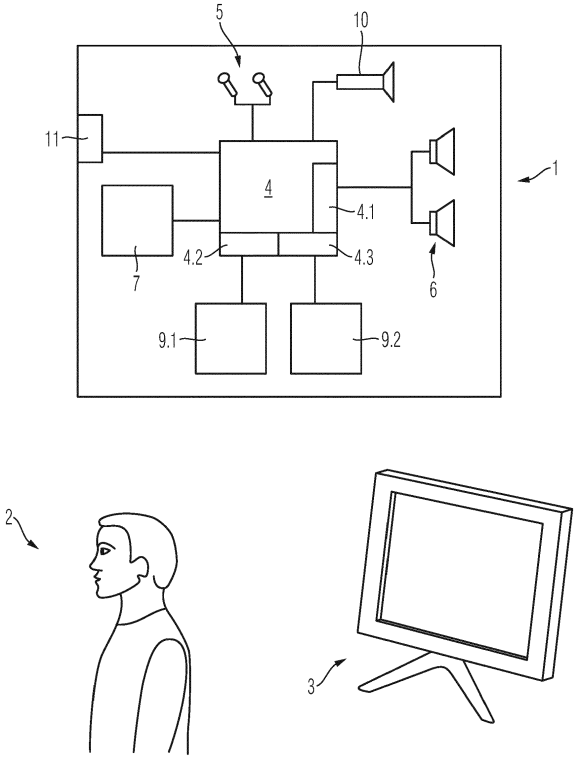

Fig. 1

EP 3 772 735 A1

Description

[0001] The present invention regards an assistance system and a corresponding method for assisting a user, wherein the system and method use speech output for providing information to a user.

[0002] Assistance systems become increasingly popular. They are developed to assist their users in many different areas of the user's daily life. For example, industrial service robots assist workers to fulfil their working tasks by handing tools, holding workpieces or providing information on how to proceed. Personal robots answer to information requests from a user, inform the user on upcoming events in his environment, inform the user about calendar entries, or remind the user to take medications. Even social tasks may be fulfilled, for example, engaging in a conversation with the user to increase his mental well-being.

[0003] The flexibility of such assistance systems is significantly influenced by its capability to interact with the user. One important channel for interaction is speech because listening and talking is a very efficient and intuitive way for communication. Thus, for future assistance systems, speech plays a key role. Unfortunately, in an aging society the number of people suffering from a hearing loss increases. Consequently, these users might be excluded or strongly impaired in their use of coming generations of intelligent assistance systems unless their needs are taken into account during the design of these systems. Usefulness of the assistance systems therefore will significantly depend on the system's capability to adapt to users with hearing impairments.

[0004] Commonly known assistance systems are not directly adapted to hearing impaired users but rather beneficially interact with hearing aid systems. Hearing aid systems that selectively amplify certain frequencies according to a hearing loss of the user are well known in the market. Recently, these hearing aid systems are provided with wireless communication technology such as Bluetooth such that the hearing aid system can wirelessly connect to telephones, television sets, computers, music players and other devices with audio output using a streaming device.

[0005] US 9, 124, 983 B2 suggests a hearing aid system, which enhances the audio signals that are transmitted from their sources to the hearing aid system such that localization of each of the one or more streaming sources is possible for the wearer of the hearing aid system. This is achieved by determining the position of the hearing aid system relative to each streaming source in real-time. However, the benefit of such a hearing aid system is still very limited because it relies on a rather simple amplification of sound.

[0006] It would thus be desirable to facilitate the perception of a speech output from an assistance system by adapting the communication between the assistance system and its user situation dependent.

[0007] This object is achieved with the inventive assistance system and assistance method according to the independent claims. Further details and aspects are defined in the dependent claims.

[0008] The inventive assistance system is capable of generating a speech presentation including speech output for providing information to a person who is assisted by the assistance system. The assistance system comprises at least one sensor for acoustically sensing an environment in which the assisted person and the assistance system are located. The sensor may comprise at least one microphone. The sensor output, indicative of the acoustic environment, is provided to a processor of the assistance system. The acoustic environment can typically be characterized by the sound sources, which are present, and the acoustic reflections from objects in the environment, e.g. walls. These sound sources and reflections manifest themselves as ambient noise and reverberations. Ambient noise and reverberations have a negative effect on the intelligibility of speech sounds. The sensor output is analyzed by the assistance system using the processor. In the analysis of the sensor output, characteristics of ambient noise and reverberations are determined. The analysis focuses especially on such characteristics of the ambient noise and reverberations that are known to significantly influence a person's auditory perception.

[0009] Based on the result of this analysis, i.e. the analysis of the characteristics of the acoustic environment, a potential interference of an intended speech output with the sensed ambient noise and reverberations in the common environment is estimated. The intended speech output is verbal information, which shall be provided to the assisted person next. By analyzing the sensed acoustic environment, the assistance system gathers information on ambient noise and reverberations in the environment of the assisted person. Additionally, the assistance system obtains information on the assisted person's hearing and in particular, hearing impairment. The methods laid out in the following can be beneficial to assisted persons with normal and impaired hearing. Also assisted persons with normal hearing will benefit from the presented methods in situations where either or both of ambient noise levels and reverberations are high. The term hearing capacity encompasses normal and impaired hearing.

[0010] This information on the assisted person's hearing capacity can be stored in a memory a priori or it can be (continuously) analyzed from an interaction between the assisted person and the assistance system. Based on the knowledge about the assisted person's hearing capacity and the estimated interference, the modality of speech presentation is determined such that an expected intelligibility of the speech output is optimized (improved). The determined modality is then used for the speech presentation. The modality defines the parameters to be used for the speech presentation including a position of a perceived origin of the speech output. Using these defined parameters, a speech presentation signal is generated. This speech presentation signal is then supplied

at least to a loudspeaker for outputting the intended speech output presentation and to other actuators of the assistance system to provide the additional multimodal information of the speech presentation to the assisted person.

[0011] The inventive assistance system targets the optimization of the intelligibility of its speech output for the assisted person taking into account the current acoustic environment the assistance system and the person are embedded in and the assisted person's hearing capacity. The assistance system achieves this by first assessing the acoustic environment, the assisted person's hearing capacity and the impact of the acoustic environment on the intelligibility given the assisted person's hearing capacity. In a subsequent step, the assistance system modifies parameters of the speech presentation such that the expected intelligibility of the output speech signal by the assisted person reaches an acceptable level. The modalities of the speech representation may, for example, be defined by parameters of the speech representation. These parameters include linguistic parameters of the speech output (e.g. shortening of sentences, use of more common words, use of words which have a higher intelligibility ...), acoustic parameters of the speech signal (e.g. sound pressure level, speech rate, prosodic variations, spectral distribution ...), recruitment of communicative gestures (e.g. pointing to objects, visual prosody ...), visual modalities (e.g. text displays, display of images) as well as the position of the perceived sound source (e.g. via virtual movements in a multi loudspeaker scenario and/or physical movements of the system or a part of it). Furthermore, the assistance system may also permanently monitor the assisted person's hearing capacity based on the assisted person's interaction with the system. Based on this monitoring, the assistance system is capable of adapting its model of the assisted person's hearing capacity accordingly.

[0012] The term speech output refers to the acoustic presentation of information to the assisted person via an acoustic output device, for example, one or more loudspeakers. However, also the case when the speech output is performed by a hearing aid worn by the assisted person is included by the term speech output. This also includes cases where the hearing aid uses other modalities than acoustic waves to transmit the speech signal to the auditory system of the assisted person, e.g. via electric nerve stimulation or bone conduction. The term speech presentation shall refer to the multi-modal presentation of the speech signal, which might include acoustics but also visually perceivable gestures, images, and/or text, etc. It is to be noted that the term "speech presentation signal" may consists of a plurality of individual signals each directed to one output device or actuator involved in outputting the speech output.

[0013] The inventive system and method have the advantage that an adaptation of the speech presentation is not limited to a pure amplification of a speech output but takes account of an individual hearing capacity and its interaction with the current environmental situation. The determination of the modality may be based on a lookup table that associates the parameters that shall be set when outputting the speech presentation with the assisted person's hearing capacity and the respective characteristics that are determined from the sensed acoustic environment. The speech presentation of the assistance system thus automatically adapts to changing environmental conditions and different hearing capacities, for example, when different assisted persons use the same assistance system and are identified by the assistance system.

[0014] It is particularly preferred, that the assistance system analyzes the sensor output by determining at least one of a frequency distribution, an intensity of sound emitted by one or more sound sources in the common environment, and a location of the sound source. For different frequency distributions, for example, different modalities for outputting the speech presentation can be determined by adapting the frequency of the speech output to shift it into a frequency range with less interference with the ambient noise. Determination of a location of the sound source allows moving the perceived speech output origin to a different position. Thus, the interference between the sound from one or more sound sources in the environment of the assisted person and the speech output will be reduced. Finally, analyzing an intensity of sound emitted by sound sources in the environment allows limiting the intensity of the speech output to such an extent that is sufficient for the user to understand easily the speech output without bothering the user.

[0015] It is particularly preferred that the determined modality defines parameters of the speech presentation including at least one of: voice, frequency, timing, combination of speech output and gestures, intensity, prosody, speech output complexity level, and position of the speech output origin as perceived by the user. Adapting the voice that is used for the speech output is one simple way of adapting to a specific hearing capacity of the assisted person. Depending on the individual hearing loss with respect to the frequency range, it is for some people easier to understand a women's voice compared to a man's voice and vice versa. Thus, having knowledge about the individual hearing capacity, the system will select a voice that can easily be understood by the assisted person. Apart from that, the frequency distribution of the speech output may also be adapted to further enhance this effect. Another aspect is timing and/or speed of the speech output. When, for example, an analysis of the ambient noise reveals that the ambient noise periodically increases and decreases, a period of time can be used at least for the speech output where a reduced intensity level of the ambient noise can be expected. This could be useful, for example, when the assisted person and the assistance system are close to a crowded street with traffic lights. Further, at least the speech output can be paused when a sudden increase of the intensity of the ambient noise is detected. In general, when interferences

with a speech output are detected during its generation, which might interfere with its intelligibility, the system might repeat this same speech output. For the repetition, it might choose an instance in time in which the interferences are smaller or it might change the speech presentation to increase intelligibility despite the interference.

[0016] Especially when the assistance system comprises a humanoid robot or only a humanoid upper body, the speech output may be combined with gestures for outputting the speech presentation. Gestures might be expressed via movements of the arms, legs, head, fingers or similar components of the assistance system. They might also be expressed via facial or similar movements e.g. implemented as lip, eye, eyebrow or ear movements. As it is known from humans, gestures emphasize spoken words, parts of it or illustrate the content of the spoken words. This can be imitated by a humanoid robot. In a case, where the humanoid robot can be seen by the assisted person, this will significantly increase intelligibility. As mentioned above already, the intensity of the speech output may also be adapted, which means that the assistance system automatically adapts to both, the hearing loss of the assisted person but also the intensity of the ambient noise.

[0017] Further, it is beneficial to adapt the speech presentation complexity level. This can be achieved by associating with words in a vocabulary used for the speech presentation a complexity level, which correlates with an intelligibility level of the word. The entire speech presentation can then be limited to use words with a low complexity level for users or situations where it can be expected that understanding of more complex words is critical. The same approach can be applied on the level of the sentence structure where sentence structures can be applied which have a lower complexity and hence provide a higher intelligibility. The limitation of complexity may also be applied to the speech output only. Another very efficient way is to move the position of the (perceived) speech output origin to a location that is assumed to make it easier for the assisted person to understand the speech output. In case that the assistance system is realized by a movable entity, like a (humanoid) robot, this can be achieved by moving the entity to the desired position before starting the speech output. Alternatively, a stationary assistance system may comprise a plurality of sound sources and this plurality of sound sources is controlled in a way to move a virtual origin of the speech output to the desired position.

[0018] The latter parameter, namely the position of the speech output origin, can be used particularly efficiently in case that the assistance system is configured to determine a position of the assisted person relative to the one or more sources of the ambient noise. In that case the position of the speech output origin as perceived by the assisted person is determined on the basis of the assisted person's relative position. For example, the speech output origin is located on the opposite side with respect to the assisted person. Thus, from this side, even without increasing intensity of the speech output, the assisted person can more easily understand the speech output.

[0019] According to another preferred embodiment, the assistance system comprises a humanoid robot that includes a head with a mouth imitation and/or at least one arm. Such humanoid robot is configured to visually assist the speech output for outputting the speech presentation using at least one of: head movement, lip movement and/or movement of the at least one arm or one or more parts thereof. The movement is coordinated with the speech output. Lip movement that is coordinated with the speech output facilitates distinction of different vowels and consonants and thereby assists the user's comprehension. Similarly, comprehension is improved when head movements like nodding or shaking the head are coordinated with the speech output, in particular with its content. In case of a humanoid robot that comprises at least one arm including a hand, the arm and/or hand as part thereof may be controlled to realize a pointing movement. Thus, the humanoid robot can point to a position or to an object to which the speech output refers. Even if the speech output was not perfectly understood by the assisted person, he can still recognize the content because of the additional information he receives by the pointing movement. Other gestures may be thought of as well, for example, when size is one aspect in the content of the speech output, a respective indication can be given using the arms of the humanoid robot. Similarly, gestures for proximity or distance can be realized easily.

[0020] Another way to visually assist the speech output, when outputting the speech presentation, uses a display. Using a display makes is possible to display animations, pictures or text. In case the assistance system is not implemented on a humanoid robot or humanoid upper body, the animations can be used to represent the movements of the missing body parts, e.g. arms, head, mouth. The text and/or one or more picture that is displayed can refer to parts of the speech output, at least. Thus, it is possible to emphasize keywords or at least ensure that these keywords are well understood by the assisted person. Since keywords can be presented either in writing or in displaying corresponding images, using a display is particularly advantageous, when the assistance system is a mobile entity like a (humanoid) robot. Moving the system to a position that allows the assisted person to look at the display avoids that the assisted person himself has to move. Further, the visual assistance and the speech output come from the same position, which makes it easier for the assisted person to understand. Additionally or alternatively, visual information can also be presented at a different location, e.g. projecting it to a wall using a projector device, or using projections on smart glasses or contact lenses.

[0021] According to a further preferred embodiment, the reactions of the assisted person on a speech presentation from the assistance system is monitored by one or more sensors of the assistance system. Such sensors

may comprise one or more microphones, which can be dedicated microphones for monitoring a spoken response from the assisted person or the same microphones as the ones used for acoustically sensing the environment. Further, the sensors may comprise one or more cameras to record movements, head pose, or the like in reaction to a speech presentation of the system. The monitored reaction of the assisted person is then compared by the processor with an expected reaction. The comparison results in a determination of deviations and the determined deviations are stored associated with the respective modality that was used for the speech presentation causing the reaction. Additionally or alternatively, the deviations are stored associated with the results of the analysis of the acoustic environment. Monitoring deviations from expected responses depending on the used modality allows to improve the determination of the best combination of modalities to present the desired information by the speech presentation. For example, the analysis allows identifying modalities, which lead to a significant improvement in the assisted person's comprehension of the speech presentation. On the other side, some parameters may work advantageously with only specific acoustic environments. Adapting the determination of the modality accordingly will thus improve comprehension of the speech presentation for the future.

[0022] Aspects and details of the invention will now be described with respect to the annexed drawings in which

figure 1 shows the general layout of the assistance system according to the invention and a situation for explaining its functionality,

figure 2 shows a top view to illustrate the adaptation of the position of the robot for the speech presentation,

figure 3 schematically illustrates a humanoid robot and explains the pointing movement,

figure 4 shows a situation comparable to the one of figure 2 but with a plurality of loudspeakers to realize a virtual speech output origin, and

figure 5 shows a flowchart illustrating the major method steps.

[0023] Figure 1 shows a block diagram of the inventive assistance system 1. The assistance system 1 is intended for assisting a person 2 in an environment comprising as a single exemplary sound source a television 3. Obviously, a plurality of different sound sources may be present in the environment. Only for simplicity of the explanation, the number of sound sources is reduced to one.

[0024] The assistance system 1 comprises a processor 4, which is connected to a sensor for acoustically sensing the environment in which the assisted person 2 and the television 3 are located. In the illustrated embodiment, the sensor comprises two microphones 5. The signals that the microphones 5 generate in response to ambient noise and reverberations is supplied to the processor 4. The processor 4 performs an analysis of the supplied signal in order to analyze the ambient noise and reverberations. The processor 4 particularly determines a frequency distribution and intensity of sound emitted by the television 3. Since in the illustrated embodiment two microphones 5 are arranged at different locations, the analysis also allows to determine the location of the sound source.

[0025] The assistance system 1 further comprises a plurality of loudspeakers 6. In the illustrated embodiment, two loudspeakers are used but it is evident, that the number of loudspeakers may be more or less. Loudspeakers 6 are driven by a speech output signal generating unit 4.1. In the block diagram of figure 1, the speech output signal generating unit 4.1 is shown as being a part of the processor 4. Obviously, the signal generated by the speech output signal generating unit 4.1 may be amplified before it is supplied to the loudspeakers 6. For simplicity of the drawing, such amplifier is not shown in the drawing.

[0026] For visually assisting the speech output, a display 9.1 and actuators 9.2 are included in the assistance system 1. The display 9.1 receives signals from a display controller 4.2 which is also illustrated as being part of the processor 4. Similarly, the processor 4 comprises a control signal generating unit 4.3 which drives the actuators 9.2. As it was explained already for the loudspeakers 6 and the respective signal generation for driving the loudspeakers 6, there may also be separate drivers for amplifying or modifying the signals so that an intended image and/or text can be displayed on display 9.1 or that the actuators 9.2 cause the desired movements. It is to be noted that only one actuator 9.2 is illustrated, but obviously, a plurality of such actuators may be used. For a humanoid robot having two arms including hands with fingers, it is evident that quite a number of actuators 9.2 must be present. Since controlling such extremities of a humanoid robot is known in the art, no specific explanations will be given thereon.

[0027] The speech output signal, the signal from the display controller 4.2 and the control signal commonly establish a speech presentation signal. Accordingly, the speech output generating unit 4.1, the display controller 4.2 and the control signal generating unit 4.3 are components of a speech presentation signal generation unit. The speech presentation generation unit may comprise less or more components but comprises at least the speech output signal generating unit 4.1.

[0028] The processor 4 is further connected to a memory 7. In the memory 7, the obtained information on a hearing capacity of the assisted person 2 may be stored. Further, all executable programs that are needed for the analysis of the acoustic environment, generation of a speech presentation, a database for storing vocabulary for the speech presentation, a table for determining a modality for the speech presentation based on the analysis result of the acoustic environment, and the like, are stored in this memory 7. The processor 4 is able to re-

trieve information from the memory 7 and store back information to the memory 7.

[0029] Finally, the assistance system 1 comprises an interface 11 that is connected to the processor 4. As it is illustrated in figure 1, the interface 11 may be unidirectional in case that it is only used for obtaining information on the assisted person's hearing capacity. The interface 11 may for example be used to read in information that is provided by the hearing aid of the assisted person. This information is then stored in the memory 7. The interface 11 may be a wireless interface. Alternatively, a wired connection to a data source, for example, received from the assisted person's audiologist may be established. A bidirectional interface 11 may also be realized. In that case, information on the assisted person's hearing capacity derived from an analysis of ongoing interaction between the assistance system 1 and the assisted person 2 may be exported for use in other systems.

[0030] The assistant system 1 may further comprise a camera 10 or even a plurality of cameras. The camera 10 on the one hand may be used in order to monitor the reactions of the assisted person 2 in response to a speech presentation from the loudspeakers 6, display 9.1, movements caused by the actuator 9.2, but also for enabling the assistance system 1 to move freely in an unknown environment. In such a case, the recorded images from the camera 10 are processed in the processor 4 and, based on such image processing, control signals for actuators 9.2 are generated that cause the assistance system 1 to move to a desired position.

[0031] Figure 2 presents one example, how the assistance system 1 makes use of its information regarding the hearing capacity and the analysis of the acoustic environment. Figure 2 shows a top view in a situation similar to the one shown in figure 1. Here, the assistance system 1 is a humanoid robot. The microphones 5 of the assistance system 1 record sound that is output by the television 3. In the processor 4, the location of the television 3 but also the position of the assisted person 2 is determined. Determining the position and in particular also the orientation of the assisted person 2, or at least the orientation of the assisted person's head, is performed by image processing of images taken by the camera 10. In the situation illustrated in figure 2, the assisted person 2 will hear the sound from the television 3 primarily with its right ear 15. Since the assistance system 1 has analyzed the relative position of the television 3 and the assisted person 2 and also the orientation of the assisted person's head, it will move its own position more towards the left ear 16 of the assisted person 2. Thus, the interference between the sound that is output by the television 3 and the speech output emitted by the loudspeakers 6 is reduced. Additionally, the assistance system 1 could move closer to the assisted person 2.

[0032] In case that the assistance system 1 is only capable of generating the speech output but cannot visually assist the speech output, the position of the assistance system 1 may even be moved more towards the left ear 16 of the person 2. In the illustrated embodiment, however, the assistance system 1 comprises the display 9.1 and thus, it has to position itself at a location such that the display 9.1 is easily visible by the person 2. The humanoid robot of the assistance system 1 has the display 9.1 attached to a head 18 of the humanoid robot that is arranged on a body 17. As indicated in the simplified top view, the humanoid robot furthermore comprises a left arm 19 and a right arm 20. Preferably, the speech presentation uses speech output that is visually assisted.

[0033] One way to use the arms of a humanoid robot is illustrated in figure 3. Figure 3 shows a front view of a humanoid robot that comprises as mentioned with respect to figure 2, a body 17, head 18, loudspeaker 6, and microphones 5, which are realized as ears attached to the head 18, a left arm 19, a right arm 20, wherein each of the arms 19, 20 includes a hand 21 and 22, respectively. The head 18 also comprises a mouth imitation 23 with two lips that can be moved individually. When outputting speech by the loudspeakers 6, the robot can therefore move the lips coordinated with the speech output. Thus, by generating coordinated movements of the lips, the speech output can be visually assisted. A further opportunity to visually assist the speech output is moving one of the arms 19, 20 or at least a part thereof, for example, the left hand 21 or the right hand 22 coordinated with the speech output.

[0034] In one simple embodiment, as shown in figure 3, the humanoid robot points into a direction towards a position of an object which is referred to by the current speech output. Alternatively, the arms 19, 20 and/or hands 21, 22 can be controlled to move resembling a person gesticulating when speaking.

[0035] The embodiment depicted in figure 3 arranges the display 9.1 at a front side of the body 17 of the humanoid robot. This alternative arrangement of the display 9.1 enables to design the head 18 with particular focus on communicating facial gestures to the assisted person 2.

[0036] Finally, the front view of the robot shows that there are two legs 24, which are used for freely positioning the humanoid robot in the environment of the assisted person 2. The legs 24 are only one example. Any actuator that allows positioning the assistance system 1 freely in the environment of the person 2 may be used instead.

[0037] Figure 4 shows a top view with a television 3 and the assisted person 2 but here, the assistance system 1 is not realized as a humanoid robot. Rather, a plurality of speakers 6.1... 6.4 are arranged at corners of a room, for example. These four speakers 6.1... 6.4 allow to virtually generate an origin of a speech output. This means, that the four speakers 6.1... 6.4 are jointly controlled by respective speech output signals such that the person 2 gets the impression as if the speech output came from a specific location within the room.

[0038] A simplified flowchart showing the major method steps for performing the inventive assistance method is shown in figure 5. At first, in step S1, information on

the assisted person's hearing capacity is obtained. The information may come from the assisted person's audiologist, who conducted a hearing test with the assisted person 2. Alternatively, information may be read out from a hearing aid of the assisted person 2. This may be done directly, using the wireless interface 11 of the assistance system 1.

[0039] In step S2, the assistance system 1 senses the acoustic environment. Based on the sensor signal, the acoustic environment is analyzed in step S3. In the analysis, properties of the interfering sound sources and acoustic reflections are determined. These properties may include the location of the sound source, the frequency content of the sound and its intensity. Although in most cases in the present description of the invention, only one sound source is mentioned for illustrating the assistance system's function and method, the same analysis may be performed in case that there is a plurality of sound sources and sources of acoustic reflections.

[0040] Advantageously, the assistance system 1 also determines if the assisted person 2 is listening to the one or more sound sources or if they are merely background noise. In case that it is determined, that one of the sound sources is a TV 3, for example, it is very likely, that the person 2 listens to a TV program. The conclusion whether the person 2 listens to the TV program may be made based on an analysis of images taken from the person 2 by the camera 10. Such images together with the determined position of the person 2 allows determining a gaze direction and a head pose from which the focus of attention of the person 2 may be derived. The assistance system 1 may then address the person 2 and give her time to shift his focus of attention towards the assistance system 1. Doing so the person 2 might also change her position in the room or at least turn his head.

[0041] In the next step, S4, the assistance system 1 estimates the intelligibility of its speech presentation for the assisted person 2. This estimation of the intelligibility is, on one hand, based on an expected frequency dependent signal-to-noise ratio at the assisted person's location inferred from the location of the assisted person, the properties of the sound sources and acoustic reflections that are determined in the analysis of step S3 and the model of the assisted person's hearing capacity in step S1. In the estimation of the expected intelligibility, the system does not only consider the acoustic part of the speech presentation, i.e. the speech output, but also the other modalities. This means the system also considers the potential improvements of the intelligibility due to e.g. additional visual signals.

[0042] Based on the estimated intelligibility, the assistance system 1 then determines, in step S5, the modality that shall be applied to the intended speech presentation including the speech output. The modality comprises a set of parameters that is used for the speech presentation but also the position where the speech output originates. The position that is selected as an origin of the speech output is optimized taking into account the gained intelligibility by the assisted person 2 resulting from an improved signal-to-noise ratio at this position. Further, the costs in terms of time needed to reach the position and energy consumed to reach the position are also taken into consideration. When calculating a trajectory for moving a mobile assistance system 1 from its current position to the selected speech output origin, a possible intrusion in the assisted person's personal space must also be taken into consideration.

[0043] Once the modality for the speech presentation is determined, the direct parameters defined by the determined modality can be applied to an intended speech presentation. Before the modality can be applied in step S8 at first, information to be provided to the user is generated in step S6. The generated information is then converted into an intended speech outputs in step S7. The parameters defined in the determined modality are then applied on this intended speech output to generate the speech presentation.

[0044] Based thereon, the processor 4, to be more precise, the speech output signal generating unit 4.1, display control 4.2 and control signal generating unit 4.3, generate the respective control signals for driving the loudspeakers 6, actuators 9.2, and display 9.1. Thus, in step S9, the speech output is executed by the loudspeaker 6, maybe assisted by a visual output in step S9.1 and controlling actuators in step S9.2.

[0045] The reaction of the person 2 is monitored in step S10 by the camera 10 and microphone 5. In Step S11, a deviation from an expected reaction of a person 2 is determined and from such deviation, a hearing capacity model is generated or updated in step S12. This updated hearing capacity model is then stored in step S13 in the memory 7 and is available for future application.

[0046] Apart from a deviation of the assisted person's reaction from an expected reaction, it is also possible that the assisted person 2 explicitly gives feedback when he did not understand the assistance system 1. Such a direct feedback could either be a sentence like "I could not understand you" or "please repeat". Additionally, from images recorded by the camera 10, the assistance system 1 may interpret facial expressions and other expressive gestures allowing to conclude that the assisted person 2 has difficulties understanding the assistance system 1.

[0047] From these reactions on speech presentation, the assisted person's hearing capacity is inferred. The assistance system 1 determines the signal-to-noise ratios of the signals of the speech output at the assisted person's location. Further, the assistance system 1 determines how reliably the assisted person 2 understood the messages dependent on the signal-to-noise ratio. The hearing capacity of the assisted person 2 is then inferred from this data and potentially additionally using models of human hearing.

[0048] Such information on hearing capacity of the assisted person 2 may be used to update the information that was initially obtained.

[0049] Detailed descriptions of a few possible embodiments of the invention are provided in the following sections. In the first embodiment the hearing capacity of the assisted person 2 is known, yet the assisted person 2 is not wearing a hearing aid. Information on the assisted person's hearing capacity might be represented in the form of an audiogram. Such audiograms are typically prepared when an assisted person with a hearing impairment sees an audiologist. This audiogram contains a specification of the assisted person's hearing capacity for each measured frequency bin. However, the information on the assisted person's hearing capacity does not have to be limited to an audiogram but might also contain the results of other assessments (e.g. hearing in noise test, modified rhyme test ...). The audiogram can be provided to the assistance system 1 in step S1 in multiple ways, e.g. attaching a removable storage device containing it, transferring it to a device which is connected to the assistance system 1 through a special service application, e.g. running on the smartphone of the assisted person 2, or also the audiologist directly sending it to the assistance system 1 or a service application of the assistance system 1. When the assistance system 1 wishes to interact with the assisted person 2 it will first sense the acoustic environment in step S2. Of course, this sensing can also be performed continuously. This sensing includes localization of sound sources either in 2D or in 3D.

[0050] Many methods for localization of sound sources are known, employing for example, different numbers and spatial arrangements of microphones (Mavridis, N. (2015). A review of verbal and non-verbal human-robot interactive communication. Robotics and Autonomous Systems, 63, 22-35.; Valin, J. M., Michaud, F., Rouat, J., & Létourneau, D. (2003, October). Robust sound source localization using a microphone array on a mobile robot. In Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453) (Vol. 2, pp. 1228-1233). IEEE; Rodemann, T., Heckmann, M., Joublin, F., Goerick, C., & Scholling, B. (2006, October). Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 860-865). IEEE; Nakashima, H., & Mukai, T. (2005, October). 3D sound source localization system based on learning of binaural hearing. In 2005 IEEE International Conference on Systems, Man and Cybernetics (Vol. 4, pp. 3534-3539). IEEE.).

[0051] Either directly or based on their determined location these sound sources can be identified and their spectral properties estimated (Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A., Gannot, S., Vincent, E., ... & Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(4), 692-730.). In case the spectral characteristics of the noise sources are stationary a prediction from the current situation to the future situation, when the assistance system 1 will produce the speech sound, can be obtained with high accuracy. In case of time variant sources, estimates of their future changes have to be made based on external information or past observations. The system also estimates the reverberations of the current acoustic environment (Gaubitch, Nikolay D., et al. (2012) "Performance comparison of algorithms for blind reverberation time estimation from speech.", Proc. 13th International Workshop on Acoustic Echo and Noise control; Lollmann, Heinrich W., et al. (2010) "An improved algorithm for blind reverberation time estimation.", Proc. 12th International Workshop on Acoustic Echo and Noise control). Additionally, the location of the assisted person 2 relative to these sound sources and the assistance system 1 has to be determined. In case the person 2 is speaking, similar methods as described above can be used. Additionally or alternatively, visual information can be used to localize the person 2 (Zhang, C., & Zhang, Z. (2010). A survey of recent advances in face detection; Darrell, T., Gordon, G., Harville, M., & Woodfill, J. (2000). Integrated person tracking using stereo, color, and pattern detection. International Journal of Computer Vision, 37(2), 175-185.; Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., & Schiele, B. (2017). Arttrack: Articulated multi-person tracking in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6457-6465); Ramanan, D., & Zhu, X. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2879-2886)). This information allows to estimate the expected signal to noise ratio for each frequency bin of a speech sound produced by the assistance system at the user's location. This information on the influence of the ambient noise and reverberations at the assisted person's location can then be combined with the audiogram of the assisted person 2 and processed by an algorithm implemented in the assistance system 1 to estimate the intelligibility (Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. The Journal of the Acoustical Society of America, 130(3), 1475-1487;, Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7), 2125-2136.; Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica, 86(1), 117-128; Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. The Journal of the Acoustical Society of America, 125(5), 3328-3345. Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks.

Computer Speech & Language, 48, 51-66).

[0052] Hence, the assistance system 1 is capable of predicting the intelligibility of a speech output it will produce for the person 2. This will allow the assistance system 1 to perform internal simulations on how the intelligibility will change when parameters of the sound production are changed. This includes changes of the voice (male, female, voice quality ...), sound level and spectral characteristics (e.g. Lombard speech). Additionally, variations in the words and sentence structure and their influence on the intelligibility can be evaluated. Furthermore, also changes in the intelligibility due to changes of the assistance system's relative position (physical or virtual) to the person 2 and the sound sources can be determined. In addition to this, the system can also evaluate changes of the estimated intelligibility due to additional multimodal information conveyed by the system in the speech presentation. For example, it can take the influence of lip, facial and head movements on the intelligibility into account (Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212-215; Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. Psychological Science, 15(2), 133-137). In a similar direction, the system can assume that objects to which it will point or words, which it will show in its display, will be understood by the assisted person despite the ambient noise. With the knowledge on the expected intelligibility depending on the speech presentation parameters a fitness function with the speech presentation parameters as input variables and the expected intelligibility as target value can be formulated and the intelligibility can be optimized. Many algorithms to perform such an optimization of a fitness function are known. This optimization is continued until the predicted intelligibility reaches the minimum intelligibility level previously determined. This minimum intelligibility level can vary with the importance of the information to be conveyed to the assisted person 2 and the prior knowledge of the assisted person on the information. In case the information is of high importance, e.g. reminding the assisted person 2 to take a certain medication, the necessary intelligibility level can be set very high. In case the information is only a confirmation of a previous command of the assisted person 2, e.g. a confirmation that the assistance system 1 will turn off the light after the assisted person 2 requested it to do so, the intelligibility level can be lower. It has to be noted that the necessary intelligibility might also not be equal for all words in the utterance, e.g. when reminding the assisted person to take his medication the name of the medication has to obtain the highest intelligibility. In case the assistance system 1 cannot determine a solution with a sufficient intelligibility level it might select the solution with the highest level or inform the assisted person 2 that it cannot produce an intelligible speech presentation. Once it determined a solution, the assistance system 1 will control the relevant output devices, in particular loudspeakers 6, display 9.1 and actuators 9.2 in such a way that the speech presentation is produced accordingly. Following the example of informing the assisted person 2 to take his medication, the assistance system 1 might find a solution in which it visually displays the packaging of the medication and its name together with acoustically producing the relevant speech output. Alternatively, the assistance system 1 might decide to move closer to the assisted person 2 until the signal to noise ratio has sufficiently increased such that the predicted intelligibility is sufficient. Social factors, e.g. acceptable interpersonal distance, and time and energy effort to move the assistance system 1 also influence this optimization. In particular if images and text are used the assisted person's visual acuity might also be a relevant factor. Furthermore, the assisted person's cognitive abilities might also influence the optimization. When available, the assistance system 1 will take this additional information into account in the optimization process.

[0053] A further possible embodiment of the invention might be similar to the one described above with the main difference that the assisted person 2 is wearing a hearing aid. In this case the audiogram of the assisted person 2 can be transmitted from the hearing aid or its supporting device, e.g. a smartphone with a corresponding hearing aid application, to the assistance system 1. While optimizing the intelligibility of the speech presentation the assistance system 1 will have to consider the assisted person's hearing capacity after the enhancement of the audio signal by the hearing aid. This might also include a feedback from the hearing aid to the assistance system 1 with respect to its current operation conditions. The assistance system 1 can then either acoustically produce the speech output or send it electronically to the hearing aid. When sending the speech output in an electronic signal to the hearing aid the assistance system 1 might support information on the relative positions of the assisted person 2 and the assistance system 1 such that the hearing aid can use this information to recreate realistic localization cues for the assisted person 2. Alternatively, the assistance system 1 might itself process the electronic signal accordingly.

[0054] A further possible embodiment of the invention might adapt its knowledge of the hearing capacity of the assisted person 2 during interaction with the assisted person 2. The assistance system 1 is able to make predictions of the intelligibility of the speech presentation. If the assistance system 1 receives information that the intelligibility was not as expected the assistance system 1 is able to adapt its model of the intelligibility. Deviations between the predicted and the actual intelligibility can be due to different reasons. Frequently, the characteristics of the noise sources or the location of the assisted person 2 might change from the time of the prediction to the time when the speech signal was received by the assisted person 2. In most cases, the assistance system 1 will be

able to quantify these changes ex post as it is possible to continuously monitor the properties of the noise sources and the location of the assisted person 2 also while producing the speech presentation. Hence, the assistance system 1 can perform an assessment of the actual intelligibility at the time of the production of the speech presentation. This will allow the assistance system 1 to infer if a misunderstanding of the assisted person was due to an improper assessment of the assisted person's hearing capacity once other influencing factors are ruled out or minimized. This will then in turn allow the assistance system 1 to adapt its model of the assisted person's hearing capacity until the predicted intelligibility is equal or lower than the actual intelligibility by the assisted person 2. The feedback from the assisted person 2 if he understood the speech presentation can be obtained in different ways. One obvious way is that the assisted person 2 gives direct verbal or gestural feedback that he did not understand the speech presentation. An additional or alternative way is to observe the assisted person's behavior and determine if the observed behavior is in accordance with the information provided in the speech presentation, e.g. if the assisted person 2 requested for the location of an object and then directs himself in a direction other than the one indicated by the assistance system 1 it can be inferred that he did not understand the speech presentation. Also the assisted person's facial gestures can be used to determine if the person has understood the speech presentation (Lang, C., Wachsmuth, S., Wersing, H., & Hanheide, M. (2010, June). Facial expressions as feedback cue in human-robot interaction-a comparison between human and automatic recognition performances. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (pp. 79-85). IEEE.). This process of adapting the model of the persons's hearing capacity is also possible if no prior information on the persons's hearing capacity is available.

Claims

1. Assistance system using speech output for providing information to an assisted person (2), the assistance system (1) comprising at least one sensor (5) for acoustically sensing an environment, in which the assistance system (1) and the assisted person are located, a processor (4) configured to analyze a sensor output from the at least one sensor (5) and to estimate a potential interference of an intended speech output with the sensed acoustic environment in the common environment of the assisted person (2) and the assistance system (1) on the basis of the analysis result, the processor (4) being further configured to obtain information on an assisted person 's hearing capacity and to optimize an expected intelligibility of the speech output by determining a modality of speech presentation on the basis of the estimated interference and the obtained information on the assisted person's hearing capacity, wherein the system (1) further comprises a speech presentation signal generation unit (4.1, 4.2, 4.3) configured to generate a speech presentation signal in accordance with the determined modality of speech representation and to supply the speech presentation signal at least to a loudspeaker (6) or hearing aid for outputting the intended speech output including the information to be provided to the assisted person (2).

2. Assistance system according to claim 1, wherein the processor (4) is configured to analyze the sensor output by determining at least one of a frequency distribution, an intensity of ambient noise emitted by at least one sound source (3) in the common environment of the assisted person (2) and the assistance system (1), a location of the sound source (3), and a reverberation time of the common environment.

3. Assistance system according to claim 1 or 2, wherein the system repeats a speech presentation in case it determined that a previous presentation did not obtain a sufficient intelligibility.

4. Assistance system according to any one of the preceding claims, wherein the optimization of the expected intelligibility modifies parameters of the speech representation including at least one of: voice, frequency, timing, combination of speech output and gestures, intensity, prosody, speech output complexity level, and position of a speech output origin as perceived by the assisted person (2).

5. Assistance system according to any one of the preceding claims, wherein the assistance system (1) is configured to determine a position of the assisted person (2) relative to the one or more sound sources (3) of the acoustic environment and to move the position of a speech output origin as perceived by the assisted person (2) on the basis of the assisted person's relative position either based on a physical movements of an output device and/or based on a virtual modification of the perceived location of the output device.

6. Assistance system according to any one of the preceding claims, wherein the assistance system (1) comprises a robot including a head (18) with a mouth imitation (23) and/or at least one arm (19, 20), the robot being configured to visually assist the speech output using at least one of head movement, lip movement and/or movement of at least one arm (19, 20) or one or more parts (21, 22) thereof, the movement being coordinated with the speech output.

7. Assistance system according to claim 6, wherein the

speech presentation is a visually assisted speech output including a pointing movement of the at least one arm (19, 20) or the one or more parts (21, 22) thereof to point at a position and/or object referred to in the speech output.

8. Assistance system according to any one of the preceding claims, wherein the speech presentation is a visually assisted speech output and the assistance system (1) comprises a display (9.1) configured to visually assist the speech output by displaying at least parts of speech output or its content as text and/or one or more pictures and/or the assistance system (1) is configured to present visual information at a different location, in particular projecting it to a wall or using projections on smart glasses or contact lenses.

9. Assistance system according to any one of the preceding claims, wherein the assistance system (1) comprises one or more sensors (10, 5) for sensing reactions of the assisted person (2) on a speech presentation, wherein the processor (4) is configured to determine a deviation of the assisted person's reaction from an expected reaction and to store a determined deviation associated with the respective modality used for the speech presentation and/or associated with the result of the analysis of the acoustic environment.

10. Assistance system according to claim 9, wherein the processor (4) is configured to generate a hearing capacity model of the assisted person based on the stored deviation and its associated modality and/or result of the analysis of the acoustic environment.

11. Method for assisting a person by providing information to the assisted person (2), the method comprising the following steps:

- acoustically sensing (S2) with at least one sensor (5) an environment in which the assistance system (1) and the assisted person (2) are located,
- analyzing the sensor output (S3) for estimating an interference of an intended speech output with the sensed acoustic environment in the common environment of the assisted person (2) and the assistance system (1),
- obtaining information on an assisted person's hearing capacity (S1),
- optimizing an expected intelligibility of the speech output by determining a modality (S5) for a speech presentation on the basis of the estimated interference and the obtained information on the assisted person's hearing capacity,
- generating a speech presentation signal (S8) in accordance with the determined modality of speech presentation, and
- outputting the intended speech (S9) including the information to be provided to the assisted person (2) at least by a loudspeaker (6) or hearing aid on the basis of the generated speech presentation signal.

12. Method according to claim 11, wherein in the analysis step (S3) at least one of a frequency distribution, an intensity of sound emitted by at least one sound source (3) in the environment of the assisted person (2) and the assistance system (1), a location(s) of the one or more sound sources, and a reverberation time of the common environmenta re determined.

13. Method according to claim 11 or 12, wherein the optimization of the expected intelligibility modifies parameters of the speech presentation including at least one of: voice, frequency, timing, combination of speech output and gestures, intensity, prosody, speech output complexity level, and position of a speech output origin as perceived by the assisted person (2).

14. Method according to any one of claims 11 to 13, wherein a position of the person (2) relative to the one or more sources (3) of the acoustic environment is determined and the position of a speech output origin as perceived by the person (2) is moved on the basis of the assisted person's relative position.

15. Method according to any of claims 11 to 14, wherein for outputting the speech presentation, speech output is visually assisted by a robot by at least one of the robot's head movement, moving lips of the robot's mouth and moving of at least one arm (19, 20) or one or more parts (21, 22) thereof coordinated with the speech output.

16. Method according to claim 15, wherein for outputting the speech presentation the robot visually assists the speech output by pointing at a position and/or object referred to in the speech output.

17. Method according to any of claims 11 to 16, wherein the assistance system (1) for outputting the speech presentation visually assists the speech output by displaying at least parts of the speech output or its content as text and/or one or more pictures.

18. Method according to any one of claims 11 to 17, wherein the assistance system (1) senses (S10) reactions of the assisted person (2) on a speech presentation and determines a deviation (S11) of the assisted person's reaction from an expected reaction and stores this deviation associated with the modality used for the underlying speech presentation

and/or associated with results of the analysis of the acoustic environment.

19. Method according to claim 18, wherein the assistance system's processor (4) generates a hearing capacity model (S12) of the assisted person on the basis of the stored deviation and its associated modality and/or result of the analysis of the acoustic environment.

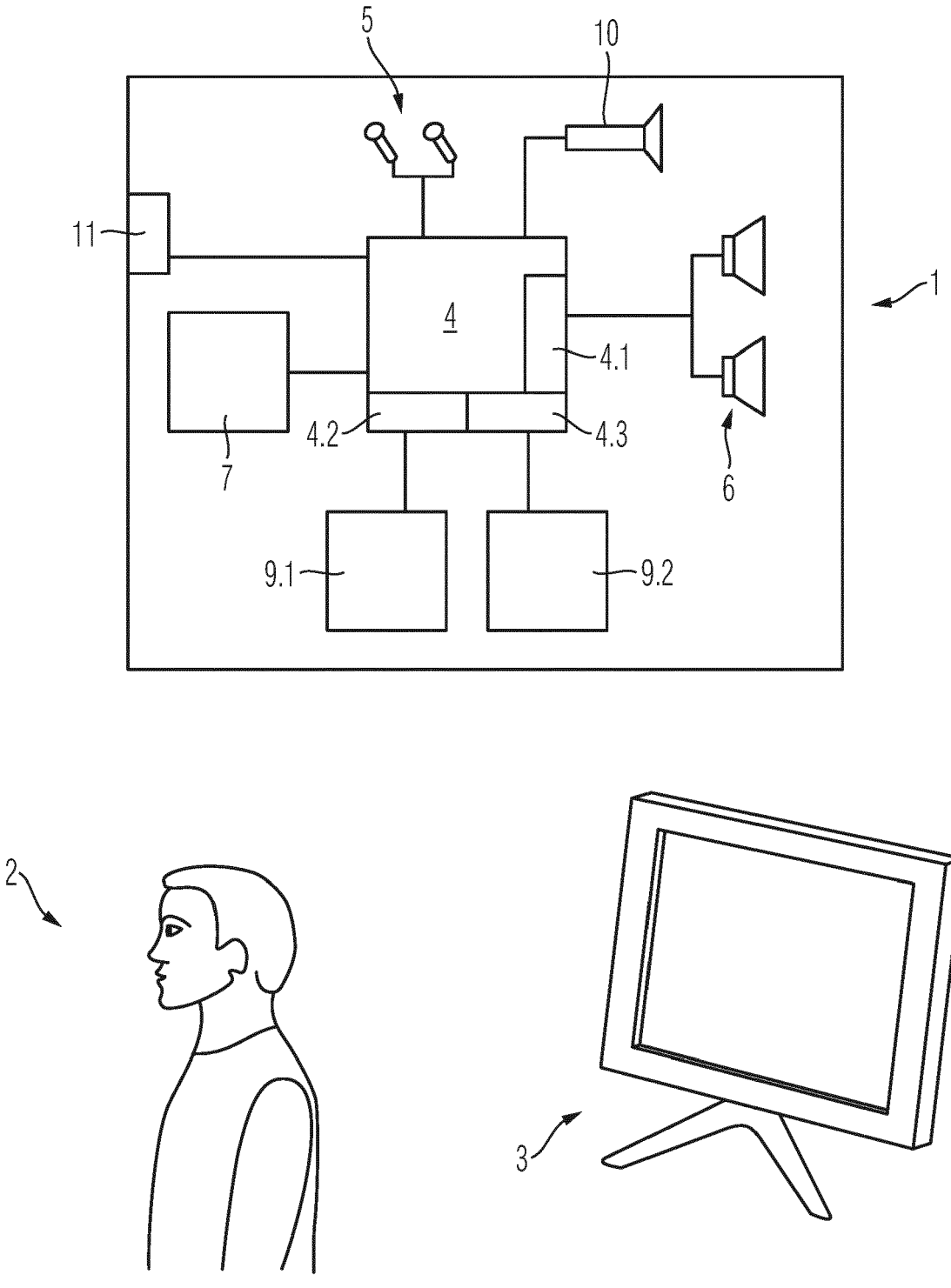

Fig. 1

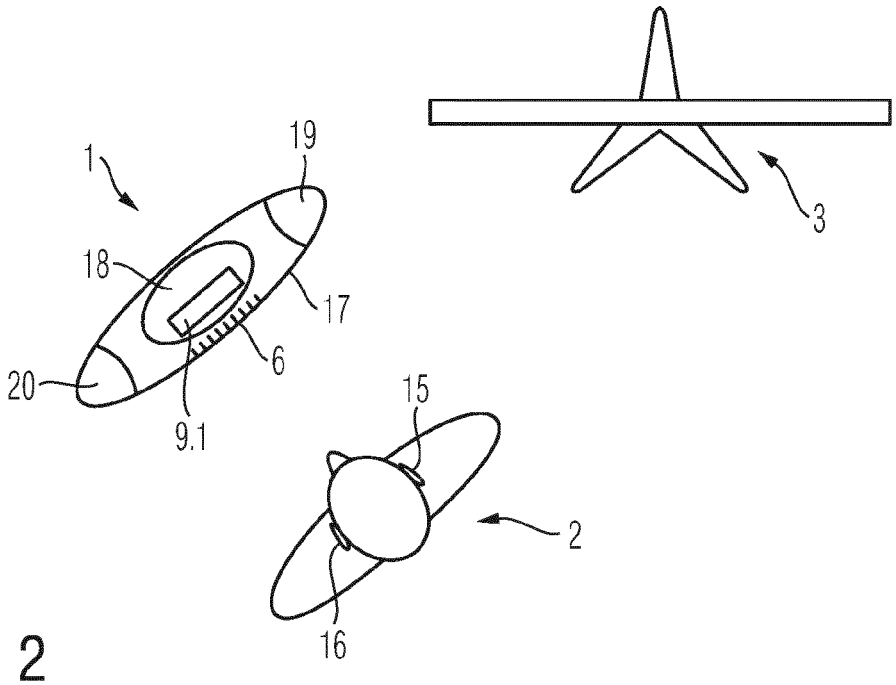

Fig. 2

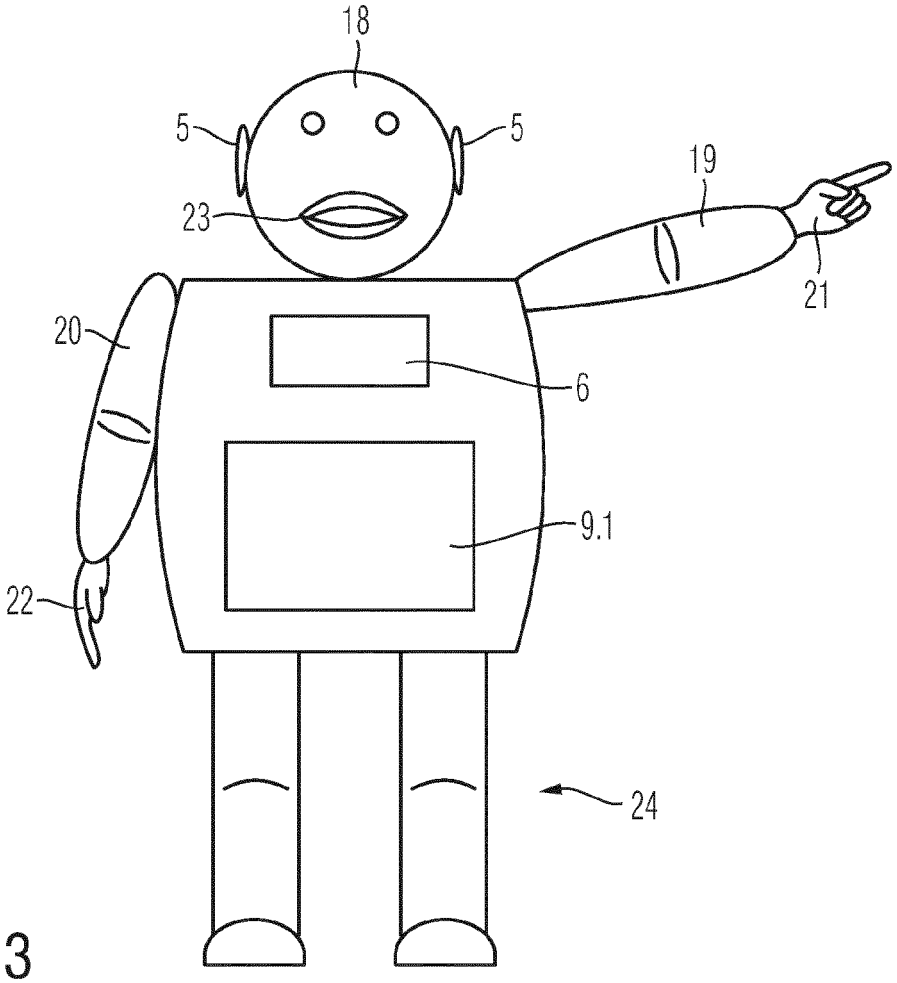

Fig. 3

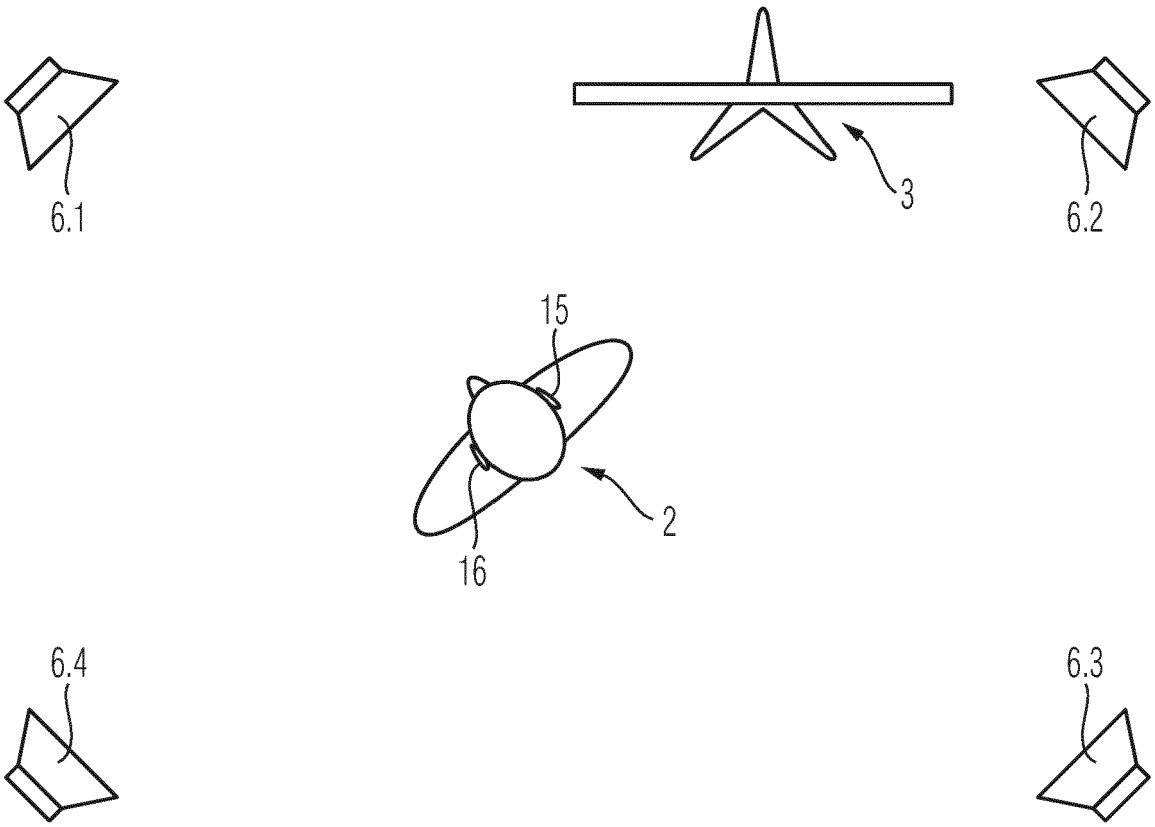

Fig. 4

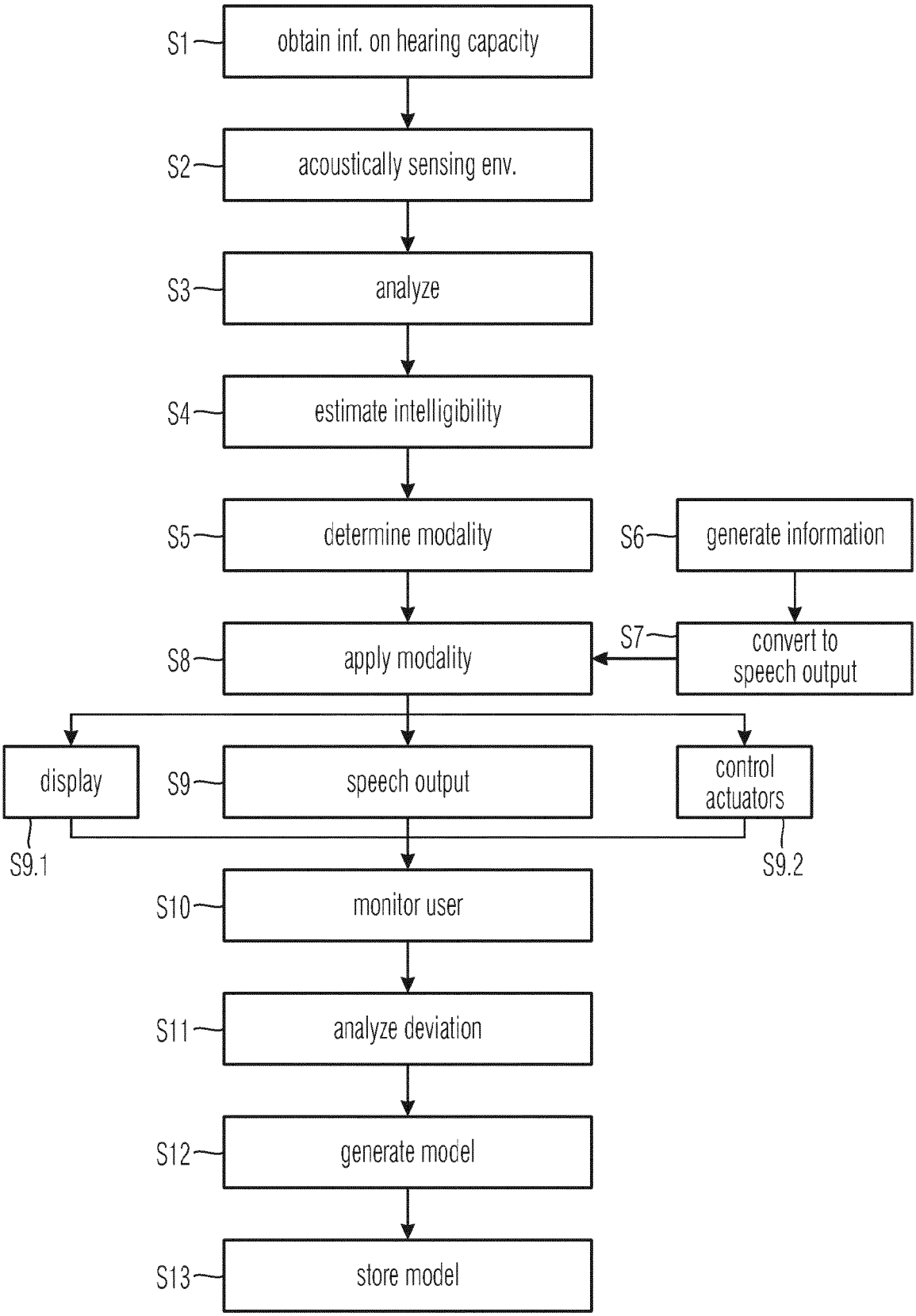
S1
obtain inf. on hearing capacity
S2
acoustically sensing env.
S3
analyze
S4
estimate intelligibility
S5
determine modality
S6
generate information
S7
convert to speech output
S8
apply modality
display
S9
speech output
control actuators
S9.1
S9.2
S10
monitor user
S11
analyze deviation
S12
generate model
S13
store model

Fig. 5

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 19 19 4153

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 7 110 951 B1 (LEMELSON LEGAL REPRESENTATIVE [US] ET AL) 19 September 2006 (2006-09-19) * column 1, line 7 - column 1, line 12 * * column 6, line 63 - column 7, line 12 * * column 7, line 26 - column 7, line 28 * * column 10, line 21 - column 10, line 35 * * column 10, line 36 - column 11, line 9 * * column 15, line 55 - column 16, line 9 * * figure 2 *	1-19	INV. G10L21/02 ADD. H04R25/00
A	US 2015/003653 A1 (RECKER KARRIE LARAE [US] ET AL) 1 January 2015 (2015-01-01) * paragraphs [0006], [0016] *	5,14	
A	US 2019/174237 A1 (LUNNER THOMAS [DK] ET AL) 6 June 2019 (2019-06-06) * paragraph [0119] - paragraph [0121] *	5,14	
A	KOAY K L ET AL: "Hey! There is someone at your door. A hearing robot using visual communication signals of hearing dogs to communicate intent", 2013 IEEE SYMPOSIUM ON ARTIFICIAL LIFE (ALIFE), IEEE, 16 April 2013 (2013-04-16), pages 90-97, XP032483664, ISSN: 2160-6374, DOI: 10.1109/ALIFE.2013.6602436 [retrieved on 2013-09-17] * abstract *	6,7,15, 16	TECHNICAL FIELDS SEARCHED (IPC) G10L H04S H04R
A	US 2009/076816 A1 (BRADFORD KIPP [US] ET AL) 19 March 2009 (2009-03-19) * figures 9B,10B,11 *	8,17	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	8 May 2020	Taddei, Hervé

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 19 19 4153

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.

08-05-2020

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7110951 B1	19-09-2006	US 7110951 B1	19-09-2006
		US 2005086058 A1	21-04-2005
US 2015003653 A1	01-01-2015	EP 2819437 A1	31-12-2014
		US 2015003653 A1	01-01-2015
		US 2016066103 A1	03-03-2016
		US 2017171672 A1	15-06-2017
US 2019174237 A1	06-06-2019	CN 109922417 A	21-06-2019
		EP 3496417 A2	12-06-2019
		US 2019174237 A1	06-06-2019
US 2009076816 A1	19-03-2009	NONE	

EPO FORM P0459

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

REFERENCES CITED IN THE DESCRIPTION

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

Patent documents cited in the description

- US 9124983 B2 **[0005]**

Non-patent literature cited in the description

- **MAVRIDIS, N.** A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems,* 2015, vol. 63, 22-35 **[0050]**
- Robust sound source localization using a microphone array on a mobile robot. **VALIN, J. M. ; MICHAUD, F. ; ROUAT, J. ; LÉTOURNEAU, D.** Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453). IEEE, October 2003, vol. 2, 1228-1233 **[0050]**
- Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. **RODEMANN, T. ; HECKMANN, M. ; JOUBLIN, F. ; GOERICK, C. ; SCHOLLING, B.** 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, October 2006, 860-865 **[0050]**
- 3D sound source localization system based on learning of binaural hearing. **NAKASHIMA, H. ; MUKAI, T.** 2005 IEEE International Conference on Systems, Man and Cybernetics. IEEE, October 2005, vol. 4, 3534-3539 **[0050]**
- **GANNOT, S. ; VINCENT, E. ; MARKOVICH-GOLAN, S. ; OZEROV, A. ; GANNOT, S. ; VINCENT, E. ; OZEROV, A.** A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* 2017, vol. 25 (4), 692-730 **[0051]**
- **GAUBITCH, NIKOLAY D. et al.** Performance comparison of algorithms for blind reverberation time estimation from speech. *Proc. 13th International Workshop on Acoustic Echo and Noise control,* 2012 **[0051]**
- **LOLLMANN, HEINRICH W. et al.** An improved algorithm for blind reverberation time estimation. *Proc. 12th International Workshop on Acoustic Echo and Noise control,* 2010 **[0051]**
- **ZHANG, C. ; ZHANG, Z.** *A survey of recent advances in face detection,* 2010 **[0051]**
- **DARRELL, T. ; GORDON, G. ; HARVILLE, M. ; WOODFILL, J.** Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision,* 2000, vol. 37 (2), 175-185 **[0051]**
- **INSAFUTDINOV, E. ; ANDRILUKA, M. ; PISHCHULIN, L. ; TANG, S. ; LEVINKOV, E. ; ANDRES, B. ; SCHIELE, B.** Arttrack: Articulated multi-person tracking in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, 6457-6465 **[0051]**
- **RAMANAN, D. ; ZHU, X.** Face detection, pose estimation, and landmark localization in the wild. *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* June 2012, 2879-2886 **[0051]**
- **JØRGENSEN, S. ; DAU, T.** Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America,* 2011, vol. 130 (3), 1475-1487 **[0051]**
- **TAAL, C. H. ; HENDRIKS, R. C. ; HEUSDENS, R. ; JENSEN, J.** An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing,* 2011, vol. 19 (7), 2125-2136 **[0051]**
- **BRONKHORST, A. W.** The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica,* 2000, vol. 86 (1), 117-128 **[0051]**
- **STRELCYK, O. ; DAU, T.** Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *The Journal of the Acoustical Society of America,* 2009, vol. 125 (5), 3328-3345 **[0051]**
- **SPILLE, C. ; EWERT, S. D. ; KOLLMEIER, B. ; MEYER, B. T.** Predicting speech intelligibility with deep neural networks. *Computer Speech & Language,* 2018, vol. 48, 51-66 **[0051]**
- **SUMBY, W. H. ; POLLACK, I.** Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America,* 1954, vol. 26 (2), 212-215 **[0052]**
- **MUNHALL, K. G. ; JONES, J. A. ; CALLAN, D. E. ; KURATATE, T. ; VATIKIOTIS-BATESON, E.** Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science,* 2004, vol. 15 (2), 133-137 **[0052]**

- Facial expressions as feedback cue in human-robot interaction-a comparison between human and automatic recognition performances. **LANG, C. ; WACHSMUTH, S. ; WERSING, H. ; HANHEIDE, M.** 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, June 2010, 79-85 **[0054]**