(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 February 2023 (09.02.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/012527 A1

(51) International Patent Classification:
G06V 40/00 (2022.01) *G06V 40/16* (2022.01)
G06V 40/10 (2022.01)

(21) International Application Number:
PCT/IB2022/054527

(22) International Filing Date:
16 May 2022 (16.05.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/229,091 04 August 2021 (04.08.2021) US

(71) Applicant: **Q (CUE) LTD.** [IL/IL]; 4 Lochamei Hagetaot Street, Apt. 5, 4730304 Ramat Hasharon (IL).

(72) Inventors: **MAIZELS, Aviad**; 4 Lochamei Hagetaot Street, Apt. 5, 4730304 Ramat Hasharon (IL). **BARLIYA, Avi**; 8 Vormaiza Street, Tel Aviv (IL). **KORNBLAU, Giora**; 12a Hanesher Street, Binyamina (IL). **WEXLER, Yonatan**; 10 Shai Agnon Boulevard, Jerusalem (IL).

(74) Agent: **KLIGLER, Daniel**; KLIGLER & ASSOCIATES PATENT, P.O. Box 20612, 6120601 Tel Aviv (IL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: DETECTION OF SILENT SPEECH

(57) Abstract: A sensing device (20, 60) includes a bracket (22) configured to fit an ear of a user (24) of the device. An optical sensing head (28) is held by the bracket in a location in proximity to a face of the user and senses light reflected from the face and to output a signal in response to the detected light. Processing circuitry (70, 75) processes the signal to generate a speech output.

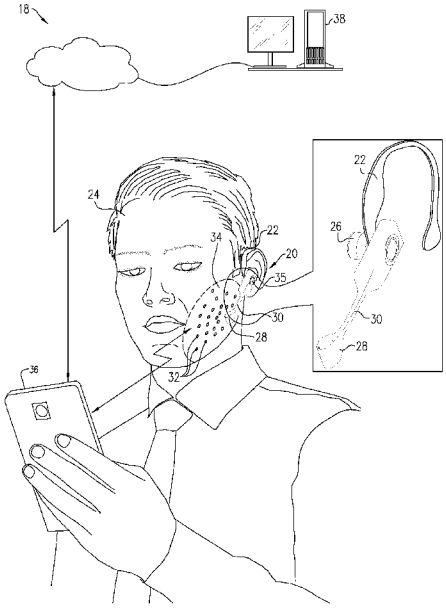


FIG. 1

WO 2023/012527 A1

DETECTION OF SILENT SPEECH

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit of U.S. Provisional Patent Application 63/229,091, filed August 4, 2021, which is incorporated herein by reference.

FIELD OF THE INVENTION

The present invention relates generally to physiological sensing, and particularly to methods and apparatus for sensing human speech.

BACKGROUND

The process of speech activates nerves and muscles in the chest, neck, and face. Thus, for example, electromyography (EMG) has been used to capture muscle impulses for purposes of speech sensing.

Secondary speckle patterns have been used for monitoring movement of skin on the human body. Secondary speckle typically occurs in diffuse reflections of a laser beam from a rough surface, such as the skin. By tracking both temporal and amplitude changes of secondary speckle produced by reflection from human skin when illuminated by a laser beam, investigators have measured blood pulse pressure and other vital signs. For example, U.S. Patent 10,398,314 describes a method for monitoring conditions of a subject's body using image data that is indicative of a sequence of speckle patterns generated by the body.

SUMMARY

Embodiments of the present invention that are described hereinbelow provide novel methods and devices for sensing human speech.

There is also provided, in accordance with an embodiment of the invention, a sensing device, including a bracket configured to fit an ear of a user of the device and an optical sensing head held by the bracket in a location in proximity to a face of the user and configured to sense light reflected from the face and to output a signal in response to the detected light. Processing circuitry is configured to process the signal to generate a speech output.

In one embodiment, the bracket includes an ear clip. Alternatively, the bracket includes a spectacle frame. In a disclosed embodiment, the optical sensing head is configured to sense the light reflected from a cheek of the user.

In some embodiments, the optical sensing head includes an emitter configured to direct coherent light toward the face and an array of sensors configured to sense a secondary speckle

pattern due to reflection of the coherent light from the face. In a disclosed embodiment, the emitter is configured to direct multiple beams of the coherent light toward different, respective locations on the face, and the array of sensors is configured to sense the secondary speckle pattern reflected from the locations. Additionally or alternatively, the locations illuminated by the beams and sensed by the array of sensors extend over an area of at least 1 cm^2. Further additionally or alternatively, the optical sensing head includes multiple emitters, which are configured to generate respective groups of the beams covering different, respective areas of the face, and the processing circuitry is configured to select and actuate a subset of the emitters without actuating all the emitters.

In a disclosed embodiment, the processing circuitry is configured to detect changes in the sensed secondary speckle pattern and to generate the speech output responsively to the detected changes.

Alternatively or additionally, the processing circuitry is configured to operate the array of sensors at a first frame rate, to sense, responsively to the signal while operating at the first frame rate, a movement of the face, and to increase the frame rate responsively to the sensed movement to a second frame rate, greater than the first frame rate, for generating the speech output.

In a disclosed embodiment, the processing circuitry is configured to generate the speech output responsively to changes in the signal output by the optical sensing head due to movements of a skin surface of the user without any utterance of sounds by the user.

Typically, the optical sensing head is held by the bracket in a position that is at least 5 mm away from a skin surface of the user.

In one embodiment, the device includes one or more electrodes configured to contact a skin surface of the user, wherein the processing circuitry is configured to generate the speech output responsively to the electrical activity sensed by the one or more electrodes together with the signal output by the optical sensing head.

Additionally or alternatively, the device includes a microphone configured to sense sounds uttered by the user. In one embodiment, the processing circuitry is configured to compare the signal output by the optical sensing head to the sounds sensed by the microphone in order to calibrate the optical sensing head. Additionally or alternatively, the processing circuitry is configured to change an operational state of the device responsively to sensing of the sounds uttered by the user.

In some embodiments, the device includes a communication interface, wherein the processing circuitry is configured to encode the signal for transmission over the communication

interface to a processing device, which processes the encoded signals to generate the speech output. In a disclosed embodiment, the communication interface includes a wireless interface.

Additionally or alternatively, the device includes a user control, which is connected to the bracket and configured to sense a gesture made by the user, wherein the processing circuitry is configured to change an operational state of the device responsively to the sensed gesture.

Further additionally or alternatively, the device includes a speaker configured to fit in the ear of the user, wherein the processing circuitry is configured to synthesize an audio signal corresponding to the speech output for playback by the speaker.

There is also provided, in accordance with an embodiment of the invention, a method for sensing, which includes sensing a movement of skin on a face of a human subject in response to words articulated by the subject without vocalization of the words by the subject and without contacting the skin. Responsively to the sensed movement, a speech output is generated including the articulated words.

In some embodiments, sensing the movement includes sensing light reflected from the face of the subject. In the disclosed embodiments, sensing the light includes directing coherent light toward the skin and sensing a secondary speckle pattern due to reflection of the coherent light from the skin. In one embodiment, directing the coherent light includes directing multiple beams of the coherent light toward different, respective locations on the face, and sensing the secondary speckle pattern reflected from each of the locations using an array of sensors.

In a disclosed embodiment, generating the speech output includes synthesizing an audio signal corresponding to the speech output. Alternatively or additionally, generating the speech output includes transcribing the words articulated by the subject.

The present invention will be more fully understood from the following detailed description of the embodiments thereof, taken together with the drawings in which:

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a schematic pictorial illustration of a system for speech sensing, in accordance with an embodiment of the invention;

Fig. 2 is a schematic sectional view of an optical sensing head, in accordance with an embodiment of the invention;

Fig. 3 is a schematic pictorial illustration of a speech sensing device, in accordance with another embodiment of the invention;

Fig. 4 is a block diagram that schematically illustrates functional components of a system for speech sensing, in accordance with an embodiment of the invention; and

Fig. 5 is a flow chart that schematically illustrates a method for speech sensing, in accordance with an embodiment of the invention.

DETAILED DESCRIPTION OF EMBODIMENTS

People communicate through their mobile telephones in nearly all locations and at all times. The widespread use of mobile telephones in public spaces creates a cacophony of noise and often raises privacy concerns, since conversations are easily overheard by passersby. At the same time, when one of the parties in a telephone conversation is in a noisy location, the other party or parties may have difficulty in understanding what they are hearing due to background noise. Text communications provide a solution to these problems, but text input to a mobile telephone is slow and interferes with the users' ability to see where they are going.

Embodiments of the present invention that are described herein address these problems using silent speech, enabling users to articulate words and sentences without actually vocalizing the words or uttering any sounds at all. The normal process of vocalization uses multiple groups of muscles and nerves, from the chest and abdomen, through the throat, and up through the mouth and face. To utter a given phoneme, motor neurons activate muscle groups in the face, larynx, and mouth in preparation for propulsion of air flow out of the lungs, and these muscles continue moving during speech to create words and sentences. Without this air flow, no sounds are emitted from the mouth. Silent speech occurs when the air flow from the lungs is absent, while the muscles in the face, larynx, and mouth continue to articulate the desired sounds.

Silent speech can arise as the result of neurological and muscular pathologies; but it can also occur intentionally, for example when we articulate words but do not wish to be heard by others. This articulation can occur even when we conceptualize spoken words without opening our mouths. The resulting activation of our facial muscles gives rise to minute movements of the skin surface. The inventors have found that by properly sensing and decoding these movements, it is possible to reconstruct reliably the actual sequence of words articulated by the user.

Thus, embodiments of the present invention that are described herein sense fine movements of the skin and subcutaneous nerves and muscles on a subject's face, occurring in response to words articulated by the subject with or without vocalization, and use the sensed movements in generating a speech output including the articulated words. These embodiments provide methods and devices for sensing these fine movements without contacting the skin, for

example by sensing light reflected from the subject's face. They thus enable users to communicate with others or to record their own thoughts silently, in a manner that is substantially imperceptible to other parties. Devices and methods in accordance with these embodiments are also insensitive to ambient noise and can be used in substantially any environment, without requiring users to divert their view and attention from their surroundings.

Some embodiments of the present invention provide sensing devices having the form of common consumer items, such as a clip-on headphone or spectacles. In these embodiments, an optical sensing head is held in a location in proximity to the user's face by a bracket that fits in or over the user's ear. The optical sensing head senses light reflected from the face, for example by directing coherent light toward an area of the face, such as the cheek, and sensing changes in the secondary speckle pattern that arises due to reflection of the coherent light from the face. Processing circuitry in the device processes the signal output by the optical sensing head due to the reflected light to generate a corresponding speech output.

Alternatively, the principles of the present invention may be implemented without an ear clip or other bracket. For example, in an alternative embodiment, a silent speech sensing module, including a coherent light source and sensors, may be integrated into a mobile communication device, such as a smartphone. This integrated sensing module senses silent speech when the user holds the mobile communication device in a suitable location in proximity to the user's face.

The term "light," as used in the present description and in the claims, refers to electromagnetic radiation in any or all of the infrared, visible, and ultraviolet ranges.

Fig. 1 is a schematic pictorial illustration of a system 18 for speech sensing, in accordance with an embodiment of the invention. System 18 is based on a sensing device 20, in which a bracket, in the form of an ear clip 22, fits over the ear of a user 24 of the device. An earphone 26 attached to ear clip 22 fits into the user's ear. An optical sensing head 28 is connected by an arm 30 to ear clip 22 and thus is held in a location in proximity to the user's face. In the pictured embodiment, device 20 has the form and appearance of a clip-on headphone, with the optical sensing head in place of (or in addition to) the microphone.

Optical sensing head 28 directs one or more beams of coherent light toward different, respective locations on the face of user 24, thus creating an array of spots 32 extending over an area 34 of the face (and specifically over the user's cheek). In the present embodiment, optical sensing head 28 does not contact the user's skin at all, but rather is held at a certain distance from the skin surface. Typically, this distance is at least 5 mm, and it may be even greater, for example at least 1 cm or even 2 cm or more from the skin surface. To enable sensing the motion of different

parts of the facial muscles, the area 34 covered by spots 32 and sensed by optical sensing head 28 typically has an extent of at least 1 cm^2; and larger areas, for example at least 2 cm^2 or even greater than 4 cm^2, can be advantageous.

Optical sensing head 28 senses the coherent light that is reflected from spots 32 the face and outputs a signal in response to the detected light. Specifically, optical sensing head 28 senses the secondary speckle patterns that arise due to reflection of the coherent light from each of spots 32 within its field of view. To cover a sufficiently large area 34, this field of view typically has a wide angular extent, typically with an angular width of at least 60^O, or possibly 70^O or even 90^O or more. Within this field of view, device 20 may sense and process the signals due to the secondary speckle patterns of all of spots 32 or of only a certain subset of spots 32. For example, device 20 may select a subset of the spots that is found to give the largest amount of useful and reliable information with respect to the relevant movements of the skin surface of user 24. Details of the structure and operation of optical sensing head 28 are described hereinbelow with reference to Fig. 2.

Within system 18, processing circuitry processes the signal that is output by optical sensing head 28 to generate a speech output. As noted earlier, the processing circuitry is capable of sensing movements of the skin of user 22 and generating the speech output, even without vocalization of the speech or utterance of any other sounds by user 22. The speech output may take the form of a synthesized audio signal or a textual transcription, or both. The synthesized audio signal may be played back via the speaker in earphone 26 (and is useful in giving user 22 feedback with respect to the speech output). Additionally or alternatively, the synthesized audio signal may be transmitted over a network, for example via a communication link with a mobile communication device, such as a smartphone 36.

The functions of the processing circuitry in system 18 may be carried out entirely within device 20, or they may alternatively be distributed between device 20 and an external processor, such as a processor in smartphone 36 running suitable application software. For example, the processing circuitry within device 20 may digitize and encode the signals output by optical sensing head 28 and transmit the encoded signals over the communication link to smartphone 36. This communication link may be wired or wireless, for example using the Bluetooth™ wireless interface provided by the smartphone. The processor in smartphone 36 processes the encoded signal in order to generate the speech output. Smartphone 36 may also access a server 38 over a data network, such as the Internet, in order to upload data and download software updates, for

example. Details of the design and operation of the processing circuitry are described hereinbelow with reference to Fig. 4.

In the pictured embodiment, device 20 also comprises a user control 35, for example in the form of a push-button or proximity sensor, which is connected to ear clip 22. User control 35 senses gestures performed by user, such as pressing on user control 35 or otherwise bringing the user's finger or hand into proximity with the user control. In response to the appropriate user gesture, the processing circuitry changes the operational state of device 20. For example, user 24 may switch device 20 from an idle mode to an active mode in this fashion, and thus signal that the device should begin sensing and generating a speech output. This sort of switching is useful in conserving battery power in device 20. Alternatively or additionally, other means may be applied in controlling the operational state of device 20 and reducing unnecessary power consumption, for example as described below with reference to Fig. 5.

Fig. 2 is a schematic sectional view of optical sensing head 28 of device 20, showing components and functional details of the optical sensing head in accordance with an embodiment of the invention. Optical sensing head 28 comprises an emitter module 40 and a receiver module 48, along with an optional microphone 54.

Emitter module 40 comprises a light source, such as an infrared laser diode 42, which emits an input beam of coherent radiation. A beamsplitting element 44, such as a Damman grating or another suitable type of diffractive optical element (DOE), splits the input beam into multiple output beams 46, which form respective spots 32 at a matrix of locations extending over area 34. In one embodiment (not shown in the figures) emitter module 40 comprises multiple laser diodes or other emitters, which generate respective groups of the output beams 46, covering different respective sub-areas within area 34 of the user's face. In this case, the processing circuitry in device 20 may select and actuate only a subset of the emitters, without actuating all the emitters. For example, to reduce the power consumption of device 20, the processing circuitry may actuate only one emitter or a subset consisting of two or more emitters that illuminates the area on the user's face that has been found to give the most useful information for generating the desired speech output.

Receiver module 48 comprises an array 52 of optical sensors, for example, a CMOS image sensor, with objective optics 50 for imaging area 34 onto array 52. Because of the small dimensions of optical sensing head 28 and its proximity to the skin surface, receiver module 48 has a sufficiently wide field of view, as noted above, and views many of spots 32 at a high angle,

far from the normal. Because of the roughness of the skin surface, the secondary speckle patterns at spots 32 can be detected at these high angles, as well.

Microphone 54 senses sounds uttered by user 24, enabling user 22 to use device 20 as a conventional headphone when desired. Additionally or alternatively, microphone 54 may be used in conjunction with the silent speech sensing capabilities of device 20. For example, microphone 54 may be used in a calibration procedure, in which optical sensing head 28 senses movement of the skin while user 22 utters certain phonemes or words. The processing circuitry may then compare the signal output by optical sensing head 28 to the sounds sensed by microphone 54 in order to calibrate the optical sensing head. This calibration may include prompting user 22 to shift the position of optical sensing head 28 in order to align the optical components in the desired position relative to the user's cheek.

In another embodiment, the audio signals output by microphone 54 can be used in changing the operational state of device 20. For example, the processing circuitry may generate the speech output only if microphone 54 does not detect vocalization of words by user 24. Other applications of the combination of optical and acoustic sensing that is provided by optical sensing head 28 with microphone 54 will be apparent to those skilled in the art after reading the present description and are considered to be within the scope of the present invention.

Fig. 3 is a schematic pictorial illustration of a speech sensing device 60, in accordance with another embodiment of the invention. In this embodiment, ear clip 22 is integrated with or otherwise attached to a spectacle frame 62. Nasal electrodes 64 and temporal electrodes 66 are attached to frame 62 and contact the user's skin surface. Electrodes 64 and 66 receive body surface electromyogram (sEMG) signals, which provide additional information regarding the activation of the user's facial muscles. The processing circuitry in device 60 uses the electrical activity sensed by electrodes 64 and 66 together with the output signal from optical sensing head 28 in generating the speech output from device 60.

Additionally or alternatively, device 60 includes one or more additional optical sensing heads 68, similar to optical sensing head 28, for sensing skin movements in other areas of the user's face. These additional optical sensing heads may be used together with or instead of optical sensing head 28.

Fig. 4 is a block diagram that schematically illustrates functional components of system 18 for speech sensing, in accordance with an embodiment of the invention. The pictured system is built around the components shown in Fig. 1, including sensing device 20, smartphone 36, and server 38. Alternatively, the functions illustrated in Fig. 4 and described below may be

implemented and distributed differently among the components of the system. For example, some or all of the processing capabilities attributed to smartphone 36 may be implemented in sensing device; or the sensing capabilities of device 20 may be implemented in smartphone 36.

In the pictured example, as explained above, sensing device 20 comprises emitter module 40, receiver module 48, speaker 26, microphone 54, and user control (UI) 35. For the sake of completeness, sensing device 20 is shown in Fig. 4 as comprising other sensors 71, as well, such as electrodes and/or environmental sensors; but as noted earlier, sensing device 20 is capable of operation based solely on non-contact measurements made by the emitter and receiver modules.

Sensing device 20 comprises processing circuitry in the form of an encoder 70 and a controller 75. Encoder 70 comprises hardware processing logic, which may be hard-wired or programmable, and/or a digital signal processor, which extracts and encodes features of the output signal from receiver module 48. Sensing device 20 transmits the encoded signals via a communication interface 72, such as a Bluetooth interface, to a corresponding communication interface 77 in smartphone 36. A battery 74 provides operating power to the components of sensing device 20.

Controller 75 comprises a programmable microcontroller, for example, which sets the operating state and operational parameters of sensing device 20 based on inputs received from user control 35, receiver module 48, and smartphone 36 (via communication interface 72). Some aspects of this functionality are described below with reference to Fig. 5. In an alternative embodiment, controller 75 comprises a more powerful microprocessor and/or a processing array, which processes the features of the output signals from receiver module 48 locally within sensing device and generates a speech output, independently of smartphone 36.

In the present embodiment, however, the encoded output signals from sensing device 20 are received in a memory 78 of smartphone 36 and processed by a speech generation application 80 running on the processor in smartphone 36. Speech generation application 80 converts the features in the output signal to a sequence of words, in the form of text and/or an audio output signal. Communication interface 77 passes the audio output signal back to speaker 26 of sensing device 20 for playback to the user. The text and/or audio output from speech generation application 80 is also input to other applications 84, such as voice and/or text communication applications, as well as a recording application. The communication applications communicate over a cellular or Wi-Fi network, for example, via a data communication interface 86.

The operations of encoder 70 and speech generation application 80 are controlled by a local training interface 82. For example, interface 82 may indicate to encoder 70 which temporal and

spectral features to extract from the signals output by receiver module 48 and may provide speech generation application 80 with coefficients of a neural network, which converts the features to words. In the present example, speech generation application 80 implements an inference network, which finds the sequence of words having the highest probability of corresponding to the encoded signal features received from sensing device 20. Local training interface 82 receives the coefficients of the inference network from server 38, which may also update the coefficients periodically.

To generate local training instructions 82, server 38 uses a data repository 88 containing speckle images and corresponding ground truth spoken words from a collection of training data 90. Repository 88 also receives training data collected from sensing devices 20 in the field. For example, the training data may comprise signals collected from sensing devices 20 while users articulate certain sounds and words (possibly including both silent and vocalized speech). This combination of general training data 90 with personal training data received from the user of each sensing device 20 enables server 38 to derive optimal inference network coefficients for each user.

Server 38 applies image analysis tools 94 to extract features from the speckle images in repository 88. These image features are input as training data to a neural network 96, together with a corresponding dictionary 104 of words and a language model 100, which defines both the phonetic structure and syntactical rules of the specific language used in the training data. Neural network 96 generates optimal coefficients for an inference network 102, which converts an input sequence of feature sets, which have been extracted from a corresponding sequence of speckle measurements, into corresponding phonemes and ultimately into an output sequence of words. Further details of the network architecture and training process are described in the above-mentioned provisional patent application. Server 38 downloads the coefficients of inference network 102 to smartphone 36 for used in speech generation application 80.

Fig. 5 is a flow chart that schematically illustrates a method for speech sensing, in accordance with an embodiment of the invention. This method is described, for the sake of convenience and clarity, with reference to the elements of system 18, as shown in Figs. 1 and 4 and described above. Alternatively, the principles of this method may be applied in other system configurations, for example using sensing device 60 (Fig. 3) or a sensing device that is integrated in a mobile communication device.

As long as user 24 is not speaking, sensing device 20 operates in a low-power idle mode in order to conserve power in battery 74, at an idling step 110. In this mode, controller 75 drives array 52 of sensors in receiver module 48 at a low frame rate, for example twenty frames/sec.

Emitter module 40 may also operate at a reduced output power. While receiver module 48 operates at this low frame rate, controller 75 processes the images output by array 52 in order to detect a movement of the face that is indicative of speech, at a motion detection step 112. When such movement is detected, controller 75 instructs receiver module 48, as well as other components of sensing device 20 to increase the frame rate, for example to the range of 100-200 frames/sec, to enable detection of changes in the secondary speckle patterns that occur due to silent speech, at an active capture step 114. Alternatively or additionally, controller 75 may increase the frame rate and power up other components of sensing device 20 in response to other inputs, such as actuation of user control 35 or instructions received from smartphone 36.

The images captured by receiver module 48 typically contain a matrix of projected laser spots 32, as illustrated in Fig. 1. Encoder 70 detects the locations of the spots in the images, at a spot detection 116. The encoder may extract features from all the spots; but to conserve power and processing resources, it is desirable that the encoder select a subset of the spots. For example, local training interface 82 may indicate which subset of spots contains the greatest amount of information with respect to the user's speech, and encoder 70 may select the spots in this subset. Encoder 70 crops small windows from each image, with each such window containing one of the selected spots, at a cropping step 118.

Encoder 70 extracts features of speckle motion from each selected spot, at a feature extraction step 120. For example, encoder 70 may estimate the total energy in each speckle, based on the average intensity of the pixels in the corresponding window, and may measure the changes in energy of each speckle over time. Additionally or alternatively, encoder 70 may extract other temporal and/or spectral features of the speckles in the selected subset of spots. Encoder 70 conveys these features to speech generation application 80 (running on smartphone 36), which inputs vectors of the feature values to the inference network 102 that was downloaded from server 38, at a feature input step 122.

Based on the sequence of feature vectors that is input to the inference network over time, speech generation application 80 outputs a stream of words, which are concatenated together into sentences, at a speech output step 124. As noted earlier, the speech output is used to synthesize an audio signal, for playback via speaker 26. Other applications 84 running on smartphone 36 post-process the speech and/or audio signal to record the corresponding text and/or to transmit speech or text data over a network, at a post-processing step 126.

It will be appreciated that the embodiments described above are cited by way of example, and that the present invention is not limited to what has been particularly shown and described

hereinabove. Rather, the scope of the present invention includes both combinations and subcombinations of the various features described hereinabove, as well as variations and modifications thereof which would occur to persons skilled in the art upon reading the foregoing description and which are not disclosed in the prior art.

5

CLAIMS

1. A sensing device, comprising:

a bracket configured to fit an ear of a user of the device;

an optical sensing head held by the bracket in a location in proximity to a face of the user and configured to sense light reflected from the face and to output a signal in response to the detected light; and

processing circuitry configured to process the signal to generate a speech output.

2. The device according to claim 1, wherein the bracket comprises an ear clip.

3. The device according to claim 1, wherein the bracket comprises a spectacle frame.

4. The device according to claim 1, wherein the optical sensing head is configured to sense the light reflected from a cheek of the user.

5. The device according to claim 1, wherein the optical sensing head comprises an emitter configured to direct coherent light toward the face and an array of sensors configured to sense a secondary speckle pattern due to reflection of the coherent light from the face.

6. The device according to claim 5, wherein the emitter is configured to direct multiple beams of the coherent light toward different, respective locations on the face, and the array of sensors is configured to sense the secondary speckle pattern reflected from the locations.

7. The device according to claim 6, wherein the locations illuminated by the beams and sensed by the array of sensors extend over a field of view having an angular width of at least 60^{o}.

8. The device according to claim 6, wherein the locations illuminated by the beams and sensed by the array of sensors extend over an area of at least 1 cm^2.

9. The device according to claim 6, wherein the optical sensing head comprises multiple emitters, which are configured to generate respective groups of the beams covering different, respective areas of the face, and wherein the processing circuitry is configured to select and actuate a subset of the emitters without actuating all the emitters.

10. The device according to claim 5, wherein the processing circuitry is configured to detect changes in the sensed secondary speckle pattern and to generate the speech output responsively to the detected changes.

11. The device according to claim 5, wherein the processing circuitry is configured to operate the array of sensors at a first frame rate, to sense, responsively to the signal while operating at the first frame rate, a movement of the face, and to increase the frame rate responsively to the sensed movement to a second frame rate, greater than the first frame rate, for generating the speech output.

12. The device according to claims 1-11, wherein the processing circuitry is configured to generate the speech output responsively to changes in the signal output by the optical sensing head due to movements of a skin surface of the user without any utterance of sounds by the user.

14. The device according to claims 1-11, wherein the optical sensing head is held by the bracket in a position that is at least 5 mm away from a skin surface of the user.

15. The device according to claims 1-11, and comprising one or more electrodes configured to contact a skin surface of the user, wherein the processing circuitry is configured to generate the speech output responsively to the electrical activity sensed by the one or more electrodes together with the signal output by the optical sensing head.

16. The device according to claims 1-11, and comprising a microphone configured to sense sounds uttered by the user.

17. The device according to claim 16, wherein the processing circuitry is configured to compare the signal output by the optical sensing head to the sounds sensed by the microphone in order to calibrate the optical sensing head.

18. The device according to claim 16, wherein the processing circuitry is configured to change an operational state of the device responsively to sensing of the sounds uttered by the user.

19. The device according to claims 1-11, and comprising a communication interface, wherein the processing circuitry is configured to encode the signal for transmission over the communication interface to a processing device, which processes the encoded signals to generate the speech output.

20. The device according to claim 17, wherein the communication interface comprises a wireless interface.

21. The device according to claims 1-11, and comprising a user control, which is connected to the bracket and configured to sense a gesture made by the user, wherein the processing circuitry is configured to change an operational state of the device responsively to the sensed gesture.

22. The device according to claims 1-11, and comprising a speaker configured to fit in the ear of the user, wherein the processing circuitry is configured to synthesize an audio signal corresponding to the speech output for playback by the speaker.

23. A method for sensing, comprising:

sensing a movement of skin on a face of a human subject in response to words articulated by the subject without vocalization of the words by the subject and without contacting the skin; and

responsively to the sensed movement, generating a speech output including the articulated words.

24. The method according to claim 23, wherein sensing the movement comprises sensing light reflected from the face of the subject.

25. The method according to claim 24, wherein sensing the light comprises directing coherent light toward the skin and sensing a secondary speckle pattern due to reflection of the coherent light from the skin.

26. The method according to claim 25, wherein directing the coherent light comprises directing multiple beams of the coherent light toward different, respective locations on the face, and sensing the secondary speckle pattern reflected from each of the locations using an array of sensors.

27. The method according to claim 26, wherein the locations illuminated by the beams and sensed by the array of sensors extend over a field of view having an angular width of at least 60^{o}.

28. The method according to claim 26, wherein the locations illuminated by the beams and sensed by the array of sensors extend over an area of at least 1 cm^2 on a cheek of the subject.

29. The method according to claim 25, wherein generating the speech output comprises detecting changes in the sensed secondary speckle pattern and generating the speech output responsively to the detected changes.

30. The method according to any of claims 23-29, wherein generating the speech output comprises synthesizing an audio signal corresponding to the speech output.

31. The method according to any of claims 23-29, wherein generating the speech output comprises transcribing the words articulated by the subject.

FIG. 1

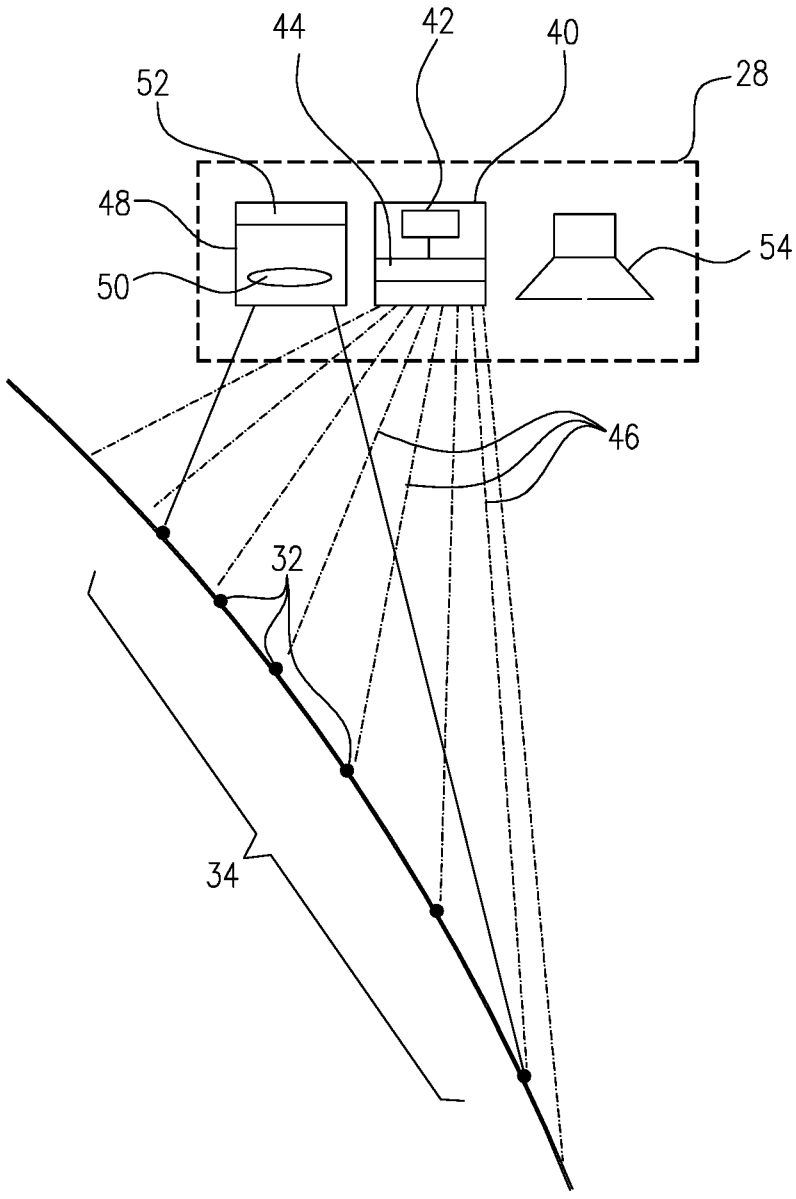


FIG. 2

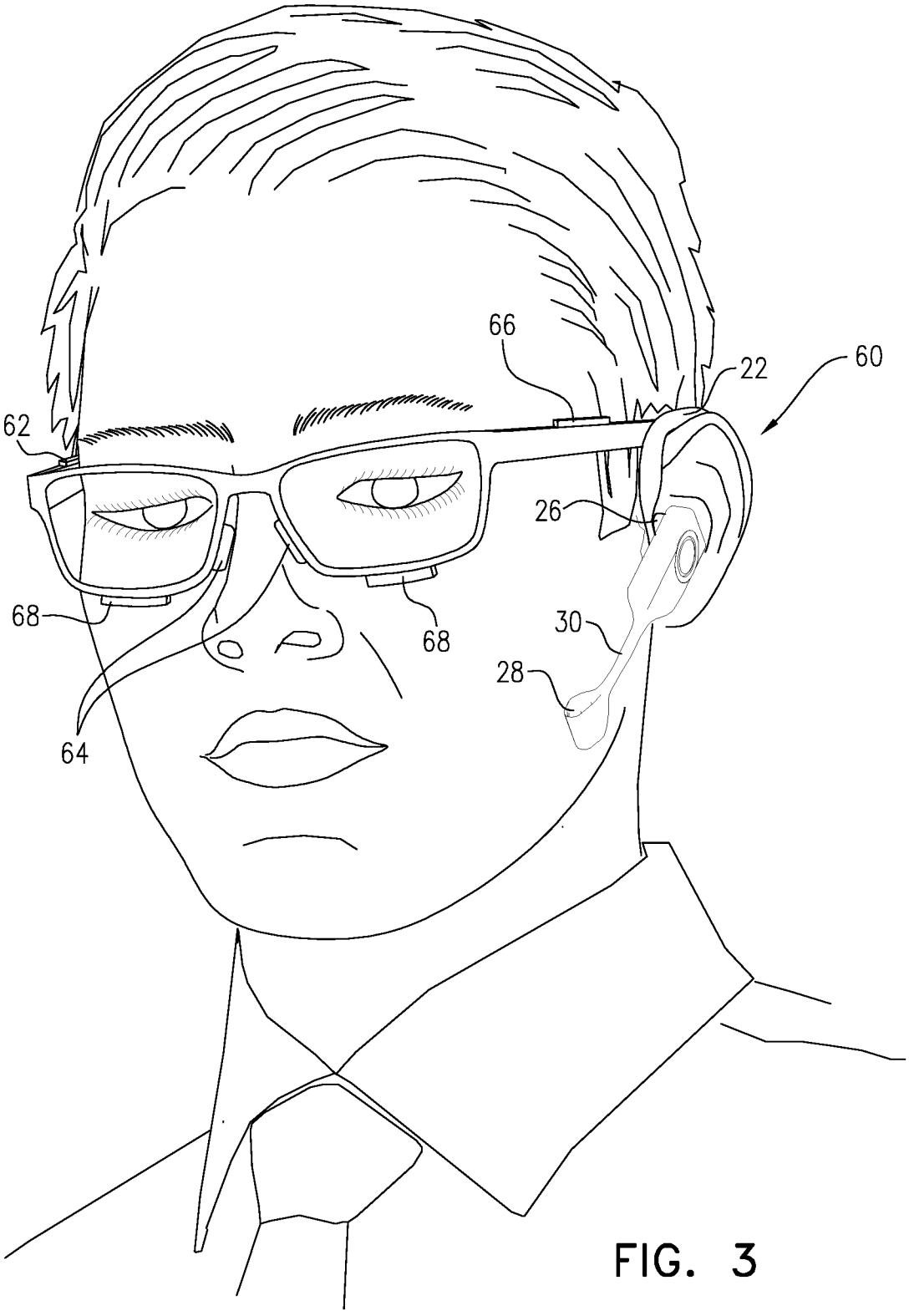


FIG. 3

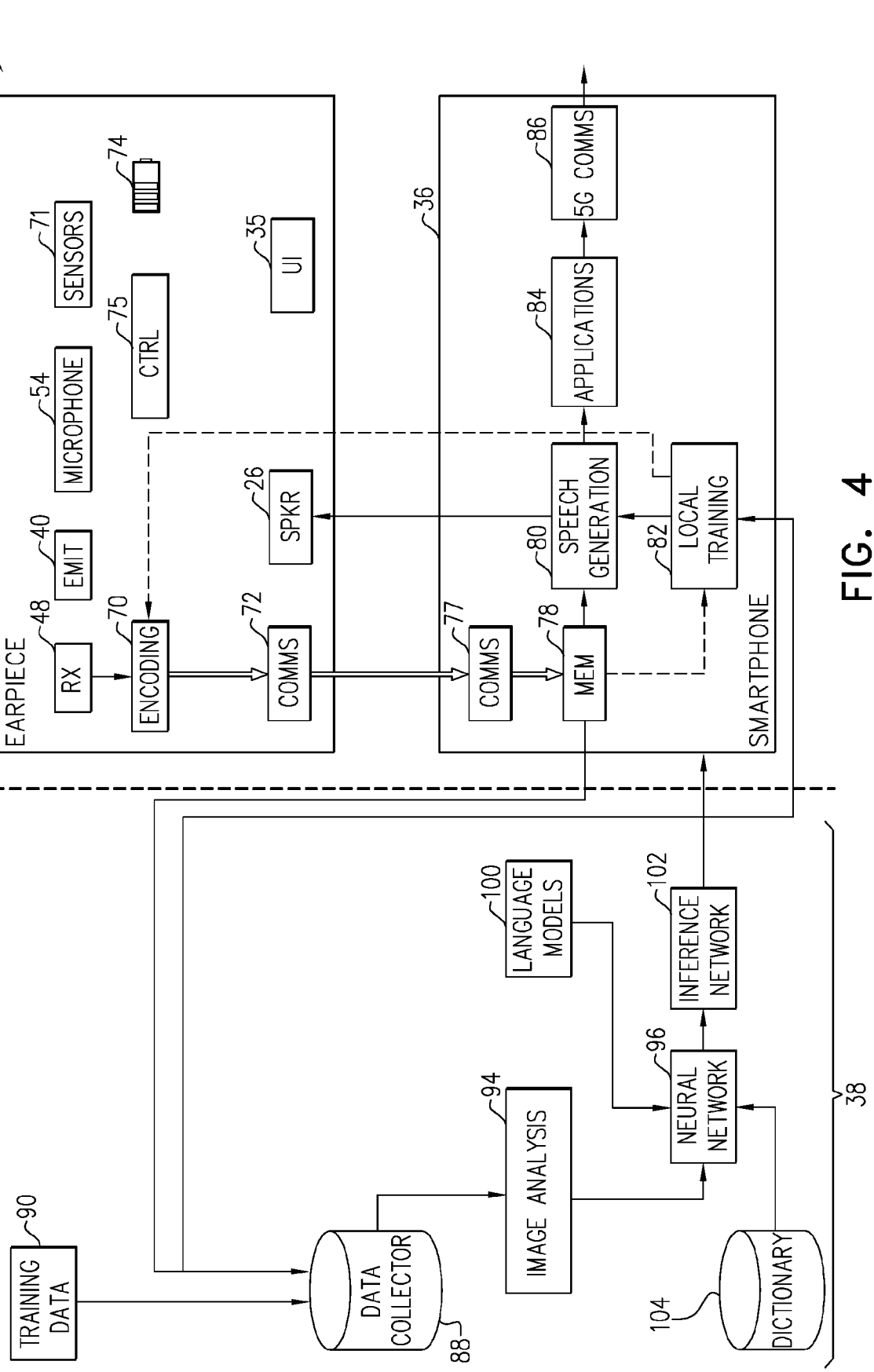

18
20
EARPIECE
48
RX
40
EMIT
54
MICROPHONE
71
SENSORS
74
70
ENCODING
75
CTRL
72
COMMS
26
SPKR
35
UI
36
SMARTPHONE
77
COMMS
78
MEM
80
SPEECH GENERATION
82
LOCAL TRAINING
84
APPLICATIONS
86
5G COMMS
90
TRAINING DATA
DATA COLLECTOR
88
94
IMAGE ANALYSIS
96
NEURAL NETWORK
100
LANGUAGE MODELS
102
INFERENCE NETWORK
104
DICTIONARY
38

FIG. 4

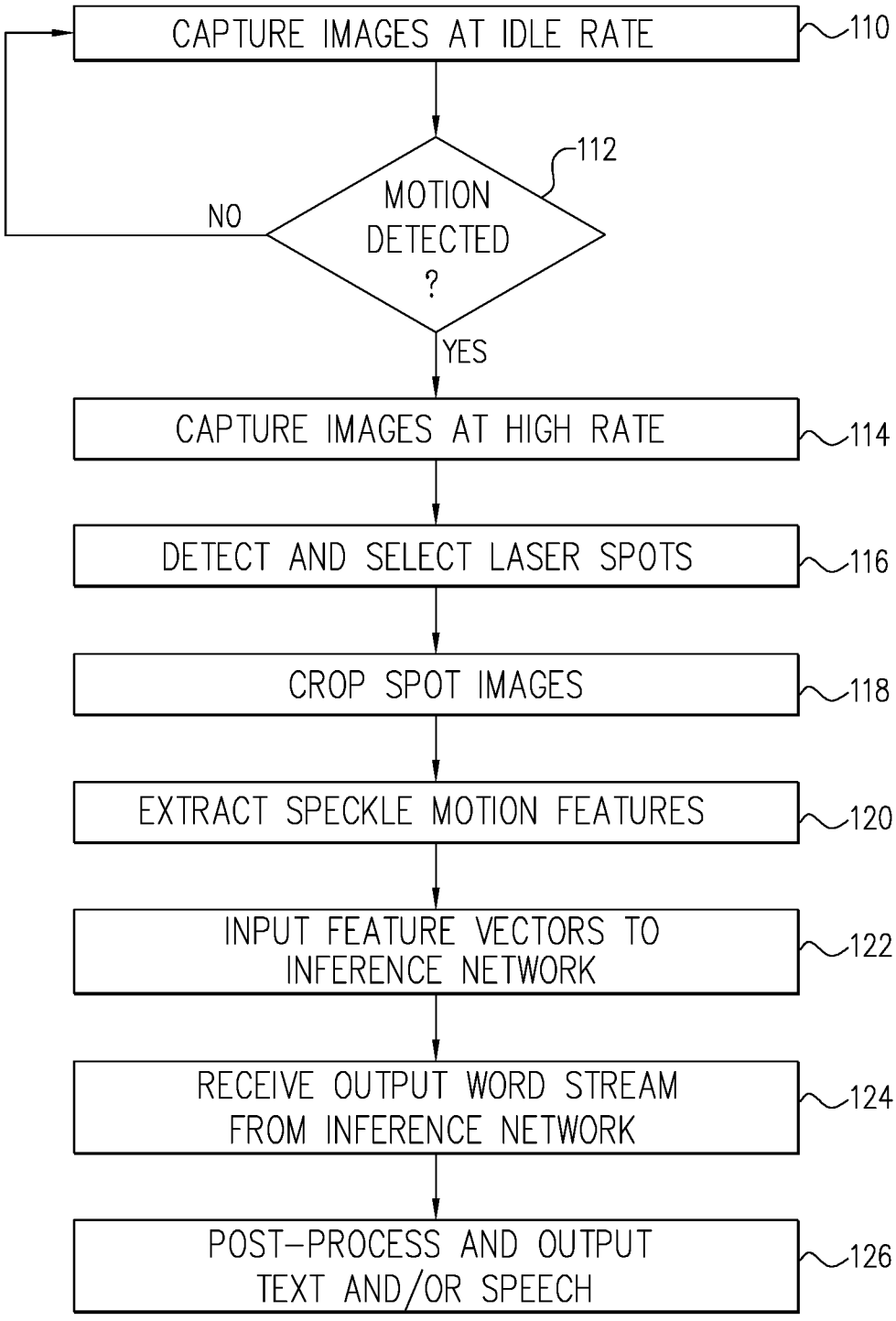

CAPTURE IMAGES AT IDLE RATE
110
MOTION DETECTED ?
112
NO
YES
CAPTURE IMAGES AT HIGH RATE
114
DETECT AND SELECT LASER SPOTS
116
CROP SPOT IMAGES
118
EXTRACT SPECKLE MOTION FEATURES
120
INPUT FEATURE VECTORS TO INFERENCE NETWORK
122
RECEIVE OUTPUT WORD STREAM FROM INFERENCE NETWORK
124
POST-PROCESS AND OUTPUT TEXT AND/OR SPEECH
126

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2022/054527

A. CLASSIFICATION OF SUBJECT MATTER

G06V 40/00(2022.01)i; ***G06V 40/10***(2022.01)i; ***G06V 40/16***(2022.01)i
CPC:G06V 40/00; G06V 40/10; G06V 40/16

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06V 40/00; G06V 40/10
CPC:G06V 40/00; G06V 40/10

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

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

Databases consulted: Esp@cenet, Google Patents, Orbit

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016027441 A1 (AL LIU CHING FENG et al) 28 January 2016 (2016-01-28) whole document	1-31
A	US 2019189145 A1 (RAKSHIT SARBAJIT) 20 June 2019 (2019-06-20) whole document	1-31

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 August 2022	**30 August 2022**

Name and mailing address of the ISA/IL	Authorized officer
Israel Patent Office **Technology Park, Bldg.5, Malcha, Jerusalem, 9695101,** **Israel** **Israel**	**ZOZULYA Irina**
Telephone No. **972-73-3927253** Email: **pctoffice@justice.gov.il**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2022/054527

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2016027441 A1	28 January 2016	US 2016027441 A1	28 January 2016
		US 9424842 B2	23 August 2016
		CN 105321519 A	10 February 2016
		CN 105321519 B	14 May 2019
		EP 2980788 A1	03 February 2016
		JP 2018028681 A	22 February 2018
		JP 6484317 B2	13 March 2019
		JP 2016031534 A	07 March 2016
		TW 201604864 A	01 February 2016
		TW I576826 B	01 April 2017
US 2019189145 A1	20 June 2019	US 2019189145 A1	20 June 2019
		US 10529355 B2	07 January 2020
		US 2019371356 A1	05 December 2019
		US 10679644 B2	09 June 2020

(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118235174 A
(43)申请公布日 2024.06.21

(21)申请号 202280052345.4
(22)申请日 2022.05.16
(30)优先权数据
63/229,091 2021.08.04 US
(85)PCT国际申请进入国家阶段日
2024.01.25
(86)PCT国际申请的申请数据
PCT/IB2022/054527 2022.05.16
(87)PCT国际申请的公布数据
WO2023/012527 EN 2023.02.09
(71)申请人 库伊有限公司
地址 以色列拉马特甘
(72)发明人 阿维亚德·梅泽尔斯
阿维·巴里亚 乔拉·科恩布劳
约纳坦·韦克斯勒
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 陆建萍 杨明钊
(51)Int.Cl.
G06V 40/00(2006.01)
G06V 40/16(2006.01)
G06V 40/10(2006.01)

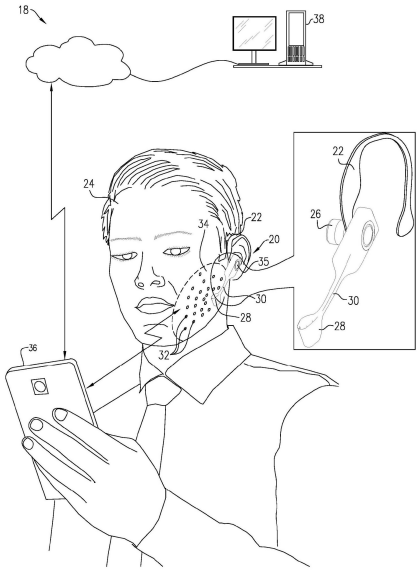
权利要求书2页 说明书7页 附图5页

(54)发明名称

无声语音检测

(57)摘要

一种感测设备(20,60)包括支架(22),该支架被配置为适配该设备的用户(24)的耳朵。光学感测头(28)由支架保持在靠近用户面部的位置,并且感测从面部反射的光,并且响应于检测到的光而输出信号。处理电路(70,75)处理该信号以生成语音输出。



CN 118235174 A

1.一种感测设备,包括:

支架,所述支架被配置为适配所述设备的用户的耳朵;

光学感测头,所述光学感测头由所述支架保持在靠近用户的面部的位置,并且被配置为感测从所述面部反射的光并响应于检测到的光而输出信号;和

处理电路,所述处理电路被配置为处理所述信号以生成语音输出。

2.根据权利要求1所述的设备,其中,所述支架包括耳夹。

3.根据权利要求1所述的设备,其中,所述支架包括眼镜框架。

4.根据权利要求1所述的设备,其中,所述光学感测头被配置为感测从用户的脸颊反射的光。

5.根据权利要求1所述的设备,其中,所述光学感测头包括发射器和感测器阵列,所述发射器被配置为将相干光引导到所述面部,所述感测器阵列被配置为感测由于所述相干光从所述面部的反射而产生的二次散斑图案。

6.根据权利要求5所述的设备,其中,所述发射器被配置为将所述相干光的多个光束引导到所述面部上的不同的相应位置,并且所述感测器阵列被配置为感测从所述位置反射的二次散斑图案。

7.根据权利要求6所述的设备,其中,由所述光束照射并由所述感测器阵列感测的所述位置在具有至少60°角宽度的视场上延伸。

8.根据权利要求6所述的设备,其中,由所述光束照射并由所述感测器阵列感测的所述位置在至少1cm^2的区域上延伸。

9.根据权利要求6所述的设备,其中,所述光学感测头包括多个发射器,所述多个发射器被配置为产生所述光束的相应组,所述光束的相应组覆盖所述面部的不同的相应区域,并且其中,所述处理电路被配置为选择和致动所述发射器的子集,而不致动所有发射器。

10.根据权利要求5所述的设备,其中,所述处理电路被配置为检测感测到的二次散斑图案的变化,并响应于检测到的变化来生成所述语音输出。

11.根据权利要求5所述的设备,其中,所述处理电路被配置为以第一帧速率操作所述感测器阵列,响应于当以所述第一帧速率操作时的所述信号来感测所述面部的运动,并且响应于感测到的运动来将帧速率增加到大于所述第一帧速率的第二帧速率,以生成所述语音输出。

12.根据权利要求1-11所述的设备,其中,所述处理电路被配置为响应于由于在用户不发出任何声音的情况下用户的皮肤表面的运动导致的、由所述光学感测头输出的所述信号的变化,生成所述语音输出。

14.根据权利要求1-11所述的设备,其中,所述光学感测头由所述支架保持在距离用户的皮肤表面至少5mm的位置。

15.根据权利要求1-11所述的设备,并且包括一个或更多个电极,所述一个或更多个电极被配置为接触用户的皮肤表面,其中,所述处理电路被配置为响应于由所述一个或更多个电极感测到的电活动以及由所述光学感测头输出的所述信号来生成所述语音输出。

16.根据权利要求1-11所述的设备,并且包括麦克风,所述麦克风被配置为感测用户发出的声音。

17.根据权利要求16所述的设备,其中,所述处理电路被配置为将所述光学感测头输出

的所述信号与所述麦克风感测到的声音进行比较,以便校准所述光学感测头。

18.根据权利要求16所述的设备,其中,所述处理电路被配置为响应于对用户发出的声音的感测来改变所述设备的操作状态。

19.根据权利要求1-11所述的设备,并且包括通信接口,其中,所述处理电路被配置为对所述信号进行编码,以便通过所述通信接口传输到处理设备,所述处理设备处理经编码的信号以生成所述语音输出。

20.根据权利要求17所述的设备,其中,所述通信接口包括无线接口。

21.根据权利要求1-11所述的设备,并且包括用户控件,所述用户控件连接到所述支架并被配置为感测用户做出的手势,其中,所述处理电路被配置为响应于感测到的手势来改变所述设备的操作状态。

22.根据权利要求1-11所述的设备,并且包括扬声器,所述扬声器被配置为适配在用户的耳朵内,其中,所述处理电路被配置为合成对应于所述语音输出的音频信号,以用于由所述扬声器回放。

23.一种感测方法,包括:

响应于人类对象说单词但所述对象不将所述单词发声出来,并且在不接触所述对象的面部上的皮肤的情况下,感测所述皮肤的运动;和

响应于感测到的运动,生成语音输出,所述语音输出包括被说出的单词。

24.根据权利要求23所述的方法,其中,感测所述运动包括感测从所述对象的面部反射的光。

25.根据权利要求24所述的方法,其中,感测所述光包括将相干光引导到所述皮肤并感测由于所述相干光从所述皮肤的反射而产生的二次散斑图案。

26.根据权利要求25所述的方法,其中,引导所述相干光包括将所述相干光的多个光束引导到所述面部上的不同的相应位置,并使用感测器阵列感测从每个位置反射的二次散斑图案。

27.根据权利要求26所述的方法,其中,由所述光束照射并由所述感测器阵列感测的所述位置在具有至少60°角宽度的视场上延伸。

28.根据权利要求26所述的方法,其中,由所述光束照射并由所述感测器阵列感测的所述位置在所述对象的脸颊上的至少1cm^2的区域上延伸。

29.根据权利要求25所述的方法,其中,生成所述语音输出包括检测感测到的二次散斑图案的变化,并响应于检测到的变化来生成所述语音输出。

30.根据权利要求23-29中任一项所述的方法,其中,生成所述语音输出包括合成对应于所述语音输出的音频信号。

31.根据权利要求23-29中任一项所述的方法,其中,生成所述语音输出包括转录由所述对象说出的单词。

无声语音检测

[0001] 相关申请的交叉引用

[0002] 本申请要求于2021年8月4日提交的美国临时专利申请63/229,091的利益,该美国临时专利申请通过引用并入本文。

发明领域

[0003] 本发明总体上涉及生理感测,尤其涉及用于感测人类语音的方法和装置。

[0004] 背景

[0005] 说话的过程会激活胸部、颈部和面部的神经和肌肉。因此,例如,肌电图(EMG)已被用于捕获肌肉脉冲以用于语音感测。

[0006] 二次散斑图案已被用于监测人体上的皮肤的运动。二次散斑通常出现在激光束从粗糙表面(例如皮肤)的漫反射中。通过跟踪由人类皮肤在被激光束照射时进行的反射产生的二次散斑的时间和振幅变化,研究人员测量了血压(blood pulse pressure)和其他生命体征。例如,美国专利10,398,314描述了一种使用图像数据监测对象身体的状况的方法,该图像数据指示由身体产生的散斑图案序列。

[0007] 概述

[0008] 下面描述的本发明的实施例提供了用于感测人类语音的新方法和设备。

[0009] 根据本发明的实施例,还提供了一种感测设备,该感测设备包括支架和光学感测头,支架被配置为适配该设备的用户的耳朵,光学感测头由支架保持在靠近用户面部的位置,并且被配置为感测从面部反射的光并响应于检测到的光而输出信号。处理电路被配置为处理该信号以生成语音输出。

[0010] 在一个实施例中,支架包括耳夹。可替代地,支架包括眼镜框架。在公开的实施例中,光学感测头被配置为感测从用户的脸颊反射的光。

[0011] 在一些实施例中,光学感测头包括发射器和感测器阵列,发射器被配置为将相干光引导到面部,感测器阵列被配置为感测由于相干光从面部的反射而产生的二次散斑图案。在公开的实施例中,发射器被配置为将相干光的多个光束引导到面部上的不同的相应位置,并且感测器阵列被配置为感测从这些位置反射的二次散斑图案。附加地或可替代地,由光束照射并由感测器阵列感测的位置在至少1cm^2的区域上延伸。此外,附加地或可替代地,光学感测头包括多个发射器,该多个发射器被配置为产生覆盖面部的不同的、相应区域的相应光束组,并且处理电路被配置为选择和致动发射器的子集,而不致动所有发射器。

[0012] 在公开的实施例中,处理电路被配置为检测感测到的二次散斑图案的变化,并响应于检测到的变化来生成语音输出。

[0013] 可替代地或附加地,处理电路被配置为以第一帧速率操作感测器阵列,响应于当以第一帧速率操作时的信号来感测面部的运动,并且响应于感测到的运动将帧速率增加到大于第一帧速率的第二帧速率,以生成语音输出。

[0014] 在所公开的实施例中,处理电路被配置为响应于由于在用户不发出任何声音的情况下用户的皮肤表面的运动导致的、由光学感测头输出的信号的变化,生成语音输出。

[0015] 通常,光学感测头由支架保持在距离用户的皮肤表面至少5mm的位置。

[0016] 在一个实施例中,该设备包括一个或更多个电极,该一个或更多个电极被配置为接触用户的皮肤表面,其中处理电路被配置为响应于由一个或更多个电极感测到的电活动以及由光学感测头输出的信号来生成语音输出。

[0017] 附加地或可替代地,该设备包括麦克风,该麦克风被配置为感测用户发出的声音。在一个实施例中,处理电路被配置为将光学感测头输出的信号与麦克风感测到的声音进行比较,以便校准光学感测头。附加地或可替代地,处理电路被配置为响应于对用户发出的声音的感测来改变设备的操作状态。

[0018] 在一些实施例中,该设备包括通信接口,其中处理电路被配置为对信号进行编码,以便通过通信接口传输到处理设备,该处理设备处理经编码的信号以生成语音输出。在公开的实施例中,通信接口包括无线接口。

[0019] 附加地或可替代地,该设备包括用户控件,该用户控件连接到支架并被配置为感测用户做出的手势,其中处理电路被配置为响应于感测到的手势来改变设备的操作状态。

[0020] 此外,附加地或可替代地,该设备包括扬声器,该扬声器被配置为适配在用户的耳朵内,其中处理电路被配置为合成对应于语音输出的音频信号,以用于由扬声器回放。

[0021] 根据本发明的实施例,还提供了一种感测方法,该方法包括响应于人类对象说(articulate)单词但该对象不将单词发声出来,并且在不接触该对象的面部上的皮肤的情况下,感测该皮肤的运动。响应于感测到的运动,生成包括被说出的单词的语音输出。

[0022] 在一些实施例中,感测该运动包括感测从对象的面部反射的光。在公开的实施例中,感测该光包括将相干光引导到皮肤,并感测由于相干光从皮肤的反射而产生的二次散斑图案。在一个实施例中,引导相干光包括将相干光的多个光束引导到面部上的不同的相应位置,并使用感测器阵列感测从每个位置反射的二次散斑图案。

[0023] 在公开的实施例中,生成语音输出包括合成对应于语音输出的音频信号。可替代地或附加地,生成语音输出包括转录由对象说出的单词。

[0024] 根据本发明的实施例的以下详细描述并结合附图,本发明将得到更充分的理解,在附图中:

[0025] 附图简述

[0026] 图1是根据本发明的实施例的用于语音感测的系统的示意性形象化图示;

[0027] 图2是根据本发明的实施例的光学感测头的示意性剖视图;

[0028] 图3是根据本发明的另一实施例的语音感测设备的示意性形象化图示;

[0029] 图4是示意性地示出根据本发明的实施例的用于语音感测的系统的功能部件的框图;和

[0030] 图5是示意性地示出根据本发明的实施例的语音感测方法的流程图。

具体实施方式

[0031] 人们几乎随时随地通过他们的移动电话进行交流。移动电话在公共场所的广泛使用带来了不和谐的噪音,并经常引起隐私问题,因为对话很容易被路人听到。同时,当电话对话中的一方处于嘈杂的位置时,另一方或多方可能由于背景噪音而难以理解他们所听到的内容。文本交流为这些问题提供了一个解决方案,但是移动电话的文本输入很慢,并且干

扰了用户查看他们要去哪里的能力。

[0032] 本文描述的本发明的实施例使用无声语音来解决这些问题,使得用户能够说出单词和句子,而无需实际上将单词发声出来或根本无需发出任何声音。正常的发声过程使用多群肌肉和神经,从胸部和腹部开始,通过喉咙,并向上通过口腔和面部。为了说出给定的音素,运动神经元激活面部、喉部和口腔中的肌肉群,为推动气流流出肺部做准备,并且这些肌肉在说话过程中继续运动,以创造单词和句子。如果没有这种气流,嘴就不会发出声音。当没有来自肺部的气流,而面部、喉部和口腔中的肌肉继续说出想要的声音时,则会出现无声语音。

[0033] 无声语音可能是由于神经疾病和肌肉疾病引起的;但它也可能是有意发生的,例如当我们说单词但不希望被别人听到时。即使当我们在不张嘴的情况下把口语单词概念化时,这种说也会发生。由此产生的我们面部肌肉的激活引起了皮肤表面的细微运动。发明人已经发现,通过适当地感测和解码这些运动,有可能可靠地重建由用户说出的单词的实际序列。

[0034] 因此,本文描述的本发明的实施例感测对象面部上的皮肤和皮下神经及肌肉的细微运动,并且使用感测到的运动来生成包括被说出的单词的语音输出,该细微运动响应于由对象在发声或不发声的情况下说出的单词而发生。这些实施例提供了用于在不接触皮肤的情况下(例如通过感测从对象的面部反射的光)感测这些细微运动的方法和设备。因此,它们使用户能够以其他方基本上察觉不到的方式无声地与其他人交流或记录他们自己的想法。根据这些实施例的设备和方法也对环境噪声不敏感,并且可以基本上在任何环境中使用,而不需要用户将他们的视线和注意力从他们的周围事物上转移开。

[0035] 本发明的一些实施例提供了具有普通消费品形式的感测设备,例如夹式头戴式耳机(headphone)或眼镜。在这些实施例中,光学感测头通过适配在用户的耳朵内或耳朵之上的支架被保持在靠近用户面部的位置。例如通过将相干光引导到面部的区域(例如脸颊),光学感测头感测从面部反射的光,并且感测由于相干光从面部的反射而产生的二次散斑图案的变化。该设备中的处理电路处理光学感测头由于反射光而输出的信号,以生成相应的语音输出。

[0036] 可替代地,本发明的原理可以在没有耳夹或其他支架的情况下实现。例如,在替代实施例中,包括相干光源和传感器的无声语音感测模块可以集成到诸如智能手机之类的移动通信设备中。当用户将移动通信设备保持在靠近用户面部的合适位置时,该集成感测模块感测无声语音。

[0037] 在本说明书和权利要求中使用的术语“光”是指红外、可见光和紫外范围中的任何或所有范围的电磁辐射。

[0038] 图1是根据本发明的实施例的用于语音感测的系统18的示意性形象化图示。系统18基于感测设备20,其中耳夹22形式的支架适配在该设备的用户24的耳朵之上。附接到耳夹22的耳机26适配到用户的耳朵内。光学感测头28通过臂30连接到耳夹22,因此保持在靠近用户面部的位置。在图示的实施例中,设备20具有夹式头戴式耳机的形式和外观,其中光学感测头代替麦克风(或除了麦克风之外还有光学感测头)。

[0039] 光学感测头28将一束或更多束相干光导向用户24面部上的不同的相应位置,从而产生在面部的区域34上(且具体是在用户的脸颊上)延伸的光斑(spot)32的阵列。在本实施

例中,光学感测头28根本不接触用户的皮肤,而是保持在距离皮肤表面的一定距离处。通常,该距离至少为5mm,并且它甚至可以更大,例如距离皮肤表面至少1cm或者甚至2cm或者更大。为了能够感测面部肌肉的不同部分的运动,由光斑32覆盖并由光学感测头28感测的区域34通常具有至少1cm^{2}的范围;并且更大的区域(例如至少2cm^{2}或者甚至大于4cm^2)可以是有利的。

[0040] 光学感测头28感测从面部上的光斑32反射的相干光,并响应于检测到的光输出信号。具体地,光学感测头28感测由于相干光从其视场内的每个光斑32的反射而产生的二次散斑图案。为了覆盖足够大的区域34,该视场通常具有宽的角范围,通常具有至少60°、或者可能是70°、或者甚至90°或者更多的角宽度。在该视场内,设备20可以感测和处理由于所有光斑32或仅光斑32的某个子集的二次散斑图案而产生的信号。例如,设备20可以选择光斑的子集,该子集被发现在用户24的皮肤表面的相关运动方面给出最大量的有用且可靠的信息。下面参照图2描述光学感测头28的结构和操作的细节。

[0041] 在系统18内,处理电路处理由光学感测头28输出的信号以生成语音输出。如先前所述,即使用户22没有将语音发声出来或说出任何其他声音,处理电路也能够感测用户22的皮肤的运动并生成语音输出。语音输出可以采取合成的音频信号或文本转录或两者兼有的形式。合成的音频信号可以经由耳机26中的扬声器回放(并且在给予用户22关于语音输出的反馈时有用)。附加地或替代地,合成的音频信号可以通过网络传输,例如经由与移动通信设备(例如智能手机36)的通信链路传输。

[0042] 系统18中的处理电路的功能可以完全在设备20内执行,或者它们可以替代地在设备20和外部处理器之间分配,该外部处理器例如为运行合适的应用软件的智能手机36中的处理器。例如,设备20内的处理电路可以对由光学感测头28输出的信号进行数字化和编码,并通过通信链路将编码信号传输到智能手机36。该通信链路可以是有线或无线的,例如使用智能手机提供的蓝牙™无线接口。智能手机36中的处理器处理编码信号,以便生成语音输出。智能手机36还可以通过诸如互联网之类的数据网络来访问服务器38,以便例如上传数据和下载软件更新。下文参照图4描述处理电路的设计和操作的细节。

[0043] 在图示的实施例中,设备20还包括例如按钮(push-button)传感器或接近传感器形式的用户控件35,该用户控件35连接到耳夹22。用户控件35感测由用户执行的手势,例如在用户控件35上按压或以其他方式使用户的手指或手靠近用户控件。响应于适当的用户手势,处理电路改变设备20的操作状态。例如,用户24可以以这种方式将设备20从空闲模式切换到活动模式,并因此发信号指示(signal)设备应该开始感测和生成语音输出。这种切换在设备20中节省电池功率方面是有用的。可替代地或附加地,可以应用其他方法来控制设备20的操作状态并减少不必要的功耗,例如如下文参考图5所述。

[0044] 图2是设备20的光学感测头28的示意性剖视图,示出了根据本发明的实施例的光学感测头的部件和功能细节。光学感测头28包括发射器模块40和接收器模块48,以及可选的麦克风54。

[0045] 发射器模块40包括光源,例如红外激光二极管42,该光源发射相干辐射的输入光束。分束元件44,例如达曼光栅或另一种合适类型的衍射光学元件(DOE),将输入光束分成多个输出光束46,这些输出光束46在区域34上延伸的位置矩阵处形成相应的光斑32。在一个实施例中(未在图中示出),发射器模块40包括多个激光二极管或其他发射器,它们产生

输出光束46的相应组,这些组覆盖用户面部的区域34内的不同的相应子区域。在这种情况下,设备20中的处理电路可以仅选择和致动发射器的子集,而不致动所有发射器。例如,为了降低设备20的功耗,处理电路可以仅致动一个发射器或由两个或更多个发射器组成的子集,该一个发射器或该子集照射用户面部上的区域,该区域已被发现给出用于生成期望语音输出的最有用信息。

[0046] 接收器模块48包括光学传感器的阵列52,例如CMOS图像传感器,其中物镜50用于将区域34成像到阵列52上。由于光学感测头28的尺寸小以及其靠近皮肤表面,如上所述,接收器模块48具有足够宽的视场,并且以远离法线的高角度观察许多光斑32。由于皮肤表面粗糙,也可以以这些高角度检测到光斑32处的二次散斑图案。

[0047] 麦克风54感测用户24发出的声音,使得用户22能够在需要时将设备20用作传统头戴式耳机。附加地或可替代地,麦克风54可以与设备20的无声语音感测能力结合使用。例如,麦克风54可以在校准过程中使用,在校准过程中,当用户22说出某些音素或单词时,光学感测头28感测皮肤的运动。然后,处理电路可以将光学感测头28输出的信号与麦克风54感测到的声音进行比较,以便校准光学感测头。该校准可以包括提示用户22移动光学感测头28的位置,以便将光学部件对准在相对于用户脸颊的期望位置。

[0048] 在另一实施例中,由麦克风54输出的音频信号可用于改变设备20的操作状态。例如,仅当麦克风54没有检测到用户24对单词的发声时,处理电路才可以生成语音输出。由光学感测头28和麦克风54提供的光学感测和声学感测的组合的其他应用,对于本领域技术人员在阅读本说明书之后将是显而易见的,并且被认为在本发明的范围内。

[0049] 图3是根据本发明的另一实施例的语音感测设备60的示意性形象化图示。在该实施例中,耳夹22与眼镜框架62集成或以其他方式附接到眼镜框架62。鼻电极64和颞电极66附接到框架62并接触用户的皮肤表面。电极64和66接收体表肌电图(sEMG)信号,该信号提供关于用户的面部肌肉激活的附加信息。设备60中的处理电路使用由电极64和66感测到的电活动以及来自光学感测头28的输出信号来生成从设备60输出的语音。

[0050] 附加地或可替代地,设备60包括一个或更多个附加的光学感测头68,其类似于光学感测头28,用于感测在用户面部的其他区域中的皮肤运动。这些附加的光学感测头可以与光学感测头28一起使用或代替光学感测头28使用。

[0051] 图4是示意性地示出根据本发明的实施例的用于语音感测的系统18的功能部件的框图。图示的系统围绕图1所示的部件而构建,包括感测设备20、智能手机36和服务器38。可替代地,图4所示和下面描述的功能可以在该系统的部件之间不同地实现和分配。例如,归属于智能手机36的一些或所有处理能力可以在感测设备中实现;或者设备20的感测能力可以在智能手机36中实现。

[0052] 在图示的示例中,如上所述,感测设备20包括发射器模块40、接收器模块48、扬声器26、麦克风54和用户控件(UI)35。为了完整起见,感测设备20在图4中被示出为也包括其它传感器71,例如电极和/或环境传感器;但是如前所述,感测设备20能够仅基于由发射器和接收器模块进行的非接触式测量来操作。

[0053] 感测设备20包括编码器70和控制器75形式的处理电路。编码器70包括硬件处理逻辑和/或数字信号处理器,硬件处理逻辑可以是硬连线的或可编程的,数字信号处理器提取来自接收器模块48的输出信号的特征并对其进行编码。感测设备20经由诸如蓝牙接口之类

的通信接口72将编码信号传输到智能手机36中的相应通信接口77。电池74向感测设备20的部件提供操作电力。

[0054] 控制器75包括可编程的微控制器,例如,该微控制器基于从用户控件35、接收器模块48和智能手机36(经由通信接口72)接收的输入来设置感测设备20的操作状态和操作参数。下面参照图5描述此功能的一些方面。在替代实施例中,控制器75包括更强大的微处理器和/或处理阵列,其独立于智能手机36,在感测设备内本地处理来自接收器模块48的输出信号的特征并生成语音输出。

[0055] 然而,在本实施例中,来自感测设备20的经编码的输出信号被接收到智能手机36的存储器78中,并由在智能手机36中的处理器上运行的语音生成应用80处理。语音生成应用80将输出信号中的特征转换成文本和/或音频输出信号形式的单词序列。通信接口77将音频输出信号传递回感测设备20的扬声器26,以便回放给用户。来自语音生成应用80的文本和/或音频输出也被输入到其他应用84,例如话音和/或文本通信应用以及记录应用。通信应用例如经由数据通信接口86通过蜂窝或Wi-Fi网络进行通信。

[0056] 编码器70和语音生成应用80的操作由本地训练接口82控制。例如,接口82可以向编码器70指示从由接收器模块48输出的信号中提取哪些时间特征和频谱特征,并且可以向语音生成应用80提供神经网络的系数,神经网络将这些特征转换成单词。在本示例中,语音生成应用80实现推断网络,该推断网络查找与从感测设备20接收的经编码的信号特征相对应的、具有最高概率的单词序列。本地训练接口82从服务器38接收推断网络的系数,服务器38也可以周期性地更新系数。

[0057] 为了生成本地训练指令82,服务器38使用数据存储库88,该数据存储库88包含来自训练数据90的集合中的散斑图像和相应的基准真值(ground truth)口语单词。存储库88还接收在现场从感测设备20收集到的训练数据。例如,训练数据可以包括当用户说某些声音和单词(可能包括无声语音和有声语音)时从感测设备20收集到的信号。一般训练数据90与从每个感测设备20的用户接收的个人训练数据的这种组合使得服务器38能够针对每个用户导出最佳的推断网络系数。

[0058] 服务器38应用图像分析工具94来从存储库88中的散斑图像中提取特征。这些图像特征与相应的单词字典104和语言模型100一起作为训练数据被输入到神经网络96,语言模型100定义了训练数据中使用的特定语言的语音学结构(phonetic structure)和句法规则。神经网络96生成用于推断网络102的最佳系数,推断网络102将从散斑测量值的相应序列中提取出的特征集的输入序列转换成相应的音素,并最终转换成单词的输出序列。网络架构和训练过程的进一步细节在上述的临时专利申请中进行了描述。服务器38将推断网络102的系数下载到智能手机36,以在语音生成应用80中使用。

[0059] 图5是示意性地示出根据本发明的实施例的用于语音感测的方法的流程图。为了方便和清楚起见,参照如图1和图4所示且上面描述的系统18的元件来描述该方法。可替代地,该方法的原理可以在其他系统配置中应用,例如使用感测设备60(图3)或集成在移动通信设备中的感测设备的系统配置。

[0060] 在空闲步骤110,只要用户24不说话,感测设备20就在低功率空闲模式下操作,以便节省电池74中的电力。在这种模式下,控制器75以低帧速率(例如20帧/秒)驱动接收器模块48中的传感器的阵列52。发射器模块40也可以以降低的输出功率来操作。在运动检测步

骤112,当接收器模块48以这种低帧速率操作时,控制器75处理阵列52输出的图像,以便检测指示语音的面部运动。在活动捕获步骤114,当检测到这种运动时,控制器75指示接收机模块48以及感测设备20的其他部件将帧速率增加到例如100-200帧/秒的范围,以便能够检测到由于无声语音而发生的二次散斑图案的变化。可替代地或附加地,控制器75可以响应于其他输入,例如用户控件35的致动或从智能手机36接收的指令,来增加帧速率并给感测设备20的其他部件通电。

[0061] 由接收器模块48捕获的图像通常包含所投射的激光光斑32的矩阵,如图1所示。在光斑检测116,编码器70检测图像中的光斑的位置。编码器可以从所有光斑中提取特征;但是为了节省功率和处理资源,希望编码器选择光斑的子集。例如,本地训练接口82可以指示哪个光斑子集包含关于用户语音的最大量的信息,并且编码器70可以选择该子集中的光斑。在裁剪步骤118,编码器70从每个图像裁剪出小窗口,其中每个这样的窗口包含所选择的光斑之一。

[0062] 在特征提取步骤120,编码器70从每个选择的光斑中提取散斑运动的特征。例如,编码器70可以基于相应窗口中的像素的平均强度来估计每个散斑中的总能量,并且可以测量每个散斑的能量随着时间推移的变化。附加地或可替代地,编码器70可以提取所选择的光斑子集中的散斑的其他时间特征和/或频谱特征。编码器70将这些特征传送到语音生成应用80(运行在智能手机36上),在特征输入步骤122,语音生成应用80将特征值的向量输入到从服务器38下载的推断网络102。

[0063] 在语音输出步骤124,基于随着时间的推移而输入到推断网络的特征向量序列,语音生成应用80输出单词的流,这些单词被拼接在一起成为句子。如先前所述,语音输出被用于合成音频信号,用于经由扬声器26回放。在后处理步骤126,在智能手机36上运行的其他应用84对语音和/或音频信号进行后处理,以记录相应的文本和/或通过网络传输语音或文本数据。

[0064] 应当理解,上述实施例是通过示例的方式引用的,并且本发明不限于已经在上文具体示出和描述的内容。更确切地说,本发明的范围包括上文所描述的各种特征的组合和子组合,以及本领域技术人员在阅读前述描述后会想到的并且在现有技术中未被公开的这些特征的变型和修改。

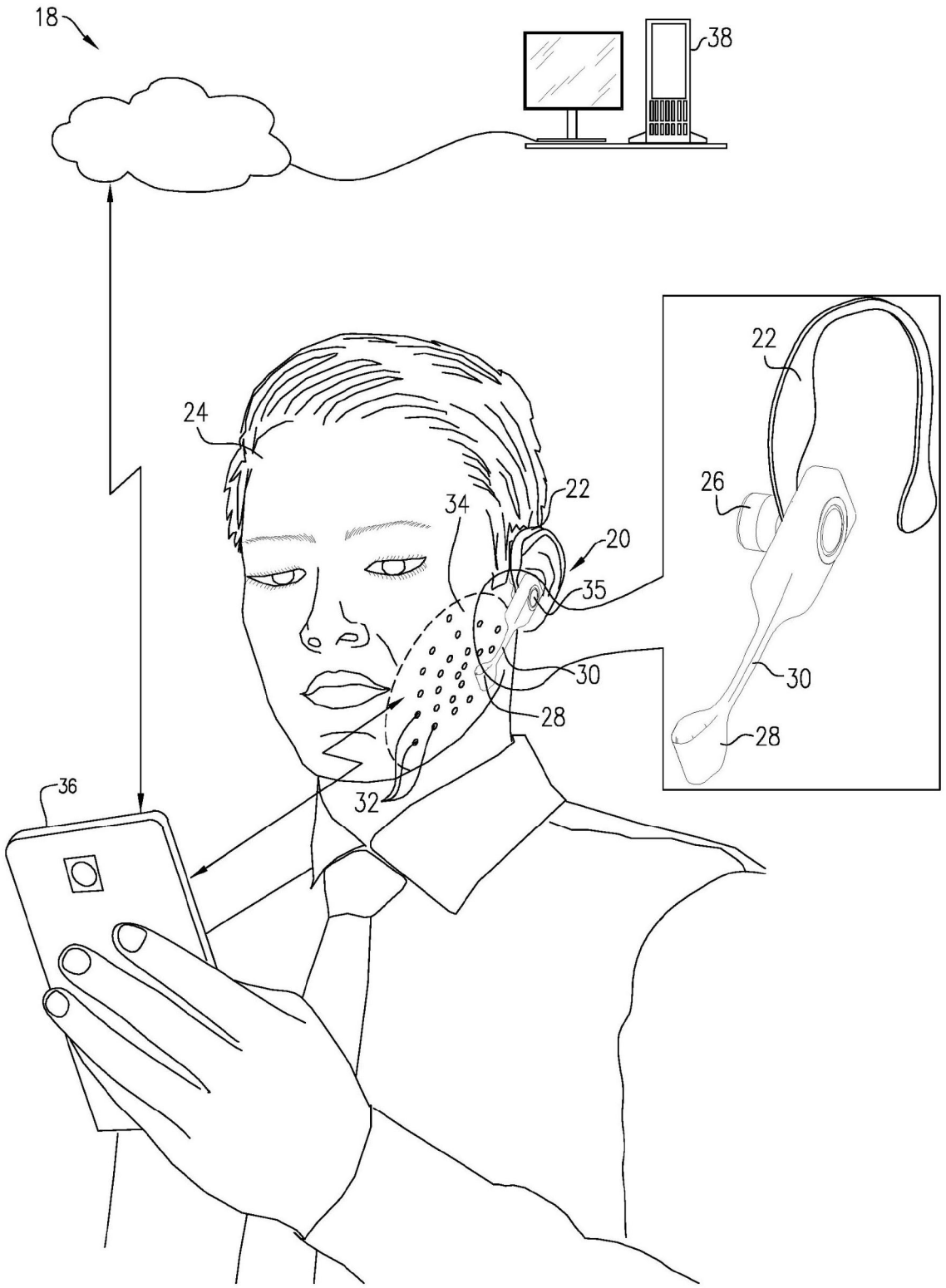


图1

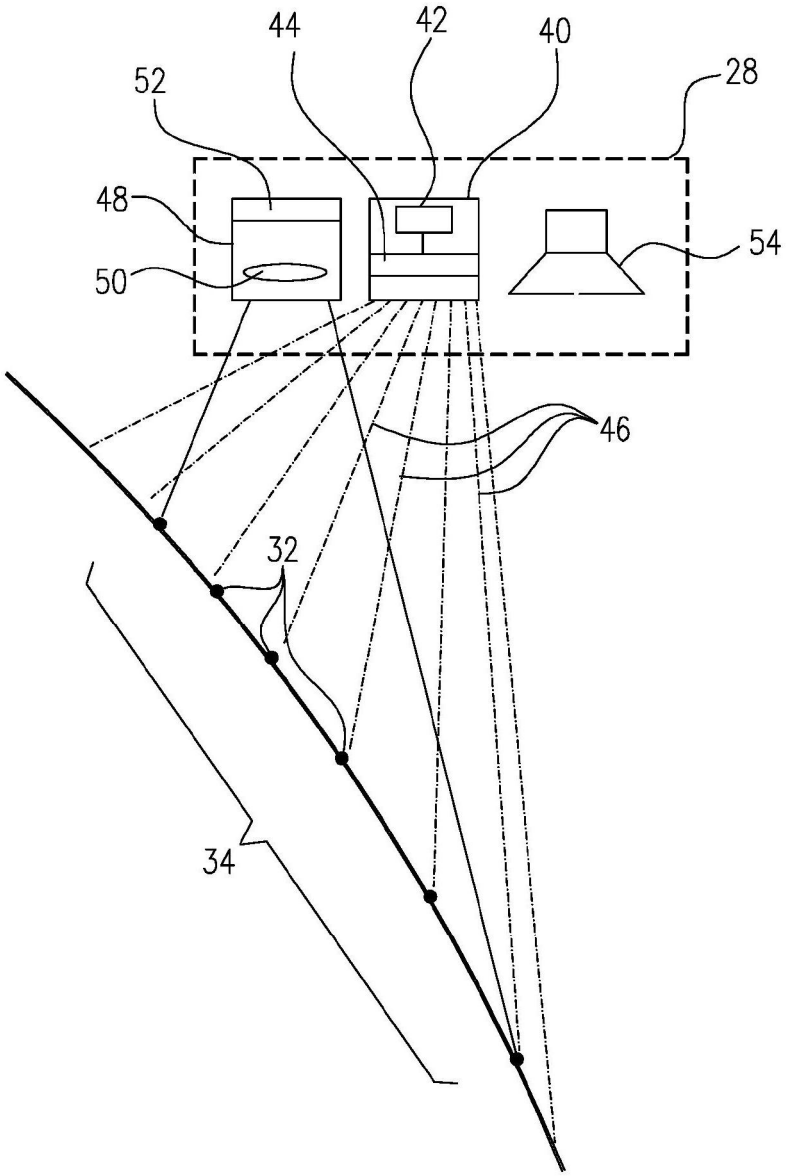


图2

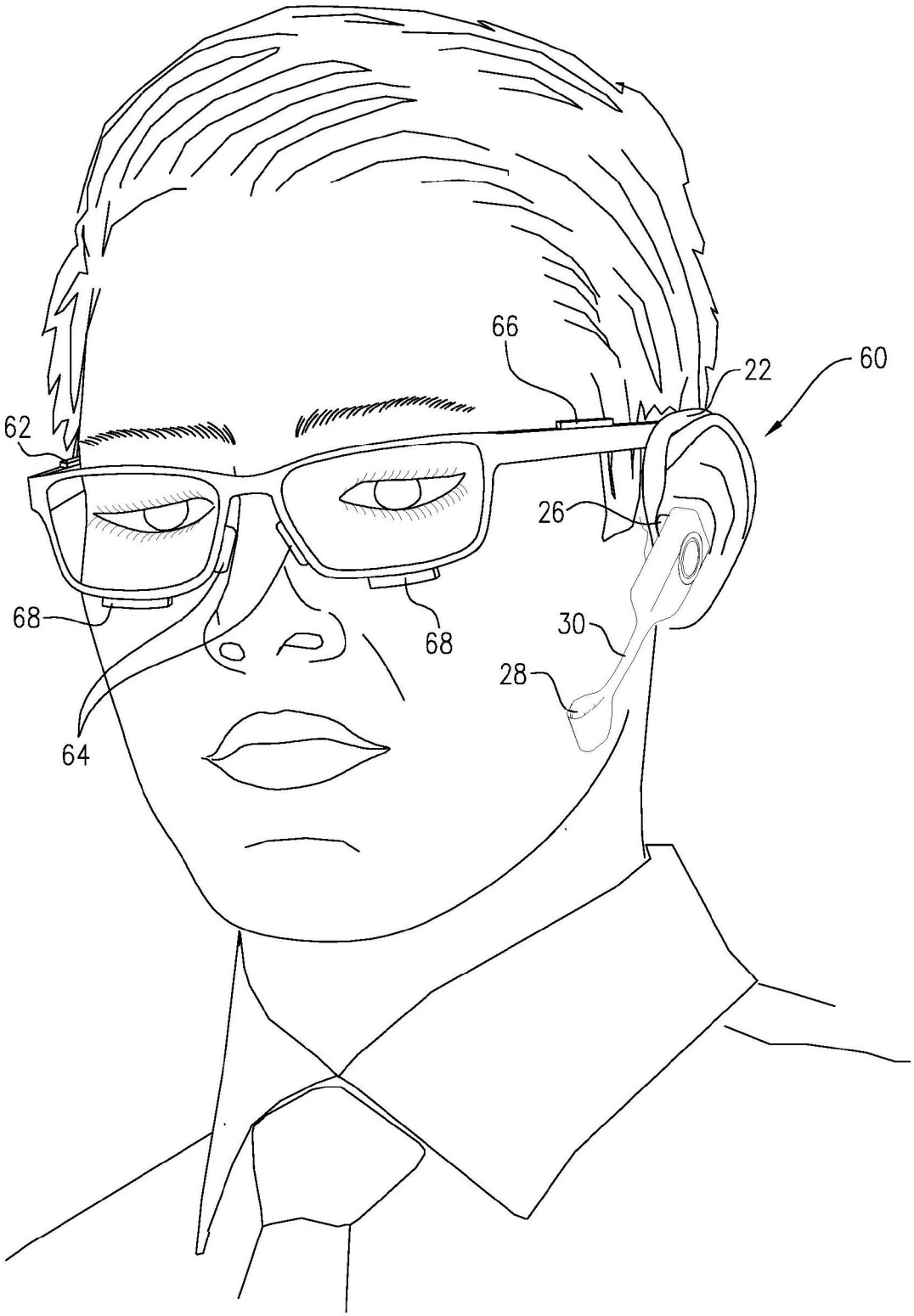


图3

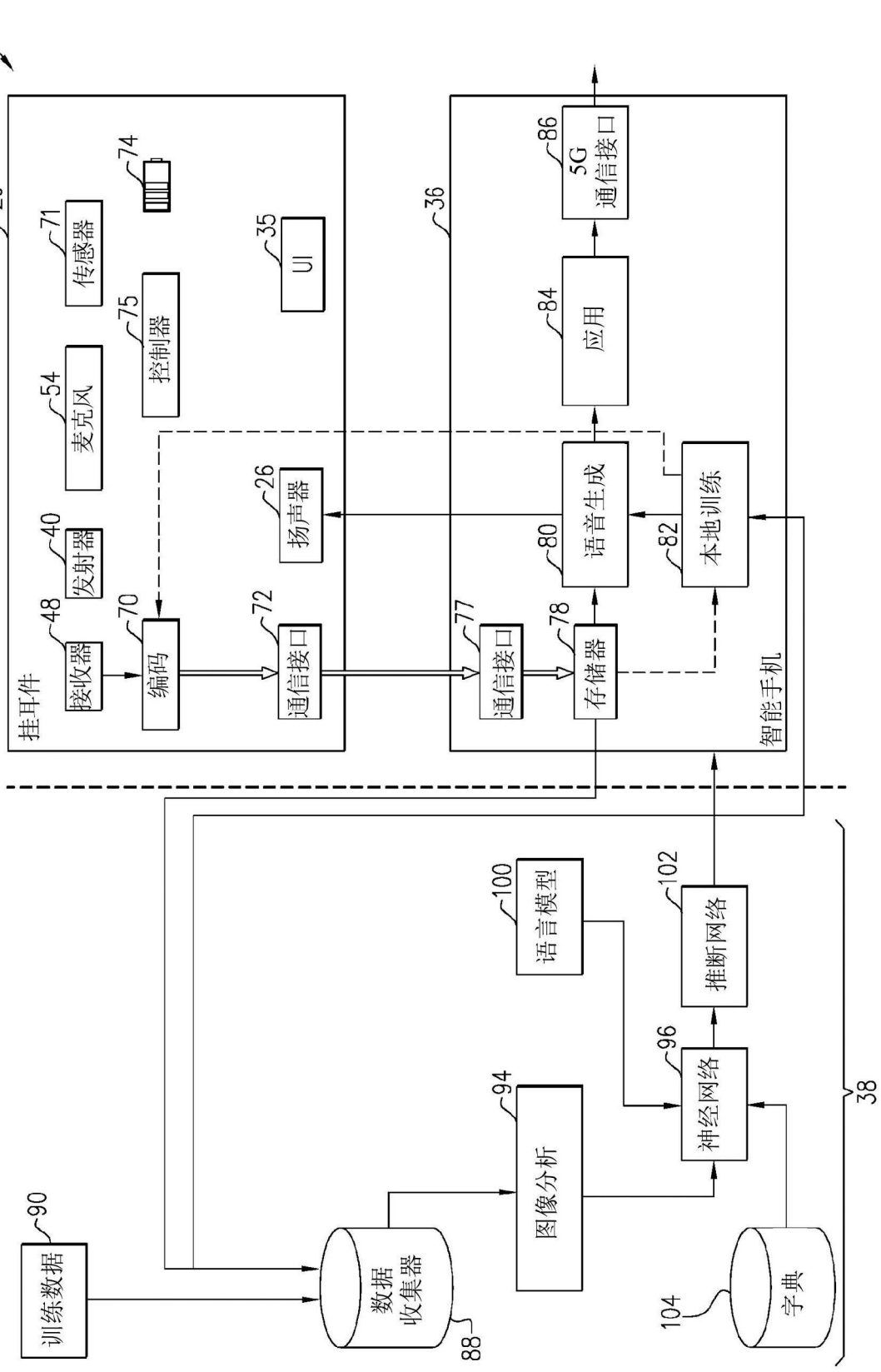

18
20
挂耳件
74
71
传感器
35
UI
75
控制器
54
麦克风
26
扬声器
40
发射器
48
接收器
70
编码
72
通信接口
36
智能手机
86
5G
通信接口
84
应用
80
语音生成
82
本地训练
77
通信接口
78
存储器
90
训练数据
数据
收集器
88
94
图像分析
100
语言模型
96
神经网络
102
推断网络
104
字典
38

图4

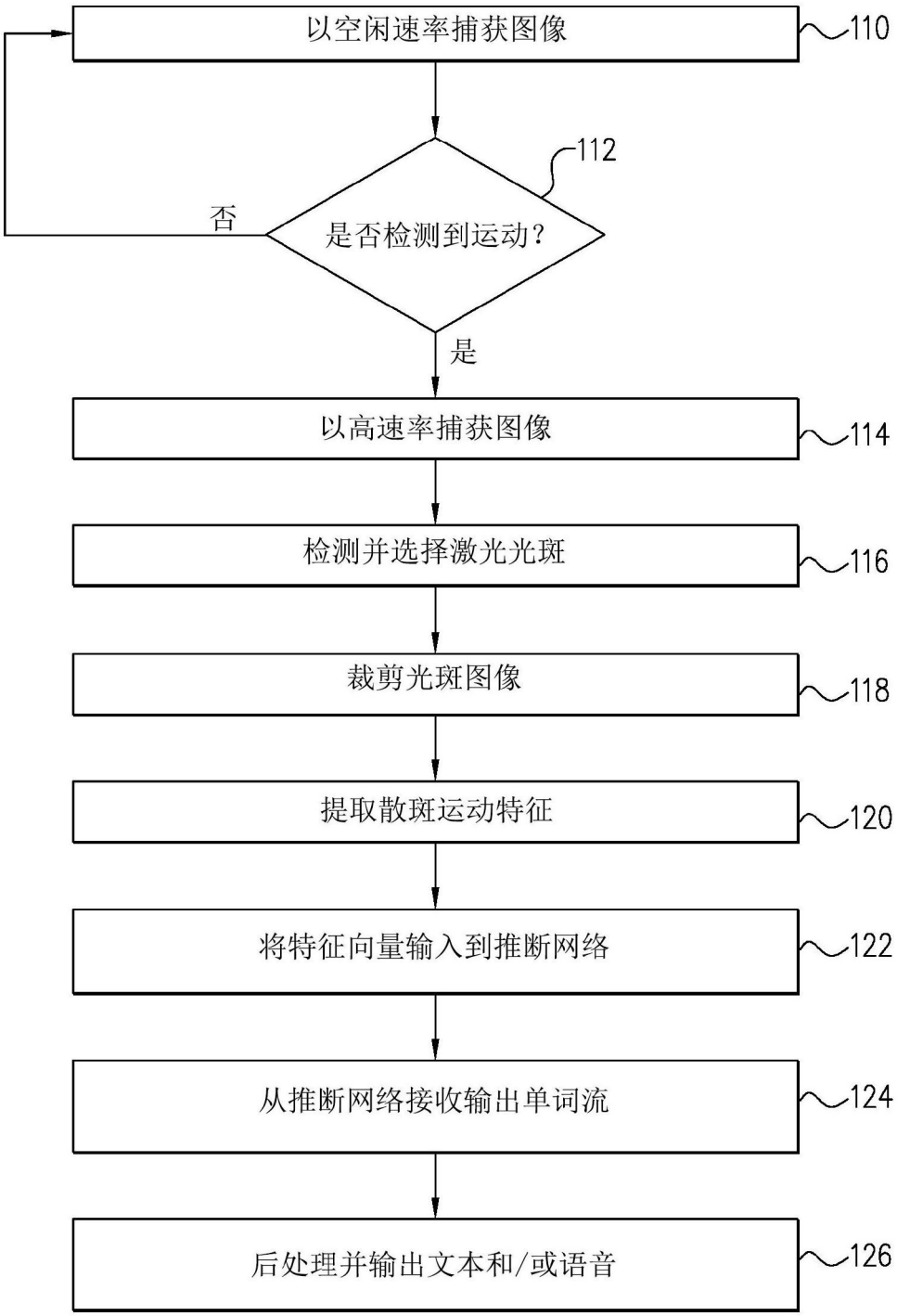

以空闲速率捕获图像
110
112
是否检测到运动？
否
是
以高速率捕获图像
114
检测并选择激光光斑
116
裁剪光斑图像
118
提取散斑运动特征
120
将特征向量输入到推断网络
122
从推断网络接收输出单词流
124
后处理并输出文本和/或语音
126

图5