



(51) International Patent Classification:

G10L 21/02 (2013.01) H04R 3/00 (2006.01)
G10L 17/22 (2013.01)

(21) International Application Number:

PCT/US2019/038546

(22) International Filing Date:

21 June 2019 (21.06.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/687,987 21 June 2018 (21.06.2018) US

(71) Applicant: **MAGIC LEAP, INC.** [US/US]; 7500 W. Sunrise Boulevard, Plantation, Florida 33322 (US).

(72) Inventor: **LEIDER, Colby Nelson**; c/o Magic Leap, Inc., 7500 W. Sunrise Boulevard, Plantation, Florida 33322 (US).

(74) Agent: **BLUMENKOPF, Bryan**; Morrison & Foerster LLP, 707 Wilshire Boulevard, Suite 6000, Los Angeles, California 90017-3543 (US).

(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, IR, IS, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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: WEARABLE SYSTEM SPEECH PROCESSING

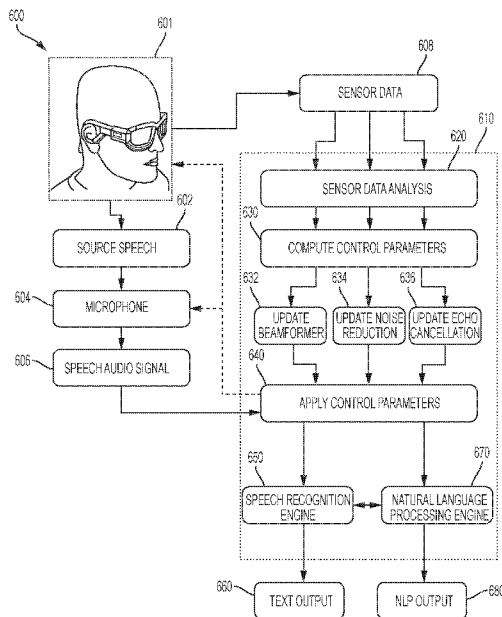


FIG. 6

(57) Abstract: A method of processing an acoustic signal is disclosed. According to one or more embodiments, a first acoustic signal is received via a first microphone. The first acoustic signal is associated with a first speech of a user of a wearable headgear unit. A first sensor input is received via a sensor, a control parameter is determined based on the sensor input. The control parameter is applied to one or more of the first acoustic signal, the wearable headgear unit, and the first microphone. Determining the control parameter comprises determining, based on the first sensor input, a relationship between the first speech and the first acoustic signal.



WEARABLE SYSTEM SPEECH PROCESSING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit of U.S. Provisional Patent Application No. 62/687,987, filed on June 21, 2018, which is hereby incorporated by reference in its entirety.

FIELD

[0002] This disclosure relates in general to systems and methods for processing acoustic speech signals, and in particular to systems and methods for processing acoustic speech signals generated by a user of a wearable device.

BACKGROUND

[0003] Tools for speech processing are tasked with receiving audio input representing human speech, typically via a microphone, and processing the audio input to determine words, logical structures, or other outputs corresponding to that audio input. For example, automatic speech recognition (ASR) tools may generate a text output based on the human speech corresponding to an audio input signal; and natural language processing (NLP) tools may generate logical structures, or computer data, corresponding to the meaning of that human speech. It is desirable for such processes to happen accurately and quickly — with some applications demanding results in real-time.

[0004] Computer speech processing systems have a history of producing dubious results. In general, the accuracy of a speech processing system can be highly dependent on the quality of the input audio signal, with the highest accuracy obtained from inputs provided under controlled conditions. For example, speech processing systems may perform admirably when the audio input is clearly enunciated speech captured by a microphone at a direct angle, and at close range, with no environmental noise, a high signal/noise ratio, and a constant volume level. However, speech processing systems may struggle to accommodate the many variables that can be

introduced into input audio signals by real-world conditions. For example, speech processing signals may demonstrate limited accuracy when multiple speech sources are present (e.g., multiple people speaking at once in the same space); when environmental noise (e.g., wind, rain, electrical interference, ambient noise) blends with a source speech signal; when the human speaker does not enunciate clearly, or speaks with an idiosyncratic or inconsistent tone, accent, or inflection; when the speaker moves or rotates relative to a microphone; when the speaker is in an acoustically reflective environment (e.g., a tiled bathroom or a large cathedral); when the speaker is a large distance from a microphone; when the speaker faces away from the microphone; or when any number of other variables are present and compromise the fidelity of the input audio signal. These problems may be magnified in mobile or outdoor applications, in which unpredictable sources of noise may be present; and in which attempts to control or understand the speaker's vicinity may be made difficult or impossible.

[0005] It is desirable to use sensor-equipped wearable systems, such as those that incorporate a head-mounted unit, to correct for the effects of such variables on input audio signals for speech processing systems. By presenting speech processing systems with more predictable, higher fidelity input, the outputs of those systems can generate more accurate and more reliable results. In addition, wearable systems are well suited for mobile, outdoor applications — precisely the type of applications in which many conventional speech processing systems may perform especially poorly.

BRIEF SUMMARY

[0006] Examples of the disclosure describe systems and methods for processing acoustic signals. According to one or more embodiments, a first acoustic signal is received via a first microphone. The first acoustic signal is associated with a first speech of a user of a wearable headgear unit. A first sensor input is received via a sensor. a control parameter is determined based on the sensor input. The control parameter is applied to one or more of the first acoustic signal, the wearable

headgear unit, and the first microphone. Determining the control parameter comprises determining, based on the first sensor input, a relationship between the first speech and the first acoustic signal.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] FIG. 1 illustrates an example wearable head device that can be used as part of a wearable system, according to some embodiments.

[0008] FIG. 2 illustrates an example handheld controller that can be used as part of a wearable system, according to some embodiments.

[0009] FIG. 3 illustrates an example auxiliary unit that can be used as part of a wearable system, according to some embodiments.

[0010] FIG. 4 illustrates an example functional block diagram for an example wearable system, according to some embodiments.

[0011] FIG. 5 illustrates a flow chart of an example speech processing system, according to some embodiments.

[0012] FIG. 6 illustrates a flow chart of an example system for processing acoustic speech signals, according to some embodiments.

DETAILED DESCRIPTION

[0013] In the following description of examples, reference is made to the accompanying drawings which form a part hereof, and in which it is shown by way of illustration specific examples that can be practiced. It is to be understood that other examples can be used and structural changes can be made without departing from the scope of the disclosed examples.

[0014] EXAMPLE WEARABLE SYSTEM

[0015] FIG. 1 illustrates an example wearable head device 100 configured to be worn on the head of a user. Wearable head device 100 may be part of a broader wearable system that comprises one or more components, such as a head device (e.g., wearable head device 100), a handheld controller (e.g., handheld controller 200 described below), and/or an auxiliary unit (e.g., auxiliary unit 300 described below). In some examples, wearable head device 100 can be used for virtual reality, augmented reality, or mixed reality systems or applications. Wearable head device 100 can comprise one or more displays, such as displays 110A and 110B (which may comprise left and right transmissive displays, and associated components for coupling light from the displays to the user's eyes, such as orthogonal pupil expansion (OPE) grating sets 112A/112B and exit pupil expansion (EPE) grating sets 114A/114B); left and right acoustic structures, such as speakers 120A and 120B (which may be mounted on temple arms 122A and 122B, and positioned adjacent to the user's left and right ears, respectively); one or more sensors such as infrared sensors, accelerometers, GPS units, inertial measurement units (IMU)(e.g. IMU 126), acoustic sensors (e.g., microphone 150); orthogonal coil electromagnetic receivers (e.g., receiver 127 shown mounted to the left temple arm 122A); left and right cameras (e.g., depth (time-of-flight) cameras 130A and 130B) oriented away from the user; and left and right eye cameras oriented toward the user (e.g., for detecting the user's eye movements)(e.g., eye cameras 128 and 128B). However, wearable head device 100 can incorporate any suitable display technology, and any suitable number, type, or combination of sensors or other components without departing from the scope of the invention. In some examples, wearable head device 100 may incorporate one or more microphones 150 configured to detect audio signals generated by the user's voice; such microphones may be positioned in a wearable head device adjacent to the user's mouth. In some examples, wearable head device 100 may incorporate networking features (e.g., Wi-Fi capability) to communicate with other devices and systems, including other wearable systems. Wearable head device 100 may further include components such as a battery, a processor, a memory, a storage unit, or various input devices (e.g., buttons, touchpads); or may be coupled to a handheld controller (e.g., handheld controller 200) or an auxiliary unit (e.g.,

auxiliary unit 300) that comprises one or more such components. In some examples, sensors may be configured to output a set of coordinates of the head-mounted unit relative to the user's environment, and may provide input to a processor performing a Simultaneous Localization and Mapping (SLAM) procedure and/or a visual odometry algorithm. In some examples, wearable head device 100 may be coupled to a handheld controller 200, and/or an auxiliary unit 300, as described further below.

[0016] FIG. 2 illustrates an example mobile handheld controller component 200 of an example wearable system. In some examples, handheld controller 200 may be in wired or wireless communication with wearable head device 100 and/or auxiliary unit 300 described below. In some examples, handheld controller 200 includes a handle portion 220 to be held by a user, and one or more buttons 240 disposed along a top surface 210. In some examples, handheld controller 200 may be configured for use as an optical tracking target; for example, a sensor (e.g., a camera or other optical sensor) of wearable head device 100 can be configured to detect a position and/or orientation of handheld controller 200 — which may, by extension, indicate a position and/or orientation of the hand of a user holding handheld controller 200. In some examples, handheld controller 200 may include a processor, a memory, a storage unit, a display, or one or more input devices, such as described above. In some examples, handheld controller 200 includes one or more sensors (e.g., any of the sensors or tracking components described above with respect to wearable head device 100). In some examples, sensors can detect a position or orientation of handheld controller 200 relative to wearable head device 100 or to another component of a wearable system. In some examples, sensors may be positioned in handle portion 220 of handheld controller 200, and/or may be mechanically coupled to the handheld controller. Handheld controller 200 can be configured to provide one or more output signals, corresponding, for example, to a pressed state of the buttons 240; or a position, orientation, and/or motion of the handheld controller 200 (e.g., via an IMU). Such output signals may be used as input to a processor of wearable head device 100, to auxiliary unit 300, or to another component of a wearable system. In some examples, handheld controller 200 can include one or more microphones to detect sounds (e.g., a user's speech, environmental sounds),

and in some cases provide a signal corresponding to the detected sound to a processor (e.g., a processor of wearable head device 100).

[0017] FIG. 3 illustrates an example auxiliary unit 300 of an example wearable system. In some examples, auxiliary unit 300 may be in wired or wireless communication with wearable head device 100 and/or handheld controller 200. The auxiliary unit 300 can include a battery to provide energy to operate one or more components of a wearable system, such as wearable head device 100 and/or handheld controller 200 (including displays, sensors, acoustic structures, processors, microphones, and/or other components of wearable head device 100 or handheld controller 200). In some examples, auxiliary unit 300 may include a processor, a memory, a storage unit, a display, one or more input devices, and/or one or more sensors, such as described above. In some examples, auxiliary unit 300 includes a clip 310 for attaching the auxiliary unit to a user (e.g., a belt worn by the user). An advantage of using auxiliary unit 300 to house one or more components of a wearable system is that doing so may allow large or heavy components to be carried on a user's waist, chest, or back — which are relatively well suited to support large and heavy objects — rather than mounted to the user's head (e.g., if housed in wearable head device 100) or carried by the user's hand (e.g., if housed in handheld controller 200). This may be particularly advantageous for relatively heavy or bulky components, such as batteries.

[0018] FIG. 4 shows an example functional block diagram that may correspond to an example wearable system 400, such as may include example wearable head device 100, handheld controller 200, and auxiliary unit 300 described above. In some examples, the wearable system 400 could be used for virtual reality, augmented reality, or mixed reality applications. As shown in FIG. 4, wearable system 400 can include example handheld controller 400B, referred to here as a “totem” (and which may correspond to handheld controller 200 described above); the handheld controller 400B can include a totem-to-headgear six degree of freedom (6DOF) totem subsystem 404A. Wearable system 400 can also include example wearable head device 400A (which may correspond to wearable headgear device 100 described above); the wearable head device 400A includes a totem-to-headgear 6DOF headgear subsystem 404B. In the example, the

6DOF totem subsystem 404A and the 6DOF headgear subsystem 404B cooperate to determine six coordinates (e.g., offsets in three translation directions and rotation along three axes) of the handheld controller 400B relative to the wearable head device 400A. The six degrees of freedom may be expressed relative to a coordinate system of the wearable head device 400A. The three translation offsets may be expressed as X, Y, and Z offsets in such a coordinate system, as a translation matrix, or as some other representation. The rotation degrees of freedom may be expressed as sequence of yaw, pitch and roll rotations; as vectors; as a rotation matrix; as a quaternion; or as some other representation. In some examples, one or more depth cameras 444 (and/or one or more non-depth cameras) included in the wearable head device 400A; and/or one or more optical targets (e.g., buttons 240 of handheld controller 200 as described above, or dedicated optical targets included in the handheld controller) can be used for 6DOF tracking. In some examples, the handheld controller 400B can include a camera, as described above; and the headgear 400A can include an optical target for optical tracking in conjunction with the camera. In some examples, the wearable head device 400A and the handheld controller 400B each include a set of three orthogonally oriented solenoids which are used to wirelessly send and receive three distinguishable signals. By measuring the relative magnitude of the three distinguishable signals received in each of the coils used for receiving, the 6DOF of the handheld controller 400B relative to the wearable head device 400A may be determined. In some examples, 6DOF totem subsystem 404A can include an Inertial Measurement Unit (IMU) that is useful to provide improved accuracy and/or more timely information on rapid movements of the handheld controller 400B.

[0019] In some examples involving augmented reality or mixed reality applications, it may be desirable to transform coordinates from a local coordinate space (e.g., a coordinate space fixed relative to wearable head device 400A) to an inertial coordinate space, or to an environmental coordinate space. For instance, such transformations may be necessary for a display of wearable head device 400A to present a virtual object at an expected position and orientation relative to the real environment (e.g., a virtual person sitting in a real chair, facing forward, regardless of the position and orientation of wearable head device 400A), rather than at a fixed position and

orientation on the display (e.g., at the same position in the display of wearable head device 400A). This can maintain an illusion that the virtual object exists in the real environment (and does not, for example, appear positioned unnaturally in the real environment as the wearable head device 400A shifts and rotates). In some examples, a compensatory transformation between coordinate spaces can be determined by processing imagery from the depth cameras 444 (e.g., using a Simultaneous Localization and Mapping (SLAM) and/or visual odometry procedure) in order to determine the transformation of the wearable head device 400A relative to an inertial or environmental coordinate system. In the example shown in FIG. 4, the depth cameras 444 can be coupled to a SLAM/visual odometry block 406 and can provide imagery to block 406. The SLAM/visual odometry block 406 implementation can include a processor configured to process this imagery and determine a position and orientation of the user's head, which can then be used to identify a transformation between a head coordinate space and a real coordinate space. Similarly, in some examples, an additional source of information on the user's head pose and location is obtained from an IMU 409 of wearable head device 400A. Information from the IMU 409 can be integrated with information from the SLAM/visual odometry block 406 to provide improved accuracy and/or more timely information on rapid adjustments of the user's head pose and position.

[0020] In some examples, the depth cameras 444 can supply 3D imagery to a hand gesture tracker 411, which may be implemented in a processor of wearable head device 400A. The hand gesture tracker 411 can identify a user's hand gestures, for example by matching 3D imagery received from the depth cameras 444 to stored patterns representing hand gestures. Other suitable techniques of identifying a user's hand gestures will be apparent.

[0021] In some examples, one or more processors 416 may be configured to receive data from headgear subsystem 404B, the IMU 409, the SLAM/visual odometry block 406, depth cameras 444, a microphone 450; and/or the hand gesture tracker 411. The processor 416 can also send and receive control signals from the 6DOF totem system 404A. The processor 416 may be coupled to the 6DOF totem system 404A wirelessly, such as in examples where the handheld

controller 400B is untethered. Processor 416 may further communicate with additional components, such as an audio-visual content memory 418, a Graphical Processing Unit (GPU) 420, and/or a Digital Signal Processor (DSP) audio spatializer 422. The DSP audio spatializer 422 may be coupled to a Head Related Transfer Function (HRTF) memory 425. The GPU 420 can include a left channel output coupled to the left source of imagewise modulated light 424 and a right channel output coupled to the right source of imagewise modulated light 426. GPU 420 can output stereoscopic image data to the sources of imagewise modulated light 424, 426. The DSP audio spatializer 422 can output audio to a left speaker 412 and/or a right speaker 414. The DSP audio spatializer 422 can receive input from processor 419 indicating a direction vector from a user to a virtual sound source (which may be moved by the user, e.g., via the handheld controller 400B). Based on the direction vector, the DSP audio spatializer 422 can determine a corresponding HRTF (e.g., by accessing a HRTF, or by interpolating multiple HRTFs). The DSP audio spatializer 422 can then apply the determined HRTF to an audio signal, such as an audio signal corresponding to a virtual sound generated by a virtual object. This can enhance the believability and realism of the virtual sound, by incorporating the relative position and orientation of the user relative to the virtual sound in the mixed reality environment — that is, by presenting a virtual sound that matches a user's expectations of what that virtual sound would sound like if it were a real sound in a real environment.

[0022] In some examples, such as shown in FIG. 4, one or more of processor 416, GPU 420, DSP audio spatializer 422, HRTF memory 425, and audio/visual content memory 418 may be included in an auxiliary unit 400C (which may correspond to auxiliary unit 300 described above). The auxiliary unit 400C may include a battery 427 to power its components and/or to supply power to wearable head device 400A and/or handheld controller 400B. Including such components in an auxiliary unit, which can be mounted to a user's waist, can limit the size and weight of wearable head device 400A, which can in turn reduce fatigue of a user's head and neck.

[0023] While FIG. 4 presents elements corresponding to various components of an example wearable system 400, various other suitable arrangements of these components will become apparent to those skilled in the art. For example, elements presented in FIG. 4 as being associated with auxiliary unit 400C could instead be associated with wearable head device 400A or handheld controller 400B. Furthermore, some wearable systems may forgo entirely a handheld controller 400B or auxiliary unit 400C. Such changes and modifications are to be understood as being included within the scope of the disclosed examples.

[0024] SPEECH PROCESSING ENGINES

[0025] Speech processing systems in general include systems that accept an input audio signal corresponding to human speech (a source signal); process and analyze the input audio signal; and produce, as a result of the analysis, an output corresponding to the human speech. A process or module for performing these tasks may be considered a speech processing engine. In the case of automatic speech recognition systems, the output of a speech processing engine may be a text transcription of the human speech. In the case of natural language processing systems, the output may be one or more commands or instructions indicated by the human speech; or some representation (e.g., a logical expression or a data structure) of the semantic meaning of the human speech. Other types of speech processing systems (e.g., automatic translation systems) are contemplated and are within the scope of the disclosure.

[0026] Speech processing systems are found in a diverse array of products and applications: conventional telephone systems; automated voice messaging systems; voice assistants (including standalone and smartphone-based voice assistants); vehicles and aircraft; desktop and document processing software; data entry; home appliances; medical devices; language translation software; closed captioning systems; and others. An advantage of speech processing systems is that they may allow users to provide input to a computer system using natural spoken language, such as presented to a microphone, instead of conventional computer input devices such as keyboards or touch panels; accordingly, speech processing systems may be particularly useful in

environments where conventional input devices (e.g., keyboards) may be unavailable or impractical. Further, by permitting users to provide intuitive voice-based input, speech recognition systems can heighten feelings of immersion. As such, speech recognition can be a natural fit for wearable systems, and in particular, for virtual reality, augmented reality, and/or mixed reality applications of wearable systems, in which user immersion is a primary goal, and in which it may be desirable to limit the use of conventional computer input devices, whose presence may detract from feelings of immersion.

[0027] FIG. 5 shows an automatic speech recognition engine 500, according to some embodiments. Engine 500 is intended to be illustrative of automatic speech recognition systems in general; other specific systems are possible and within the scope of the disclosure. Engine 500 may be implemented using one or more processors (e.g., CPUs, GPUs, and/or DSPs), a memory, an input device (e.g., a microphone), an output device (e.g., a display, a speaker), a network, a database, and/or other suitable components. In engine 500, an audio signal 510, corresponding to a source human speech signal, is presented to a signal preprocessing stage 520. In some examples, signal preprocessing stage 520 can apply one or more signal processing functions to audio signal 510. For example, preprocessing functions can include audio processing functions, such as peak compression; noise reduction; bandlimiting; equalization; signal attenuation; or other suitable functions. These preprocessing functions can simplify the task of later processing and analyzing audio signal 510. For instance, feature extraction algorithms can be calibrated to perform best on input signals having certain audio characteristics, such as gain and frequency characteristics, and where a signal/noise ratio of the input signal is maximized. In some examples, preprocessing audio signal 510 can condition the signal such that the signal can be more reliably analyzed elsewhere in engine 500. For instance, signal preprocessing stage 520 can re-encode audio signal 510 (e.g., re-encoding at a specific bitrate); or convert audio signal 510 from a first form (e.g., a time-domain signal) into a second form (e.g., a frequency-domain signal; or a parametric representation) that may simplify the later processing of audio signal 510. In some examples, one or more functions of preprocessing stage 520 may be performed by a DSP.

[0028] At stage 530, a feature extraction process can be applied to audio signal 510 (as preprocessed at stage 520). A goal of feature extraction is to identify individual speech features of audio signal 510, and reduce or eliminate variations in these features so that the features can be effectively and consistently processed (e.g., compared against stored patterns in a database). For example, feature extraction can reduce, eliminate, or control variations in the speaker's pitch, gender, accent, pronunciation, and pacing. Feature extraction can also reduce, eliminate, or control variations in recording equipment (e.g., microphone type), signal transmission (e.g., via land-based telephone lines or cellular telephone networks), or recording environment (e.g., room acoustics; background noise levels; speaker's distance from microphone; speaker's angle to microphone). Various suitable techniques for feature extraction are known in the art.

[0029] Speech features extracted from audio signal 510 at stage 530 can be presented to a decoder stage 540. The goal of decoder stage 540 is to determine a text output 570 that corresponds to the source human speech from which audio signal 510 was generated. In some examples, text output 570 need not be text, but can be another data representation of the source speech. Various techniques exist for decoding speech features into text; for example, hidden Markov models; Viterbi decoding; beam search; dynamic search; multipass search; weighted finite state transducers (WFST); or any suitable combination of the above. Other suitable techniques will be familiar to those skilled in the art.

[0030] In some examples, decoder 540 can make use of an acoustic modeling stage 550 to facilitate generating text output 570. Acoustic modeling stage 550 can use a model of the relationships between speech signals and linguistic units (e.g., phonemes) to identify one or more linguistic units from an audio signal 510 (including one or more features extracted at stage 530). Various suitable acoustic modeling techniques that can be applied at acoustic modeling stage 550 will be familiar to those skilled in the art.

[0031] In some examples, decoder 540 can make use of a language modeling stage 560 to facilitate generating text output 570. Linguistic modeling stage 560 can use a model of a

language's grammar, vocabulary, and other characteristics to determine linguistic units (e.g., phonemes) that most probably correspond to features of audio signal 510. For example, a linguistic model applied by stage 560 could conclude that a particular extracted feature is more likely to correspond to a word that appears with a high frequency in the speaker's language than to a word that appears with a low frequency in that language. Various suitable linguistic modeling techniques that can be applied at linguistic modeling stage 560 will be familiar to those skilled in the art. In addition, decoder 540 may make use of other suitable techniques or models to facilitate the generation of text output 570; the disclosure is not limited to any particular technique or group of techniques.

[0032] Typically, text output 570 does not correspond to the source human speech with perfect certainty; instead, the likelihood that text output 570 correctly corresponds to the source human speech can be expressed as some probability or confidence interval. Because of the many variables that can affect audio signal 510, even sophisticated speech recognition systems do not consistently produce perfect text output for all speakers. For example, the reliability of speech recognition systems such as engine 500 may be highly dependent on the quality of the input audio signal 510. Where audio signal 510 is recorded in ideal conditions — for example, in acoustically controlled environments, with the human speaker enunciating clearly and directly into a microphone from a close distance — the source speech can be more readily determined from the audio signal. For example, features can be more reliably extracted from audio signal 510 at stage 530; and decoder 540 can more effectively determine which text output 570 corresponds to those features (e.g., acoustic modeling can be more reliably applied to the features at stage 550, and/or linguistic modeling can be more reliably applied to the features at stage 560).

[0033] In real-world applications, however, audio signal 510 may deviate from ideal conditions, such that determining the source human speech becomes more difficult. For example, audio signal 510 may incorporate environmental noise, such as may be introduced by outdoor environments, or by substantial distances between a human speaker and a microphone; electrical

noise, such as from electrical interference (e.g., battery chargers for a smartphone); natural reverberations, such as from nearby surfaces (e.g., concrete, bathroom tile) or acoustic spaces (e.g., caverns, cathedrals); or other unwanted effects. In addition, audio signal 510 may suffer from attenuation of certain frequencies, such as may result when a human speaker faces away from a microphone; this is particularly problematic when the frequencies attenuated are those that carry significant speech-related information (e.g., formant frequencies that can be used to distinguish between vowel sounds). Similarly, audio signal 510 may suffer from overall low amplitude, or low signal/noise ratio, which may be caused when a large distance sits between a human speaker and a microphone. In addition, if the human speaker moves and reorients himself or herself while speaking, the audio signal 510 may change characteristics over the course of the signal, further complicating efforts to determine the underlying speech.

[0034] Although example system 500 illustrates an example speech recognition engine, other types of speech processing engines may follow an analogous structure: for example, a natural language processing engine, upon accepting an input audio signal corresponding to human speech, may perform a signal preprocessing stage; extract components from the signal (e.g., via segmentation and/or tokenization stages); and perform detection/analysis of the components, in some cases with the assistance of one or more linguistic modeling subsystems. Moreover, in some examples, the output of an automatic speech recognition engine, such as shown in example system 500, may be used as input to a further language processing engine; for instance, a natural language processing engine could accept as input the text output 570 of example system 500. Such systems may experience challenges analogous to those faced by example system 500. For example, variations in an input audio signal that make it more difficult to recover a source speech signal underlying the audio signal, such as described above, may also make it more difficult to provide other forms of output (e.g., logical expressions or data structures in the case of natural language processing engines). Accordingly, such systems also stand to benefit from the present invention, such as described below.

[0035] IMPROVING SPEECH PROCESSING USING WEARABLE SYSTEMS

[0036] The present disclosure is directed to systems and methods for improving the accuracy of speech processing systems by using input from sensors, for example those associated with wearable devices (e.g., head-mounted devices such as described above with respect to FIG. 1) to reduce, eliminate, or control variations in input audio signals, such as those described above with respect to audio signal 510. Such variations may be especially pronounced in mobile applications of speech processing, or in applications of speech processing in uncontrolled environments, such as outdoor environments. Wearable systems are frequently intended for use in such applications, and may experience such variations. The wearable system may generally refer to any combination of a head device (e.g., wearable head device 100), a handheld controller (e.g., handheld controller 200), an auxiliary unit (e.g., auxiliary unit 300), and/or an environment of the head device. In some embodiments, the sensors of the wearable system may be on the head device, the handheld controller, the auxiliary unit, and/or in the environment of the head device. For example, because wearable systems may be designed to be mobile, audio signals recorded at a single stationary microphone (e.g., at a standalone voice assistant device) from a user of a wearable system may suffer from a low signal/noise ratio if the user is far from the single stationary microphone, or from “acoustic shadows” or unwanted frequency response if the user faces away from the microphone. Moreover, the audio signal may change characteristics over time as the user moves and turns with respect to the single stationary microphone, as may be expected by a mobile user of a wearable system. In addition, because some wearable systems are intended for use in uncontrolled environments, a high potential exists for environmental noise (or speech of other humans) to be recorded along with the target human speech. Similarly, such uncontrolled environments can introduce unwanted echoes and reverberations into audio signal 510 that can obscure the underlying speech.

[0037] As described above with respect to example wearable head device 100 in FIG. 1, a wearable system can include one or more sensors that can provide input about the user and/or the environment of the wearable system. For instance, wearable head device 100 can include a camera (e.g., camera 444 described in FIG. 4) to output visual signals corresponding to the environment; in some examples, the camera can be a forward-facing camera on a head-mounted

unit that shows what is currently in front of the user of the wearable system. In some examples, wearable head device 100 can include a LIDAR unit, a radar unit, and/or acoustic sensors, which can output signals corresponding to the physical geometry (e.g., walls, physical objects) of the user's environment. In some examples, wearable head device 100 can include a GPS unit, which can indicate geographic coordinates corresponding to the wearable system's current location. In some examples, wearable head device 100 can include an accelerometer, a gyroscope; and/or an inertial measurement unit (IMU) to indicate an orientation of the wearable head device 100. In some examples, wearable head device 100 can include environmental sensors, such as temperature or pressure sensors. In some examples, wearable head device 100 can include biometric sensors, such as iris cameras; fingerprint sensors; eye tracking sensors; or sensors to measure a user's vital signs. In examples where wearable head device 100 includes a head-mounted unit, such orientation can correspond to an orientation of the user's head (and, by extension, the user's mouth and a direction of the user's speech). Other suitable sensors can be included. In some embodiments, the handheld controller 200, the auxiliary unit 300, and/or the environment of the wearable head device 100 can include any suitable one or more of the sensors mentioned above with respect to the wearable head device 100. In addition, in some cases, one or more sensors may be placed in an environment with which the wearable system interacts. For instance, a wearable system may be designed to be worn by a driver of an automobile, and appropriate sensors (e.g., depth cameras, accelerometers, etc.) may be placed in the interior of the automobile. One advantage of this approach is that the sensors can occupy a known location within the environment; compared to sensors that can be attached to a wearable device that moves within the environment, this configuration can simplify the interpretation of data provided by those sensors.

[0038] The signals provided by such sensors of the wearable system (e.g., the wearable head device 100, the handheld controller 200, the auxiliary unit 300, and/or the environment of the wearable head device 100) can be used to provide information about the characteristics of an audio signal recorded by the wearable system, and/or information about the relationship between

the audio signal and an underlying source speech signal. This information can in turn be used to more effectively determine source speech underlying that audio signal.

[0039] To illustrate, FIG. 6 shows an example speech recognition system 600 that incorporates a wearable system to improve speech recognition of an audio signal recorded by one or more microphones. FIG. 6 shows a user of a wearable system 601, which may correspond to wearable system 400 described above, and which may include one or more of example wearable head device 100; handheld controller 200; and auxiliary unit 300. The user of wearable system 601 provides a verbal utterance 602 (the “source speech”) which is detected at one or more microphones 604, which output corresponding audio signal 606. Wearable system 601 can include one or more sensors described above, including one or more of a camera; a LIDAR unit, a radar unit; an acoustic sensor; a GPS unit; an accelerometer; a gyroscope; an IMU; a microphone (which may be one of microphones 604); a temperature sensor; biometric sensors; or any other suitable sensor or combination of sensors. Wearable system 601 can also include one or more processors (e.g., CPUs, GPUs, and/or DSPs), a memory, an input device (e.g., a microphone), an output device (e.g., a display, a speaker), a network, and/or a database. These components may be implemented using a combination of a wearable head device 100, a handheld controller 200, and an auxiliary unit 300. In some examples, sensors of wearable system 601 provide sensor data 608 (which may be multichannel sensor data such as in examples where one or more sensors, including sensors from two or more sensor types, are presented in parallel). Sensor data 608 may be provided concurrently with microphones 604 detecting the source speech; that is, the sensors of wearable system 601 can provide sensor data 608 corresponding to conditions at the time the source speech is provided. In some examples, one or more of microphones 604 can be included in a wearable head device, a handheld controller, and/or an auxiliary unit of wearable system 601 and/or in an environment thereof.

[0040] In example system 600, subsystem 610 can accept as input sensor data 608 and audio signal 606; determine control parameters to apply to process audio signal 606; and provide the processed audio signal as input to a signal processing engine (e.g., speech recognition engine 650

and/or natural language processing (NLP) engine 670), which may generate an output (e.g., text output 660 and/or natural language processing output 680, respectively). Subsystem 610 includes one or more processes, stages, and/or modules described below and shown in FIG. 6. Subsystem 610 can be implemented using one or more processors (e.g., CPUs, GPUs, and/or DSPs), a memory, a network, a database, and/or any combination of other suitable components. In some examples, some or all of subsystem 610 can be implemented on wearable system 601: for example, on one or more of a head device (e.g., wearable head device 100); a handheld controller (e.g., handheld controller 200); and/or an auxiliary unit (e.g., auxiliary unit 300). In some examples, some or all of subsystem 610 can be implemented on a device containing microphone 604 (e.g., a smartphone, or a standalone voice assistant device). In some examples, some or all of subsystem 610 can be implemented on a cloud server or another network-enabled computing device. For instance, a local device could perform latency-sensitive functions of subsystem 610, such as those related to processing audio signals and/or sensor data; while a cloud server or other network-enabled device could perform functions of subsystem 610 that require large computational or memory resources (e.g., training or applying a complex artificial neural network), and transmit the output to the local device. Other implementations of subsystem 610 will be apparent to those skilled in the art and are within the scope of the disclosure.

[0041] In example system 600, subsystem 610 includes a sensor data analysis stage 620. Sensor data analysis stage 620 can process and analyze sensor data 608, for example in order to determine information about the environment of wearable system 601. In examples where sensor data 608 includes sensor data from disparate sources (e.g., camera data and GPS data), sensor data analysis stage 620 can combine the sensor data (“sensor fusion”) according to techniques known to those in the art (e.g., Kalman filters). In some examples, sensor data analysis stage 620 can incorporate data from other sources in addition to sensor data 608. For instance, sensor data analysis stage 620 can combine sensor data 608 from a GPS unit with map data and/or satellite data (e.g., from a memory or database storing such data) in order to determine location information based on the output of the GPS unit. As an example, the GPS

unit could output, as part of sensor data 608, GPS coordinates corresponding to the latitude and longitude of wearable system 601. Sensor data analysis stage 620 could use the latitude and longitude with map data to identify in what country, town, street, unit (e.g., commercial unit or residential unit), or room the wearable system 601 is located; or to identify nearby businesses or points of interest. Similarly, architectural data (e.g., from public structural records) can be combined with the sensor data 608 to identify a building in which wearable system 601 is located; or weather data (e.g., from a real-time feed of satellite data) can be combined with the sensor data 608 to identify current weather conditions at that location. Other example applications will be apparent and are within the scope of the disclosure.

[0042] On a smaller scale, sensor data analysis stage 620 can analyze sensor data 608 to generate information relating to objects and geometry in the immediate vicinity of wearable system 601, or relating to a user of wearable system 601. For example, using techniques known in the art, data from a LIDAR sensor or radar unit of wearable system 601 can indicate that wearable system 601 is facing a wall located eight feet away, at an angle θ to a normal of that wall; and image data from a camera of wearable system 601 can identify that wall as likely being made of ceramic tile (an acoustically reflective material). In some examples, an acoustic sensor of wearable system 601 can be used to measure acoustic effects that a surface may have on an acoustic signal (e.g., by comparing a reflected signal from the surface to a source signal transmitted to the surface). In some examples, sensor data analysis stage 620 can use sensor data 608 to determine a position and/or orientation of the user of wearable system 601, for example using an accelerometer, gyroscope, or IMU associated with wearable system 601. In some examples, such as in augmented reality or mixed reality applications, stage 620 can incorporate a map or other representation of the user's current environment; for example, in cases where sensors of wearable system 601 have been used to construct a 3D representation of the geometry of a room, that 3D representation data can be used in conjunction with sensor data 608. Similarly, stage 620 can incorporate information such as the materials of nearby surfaces, and acoustic properties of those surfaces; information relating to other users in the environment (e.g., their locations and orientations, and/or the acoustic characteristics of their voices); and/or

information about the user of wearable system 601 (e.g., the user's age group, gender, native language, and/or vocal characteristics).

[0043] In some examples, sensor data analysis stage 620 can analyze sensor data 608 to generate information relating to microphones 604. For example, sensor data 608 may provide a position and/or orientation of one or more microphones 604, such as the position and/or orientation relative to wearable system 601. In examples where wearable system 601 includes one or more microphones 604, the position and orientation of one or more microphones 604 may be directly linked to the position and orientation of the wearable system 601. In some examples, wearable system 601 can include one or more additional microphones that are not one or more microphones 604; such additional one or more microphones can be used to provide a baseline audio signal, for example corresponding to a user's speech as detected from a known position and orientation, and from a short distance. For example, the additional one or more microphones may be in known positions with known orientations in an environment of the user. The amplitude, phase, and frequency characteristics of this baseline audio signal can be compared to audio signal 606 to identify a relationship between the source speech 602 and audio signal 606. For example, if a first audio signal detected at a first time has an amplitude one-half that of the baseline audio signal, and a second audio signal detected at a second time has an amplitude one-fourth that of the baseline audio signal, it may be inferred that the user has moved away from microphone 604 during the interval between the first time and the second time. This can be extended to any suitable number of microphones (e.g., an initial microphone, and two or more additional microphones).

[0044] Information output by sensor data analysis stage 620, based on sensor data 608 and/or other data as described above, can identify a relationship between source speech 602 and corresponding audio signal 606. Information describing this relationship can be used at stage 630 to compute one or more control parameters that can be applied to audio signal 606, to one or more microphones 604, and/or to wearable system 601. Application of these control parameters can improve the accuracy with which system 600 can recover the underlying source speech 602.

from audio signal 606. In some examples, control parameters computed at stage 630 can include digital signal processing (DSP) parameters for that can be applied to process audio signal 606. For instance, such control parameters can include parameters for a digital signal processing (DSP) noise reduction process (e.g., a signal threshold below which gated noise reduction will be applied to audio signal 606, or a noisy frequency of audio signal 606 to be attenuated); parameters for a DSP echo cancellation or dereverberation process (e.g., a time value corresponding to a delay between audio signal 606 and an echo of that signal); or parameters for other audio DSP processes (e.g., phase correction, limiting, pitch correction).

[0045] In some examples, control parameters can define a DSP filter to be applied to audio signal 606. For instance, sensor data 608 (e.g., from a microphone of a head-mounted unit of wearable system 601) can indicate a characteristic frequency curve corresponding to the voice of the user of the wearable system 601 (i.e., the user generating source speech 602). This frequency curve can be used to determine control parameters defining a digital bandpass filter to apply to audio signal 606; this bandpass filter can isolate frequencies corresponding closely to source speech 602, in order to make source speech 602 more prominent in audio signal 606. In some examples, sensor data 608 (e.g., from a microphone of a head-mounted unit of wearable system 601) can indicate a characteristic frequency curve corresponding to the voice of a different user (other than the user of the wearable system 601) in the vicinity of the wearable system 601. This frequency curve can be used to determine control parameters defining a digital notch filter to apply to audio signal 606; this notch filter can remove unwanted voices from audio signal 606, in order to render source speech 602 more prominent in audio signal 606. Similarly, sensor data 608 (e.g., a camera of wearable system 601) can identify specific other individuals in the vicinity, and their positions relative to the wearable system 601; this information can determine a level of the notch filter (e.g., the closer an individual is, the louder his or her voice is likely to be in audio signal 606, and the greater a level attenuation that may need to be applied). As another example, the presence of certain surfaces and/or materials in the vicinity of a user may affect the frequency characteristics of that user's voice as detected by microphone 604; for instance, if the user is standing in the corner of a room, certain low frequencies of the user's voice may be

accentuated in audio signal 606; this information can be used to generate parameters (e.g., a cutoff frequency) of a high-pass filter to be applied to audio signal 606. These control parameters can be applied to audio signal 606 at stage 640; or as part of speech recognition engine 650 and/or natural language processing engine 670 (e.g., at a stage corresponding to signal preprocessing stage 520 or feature extraction stage 530 described above).

[0046] In some examples, control parameters computed at stage 630 can be used to configure microphone 604. For example, such control parameters can include hardware configuration parameters, such as gain levels for a hardware amplifier coupled to microphone 604; beamforming parameters to adjust a directionality of microphone 604 (e.g., a vector at which microphone 602 should be directed); parameters to determine which of a plurality of microphones 604 should be enabled or disabled; or parameters to control where a microphone 604 should be positioned or oriented (e.g., in examples where microphone 604 is attached to a mobile platform). In some examples, microphone 604 may be a component of a smartphone, or another mobile device; and the control parameters computed at stage 630 can be used to control the mobile device (e.g., to enable various components of the mobile device, or to configure or operate software on the mobile device).

[0047] In some examples, control parameters computed at stage 630 can be used to control wearable system 601 itself. For example, such control parameters could include parameters for presenting a message to a user of wearable system 601, such as via displays 110A/110B or speakers 120A/120B (e.g., an audio or video message that the user should move away from a nearby wall in order to improve speech recognition accuracy); or parameters to enable, disable, or reconfigure one or more sensors of wearable system 601 (e.g., to reorient a servo-mounted camera in order to obtain more useful camera data). In examples where the wearable system 601 includes microphones 604, control parameters can be sent to wearable system 601 in order to control microphones 604 such as described above.

[0048] In some examples, control parameters computed at stage 630 can be used to affect a decoding process (e.g., decoding process 540 described above) of a speech processing system (e.g., speech recognition engine 650 and/or natural language processing engine). For example, sensor data 608 may indicate characteristics of a user's environment, behaviors, or mental state that may affect the user's use of language. For instance, sensor data 608 (e.g., from a camera and/or a GPS unit) may indicate that the user is attending a football game; because the user's speech (i.e., source speech 602) may be much more likely than usual to include football-related words (e.g., "coach," "quarterback," "touchdown") while the user is attending a football game, a control parameter of a speech processing system (e.g., language modeling stage 560) can be temporarily set to reflect a higher probability that an audio signal corresponds to a football-related word.

[0049] In some examples, such as shown in FIG. 6, individual update modules 632, 634, and 636 can determine control parameters computed at stage 630, or can apply the control parameters computed at stage 630, and/or sensor data 608, to audio signal 606; microphone 604; wearable system 601; or to any hardware or software subsystem of the above. For example, a beamformer update module 632 can determine, based on sensor data 608 or one or more control parameters computed at stage 630, how a beamforming module (e.g., a beamforming module of microphone 604) can be updated to improve the recognition or natural language processing of audio signal 606. In some examples, beamforming update module 632 can control the directivity of a sensor array (e.g., an array of microphones 604) in order to maximize a signal/noise ratio of a signal detected by the sensor array. For instance, the beamforming update module 632 can adjust the directivity of microphones 604 such that source speech 602 is detected with a minimum of noise and distortion; for example, in a room with multiple voices, software of an adaptive beamforming module can direct microphones 604 to maximize the signal power corresponding to a voice of interest (e.g., the voice corresponding to source speech 602). For instance, sensors of wearable system 601 can output data indicating that wearable system 601 is located a certain distance from microphones 604, from which a time-of-flight value of the source speech 602 to microphones 604 can be determined. This time-of-flight value can be used to calibrate the

beamforming, using techniques familiar to those skilled in the art, in order to maximize the ability of speech recognition engine 650 to identify source speech 602 from audio signal 606.

[0050] In some examples, a noise reduction update module 634 can determine, based on sensor data 608 or one or more control parameters computed at stage 630, how a noise reduction process can be updated to improve the recognition or natural language processing of audio signal 606. In some examples, noise reduction update module 634 can control parameters of a noise reduction process applied to audio signal 606 in order to maximize a signal/noise ratio of audio signal 606; this in turn can facilitate automatic speech recognition performed by speech recognition engine 650. For instance, noise reduction update module 634 can selectively apply signal attenuation to frequencies of audio signal 606 on which noise is likely to be present, while boosting (or declining to attenuate) frequencies of audio signal 606 which carry information of the source speech 602. Sensors of wearable system 601 can provide data to help noise reduction update module 634 identify which frequencies of audio signal 606 are likely to correspond to noise, and which frequencies are likely to carry information about the source speech 602. For instance, sensors of wearable system 601 (e.g., GPS, LIDAR, etc.) may identify that wearable system 601 is located on an airplane. An airplane may be associated with background noise having certain characteristic frequencies; for example, airplane engine noise may be concentrated primarily around a known frequency f_0 . Based on this information from the sensors, noise reduction update module 634 can attenuate frequency f_0 of audio signal 606. Similarly, sensors of wearable system 601 (e.g., a microphone mounted to wearable system 601) can identify a frequency signature corresponding to the voice of a user of wearable system 601; noise reduction update module 634 can apply a bandpass filter to a frequency range corresponding to that frequency signature, or can ensure that noise reduction is not applied to that frequency range.

[0051] In some examples, an echo cancellation (or dereverberation) update module 636 can determine, based on sensor data 608 or one or more control parameters computed at stage 630, how an echo cancellation unit can be updated to improve the recognition or natural language processing of audio signal 606. In some examples, echo cancellation update module 636 can

control parameters of an echo cancellation unit applied to audio signal 606 in order to maximize the ability of a speech recognition engine 650 to determine source speech 602 from audio signal 606. For instance, echo cancellation update module 636 can instruct an echo cancellation unit to detect and correct for (e.g., via a comb filter) echoes in audio signal 606 that follow the source speech by 100 milliseconds. Because such echoes can interfere with the ability of speech processing systems (e.g., speech recognition engine 650, natural language processing engine 670) to determine source speech (e.g., by affecting the ability to extract features, such as described above with respect to FIG. 5 at stage 530, from audio signal 606), removing these echoes can result in higher accuracy of speech recognition, natural language processing, and other speech processing tasks. In some examples, sensors of wearable system 601 can provide sensor data 608 that can be used to determine control parameters for echo cancellation. For example, such sensors (e.g., camera, LIDAR, radar, acoustic sensors) can determine that a user of wearable system 601 is located ten feet from a surface, and faces that surface at an angle θ_1 to a normal of that surface, and further that microphone 604 is located twenty feet from that surface, and faces the surface at an angle θ_2 . It can be computed from this sensor data that the surface is likely to produce an echo that reaches microphone 604 at a certain time after the source signal 602 (i.e., the time-of-flight from the user to the surface plus the time-of-flight from the surface to microphone 604). Similarly, it can be determined from the sensor data that the surface corresponds to a bathroom tile surface, or another surface with known acoustic reflective properties; and those reflective properties can be used to compute control parameters of a dereverberation unit that will attenuate the resultant acoustic reflections in audio signal 606.

[0052] Similarly, in some examples, signal conditioning can be applied to speech audio signal 606 to account for equalization applied to the speech audio signal 606 by an acoustic environment. For instance, a room can increase or attenuate certain frequencies of speech audio signal 606, for example by virtue of the room's geometry (e.g., dimensions, cubic volume), materials (e.g., concrete, bathroom tile), or other characteristics that can affect the signal as it is detected by a microphone (e.g., microphone 604) in the acoustic environment. These effects can complicate the ability of speech recognition engine 650 to perform consistently across different

acoustic environments. Sensors of wearable system 601 can provide sensor data 608 which can be used to counteract such effects. For instance, sensor data 608 can indicate, as described above, a size or shape of a room, or the presence of acoustically significant materials, from which one or more filters can be determined, and applied to the audio signal to counteract room effects. In some examples, sensor data 608 can be provided by sensors (e.g., camera, LIDAR, radar, acoustic sensors) that indicate a cubic volume of a room in which the user is present; an acoustic effect of that room can be modeled as a filter; and an inversion of that filter can be applied to an audio signal to compensate for that acoustic effect.

[0053] In some examples, modules 632, 634, or 636, or other suitable elements of example system 600 (e.g., stage 620 or stage 640) can make use of predetermined mappings to determine control parameters based on sensor data 608. In some examples described above, control parameters are computed directly based on sensor data 608. For instance, as described above, an echo cancellation update module 636 can determine control parameters for an echo cancellation unit that can be applied to attenuate echoes present in audio signal 606. As described above, such control parameters can be computed by geometrically determining the distance between wearable system 601 and a microphone 604; computing the time-of-flight that an audio signal traveling at the speed of sound in air would require to travel from the wearable system 601 to the microphone; and setting an echo period of the echo cancellation unit to correspond to that time-of-flight. In some cases, though, the control parameters can be determined by comparing sensor data 608 to a mapping of sensor data to control parameters. Such a mapping could be stored in a database, such as on a cloud server or another networked device. In some examples, one or more elements of example system 601 (e.g., beamformer update module 632, noise reduction update module 634, echo cancellation update module 636, stage 630, stage 640, speech recognition engine 650, and/or natural language processing engine 670) can query a database to retrieve one or more control parameters that correspond to sensor data 608. This process can occur instead of, or in addition to, direct computation of control parameters such as described above. In such a process, sensor data 608 could be provided to the predetermined mapping; and one or more control parameters that most closely correspond to sensor data 608 in the predetermined mapping

can be returned. Using a predetermined mapping of sensor data to parameters can carry several advantages. For instance, performing a lookup from a predetermined mapping can be computationally less expensive than processing sensor data in real-time — particularly when the computations may involve complex geometric data, or where sensors of wearable system 601 may suffer from significant latency or bandwidth limitations. Further, predetermined mappings can capture relationships between sensor data and control parameters that may be difficult to compute from sensor data alone (e.g. by mathematical modeling).

[0054] In some examples, machine learning techniques can be used to generate or refine mappings of sensor data to control parameters, or to otherwise determine control parameters from sensor data 608. For example, a neural network (or other appropriate machine learning technology) can be trained to identify desired control parameters based on sensor data input, according to techniques familiar to those skilled in the art. Desired control parameters can be identified, and the neural network further refined, through user feedback. For example, users of system 600 can be prompted to rate the quality of speech recognition output (e.g., text output 660) and/or natural language processing output (e.g., natural language processing output 680). Such user ratings can be used to adjust the likelihood that a given set of sensor data will result in a particular set of control parameters. For example, if users report a high rating of text output 660 for a particular set of control parameters and a particular set of sensor data, a mapping of that sensor data to those control parameters can be created (or a link between them strengthened). Conversely, a low rating can cause a mapping of the sensor data to the control parameters to be weakened or deleted.

[0055] Similarly, machine learning techniques can be utilized to improve the ability of a speech recognition engine (e.g., 650) to discriminate between speech belonging to a user, and speech belonging to other entities — for example, speech emanating from a television or a stereo system. As described above, neural network (or other appropriate machine learning technology) can be trained to perform this discrimination (i.e., determining whether input audio belongs to the user, or to some other source). In some cases, a calibration routine, in which the user

provides a set of input audio known to belong to that user, can be used to train the neural network. Other suitable machine learning technologies can be used for the same purpose.

[0056] While update modules 632, 634, and 636 are described to comprise a beamforming update module, a noise reduction update module, and an echo cancellation module, respectively, other suitable modules may be included in any combination. For instance, in some examples, an EQ update module can determine how filtering processes (such as described above) can be updated, based on sensor data 608 and/or one or more control parameters computed at stage 630. Moreover, the functions described above with respect to modules 632, 634, and 636 can be performed as part of other elements of example system 600, such as stage 630 and/or stage 640, or speech recognition engine 650 or natural language processing engine 670.

[0057] Although the disclosed examples have been fully described with reference to the accompanying drawings, it is to be noted that various changes and modifications will become apparent to those skilled in the art. For example, elements of one or more implementations may be combined, deleted, modified, or supplemented to form further implementations. Such changes and modifications are to be understood as being included within the scope of the disclosed examples as defined by the appended claims.

CLAIMS

What is claimed is:

1. A method of processing an acoustic signal, the method comprising:
 - receiving, via a first microphone, a first acoustic signal associated with a first speech of a user of a wearable headgear unit;
 - receiving, via a sensor, a first sensor input;
 - determining a control parameter based on the sensor input; and
 - applying the control parameter to one or more of the first acoustic signal, the wearable headgear unit, and the first microphone,wherein:
 - determining the control parameter comprises determining, based on the first sensor input, a relationship between the first speech and the first acoustic signal.
2. The method of claim 1, wherein the control parameter is applied to the first acoustic signal to generate a second acoustic signal, and the method further comprises providing the second acoustic signal to a speech recognition engine to generate text output corresponding to the first speech.
3. The method of claim 1, wherein the control parameter is applied to the first acoustic signal to generate a second acoustic signal, and the method further comprises providing the second acoustic signal to a natural language processing engine to generate natural language data corresponding to the first speech.
4. The method of claim 1, wherein the wearable headgear unit comprises the first microphone.

5. The method of claim 1, wherein determining a control parameter based on the sensor input comprises:
 - detecting, based on the sensor input, a surface;
 - determining, an effect of the surface on a relationship between the first speech and the first acoustic signal; and
 - determining a control parameter that, when applied to the one or more of the first acoustic signal, the wearable headgear unit, and the first microphone, will reduce the effect of the surface on the relationship between the first speech and the first acoustic signal.
6. The method of claim 5, further comprising determining an acoustic property of the surface, wherein the effect of the surface on the relationship between the first speech and the first acoustic signal is determined based on the acoustic property.
7. The method of claim 1, wherein determining a control parameter based on the sensor input comprises:
 - detecting, based on the sensor input, a person different from the user;
 - determining an effect of a speech of the person on a relationship between the first speech and the first acoustic signal; and
 - determining a control parameter that, when applied to the one or more of the first acoustic signal, the wearable headgear unit, and the first microphone, will reduce the effect of the speech on the relationship between the first speech and the first acoustic signal.
8. The method of claim 1, wherein determining a control parameter based on the sensor input comprises applying the sensor input to the input of an artificial neural network.
9. The method of claim 1, wherein the control parameter is a control parameter for an echo cancellation module, and determining the control parameter based on the sensor input comprises:
 - detecting, based on the sensor input, a surface; and

determining a time of flight between the surface and the first microphone.

10. The method of claim 1, wherein the control parameter is a control parameter for a beamforming module, and determining the control parameter based on the sensor input comprises determining a time of flight between the user and the first microphone.
11. The method of claim 1, wherein the control parameter is a control parameter for a noise reduction module, and determining the control parameter based on the sensor input comprises determining a frequency to be attenuated in the first acoustic signal.
12. The method of claim 1, wherein:
 - the wearable headgear unit comprises a second microphone,
 - the sensor input comprises a second acoustic signal detected via the second microphone,
 - and
 - the control parameter is determined based on a difference between the first acoustic signal and the second acoustic signal.
13. The method of claim 1, wherein:
 - the wearable headgear unit comprises a plurality of microphones not including the first microphone,
 - the method further comprises receiving, via the plurality of microphones, a plurality of acoustic signals associated with the first speech; and
 - the control parameter is determined based on a difference between the first acoustic signal and the plurality of acoustic signals.
14. The method of claim 1, wherein the sensor is coupled to the wearable headgear unit.
15. The method of claim 1, wherein the sensor is positioned in an environment of the user.

16. A system comprising:
a wearable headgear unit including:
a display for displaying a mixed reality environment to a user;
a speaker; and
one or more processors configured to perform a method comprising:
receiving, via a first microphone, a first acoustic signal associated with a first speech of the user;
receiving, via a sensor, a first sensor input;
determining a control parameter based on the sensor input; and
applying the control parameter to one or more of the first acoustic signal, the wearable headgear unit, and the first microphone,
wherein:
determining the control parameter comprises determining, based on the first sensor input, a relationship between the first speech and the first acoustic signal.
17. The system of claim 16, wherein the control parameter is applied to the first acoustic signal to generate a second acoustic signal, and the method further comprises providing the second acoustic signal to a speech recognition engine to generate text output corresponding to the first speech.
18. The system of claim 16, wherein the control parameter is applied to the first acoustic signal to generate a second acoustic signal, and the method further comprises providing the second acoustic signal to a natural language processing engine to generate natural language data corresponding to the first speech.
19. The system of claim 16, wherein the wearable headgear unit further includes the first microphone.

20. The system of claim 16, wherein determining a control parameter based on the sensor input comprises:

detecting, based on the sensor input, a surface;

determining, an effect of the surface on a relationship between the first speech and the first acoustic signal; and

determining a control parameter that, when applied to the one or more of the first acoustic signal, the wearable headgear unit, and the first microphone, will reduce the effect of the surface on the relationship between the first speech and the first acoustic signal.

21. The system of claim 20, wherein the method further comprises determining an acoustic property of the surface, and wherein the effect of the surface on the relationship between the first speech and the first acoustic signal is determined based on the acoustic property.

22. The system of claim 16, wherein determining a control parameter based on the sensor input comprises:

detecting, based on the sensor input, a person different from the user;

determining an effect of a speech of the person on a relationship between the first speech and the first acoustic signal; and

determining a control parameter that, when applied to the one or more of the first acoustic signal, the wearable headgear unit, and the first microphone, will reduce the effect of the speech on the relationship between the first speech and the first acoustic signal.

23. The system of claim 16, wherein determining a control parameter based on the sensor input comprises applying the sensor input to the input of an artificial neural network.

24. The system of claim 16, wherein the control parameter is a control parameter for an echo cancellation module, and determining the control parameter based on the sensor input comprises:

detecting, based on the sensor input, a surface; and
determining a time of flight between the surface and the first microphone.

25. The system of claim 16, wherein the control parameter is a control parameter for a beamforming module, and determining the control parameter based on the sensor input comprises determining a time of flight between the user and the first microphone.

26. The system of claim 16, wherein the control parameter is a control parameter for a noise reduction module, and determining the control parameter based on the sensor input comprises determining a frequency to be attenuated in the first acoustic signal.

27. The system of claim 16, wherein:
the wearable headgear unit comprises a second microphone,
the sensor input comprises a second acoustic signal detected via the second microphone,
and
the control parameter is determined based on a difference between the first acoustic signal and the second acoustic signal.

28. The system of claim 16, wherein:
the wearable headgear unit comprises a plurality of microphones not including the first microphone,
the method further comprises receiving, via the plurality of microphones, a plurality of acoustic signals associated with the first speech; and
the control parameter is determined based on a difference between the first acoustic signal and the plurality of acoustic signals.

29. The system of claim 16, wherein the sensor is coupled to the wearable headgear unit.

30. The system of claim 16, wherein the sensor is positioned in an environment of the user.

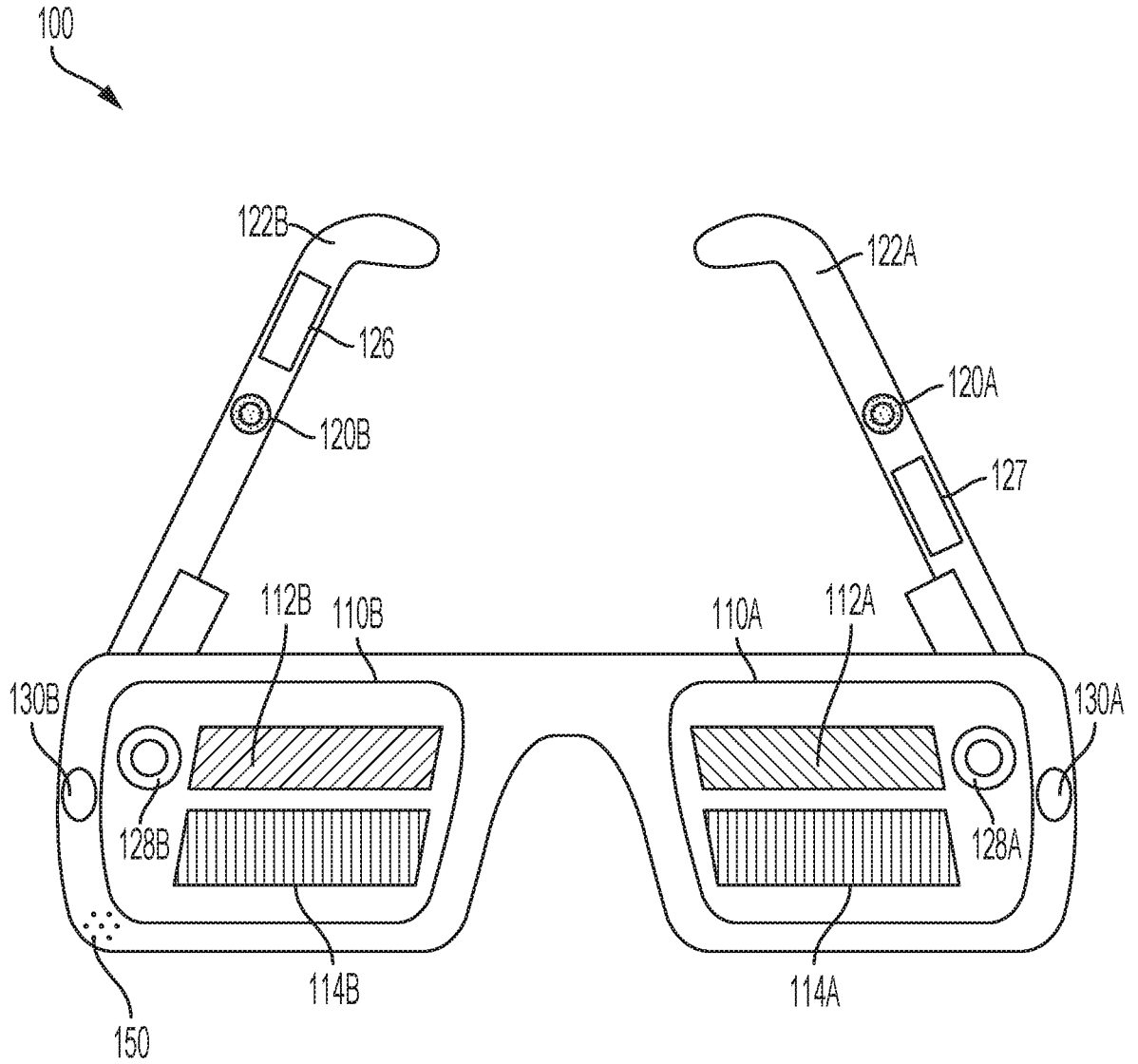


FIG. 1

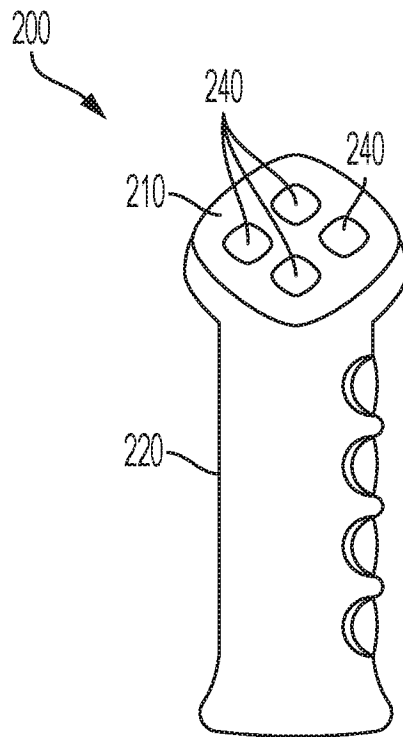


FIG. 2

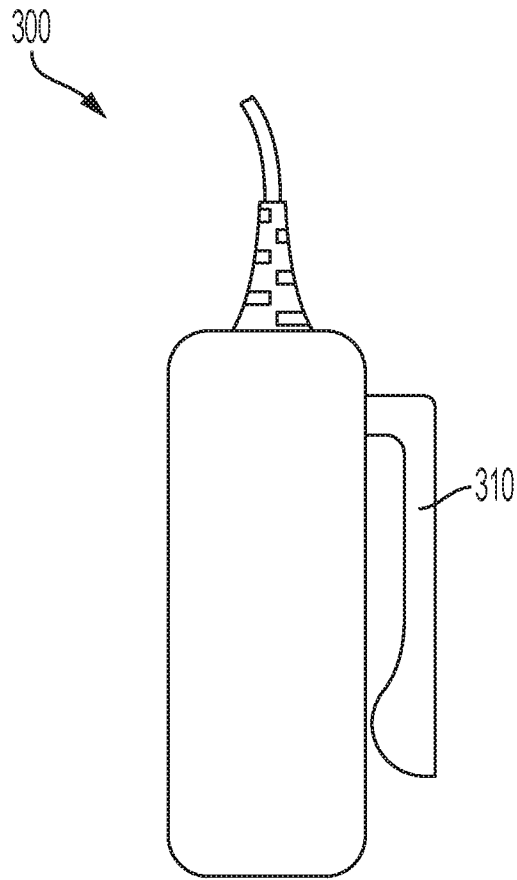


FIG. 3

400

4/6

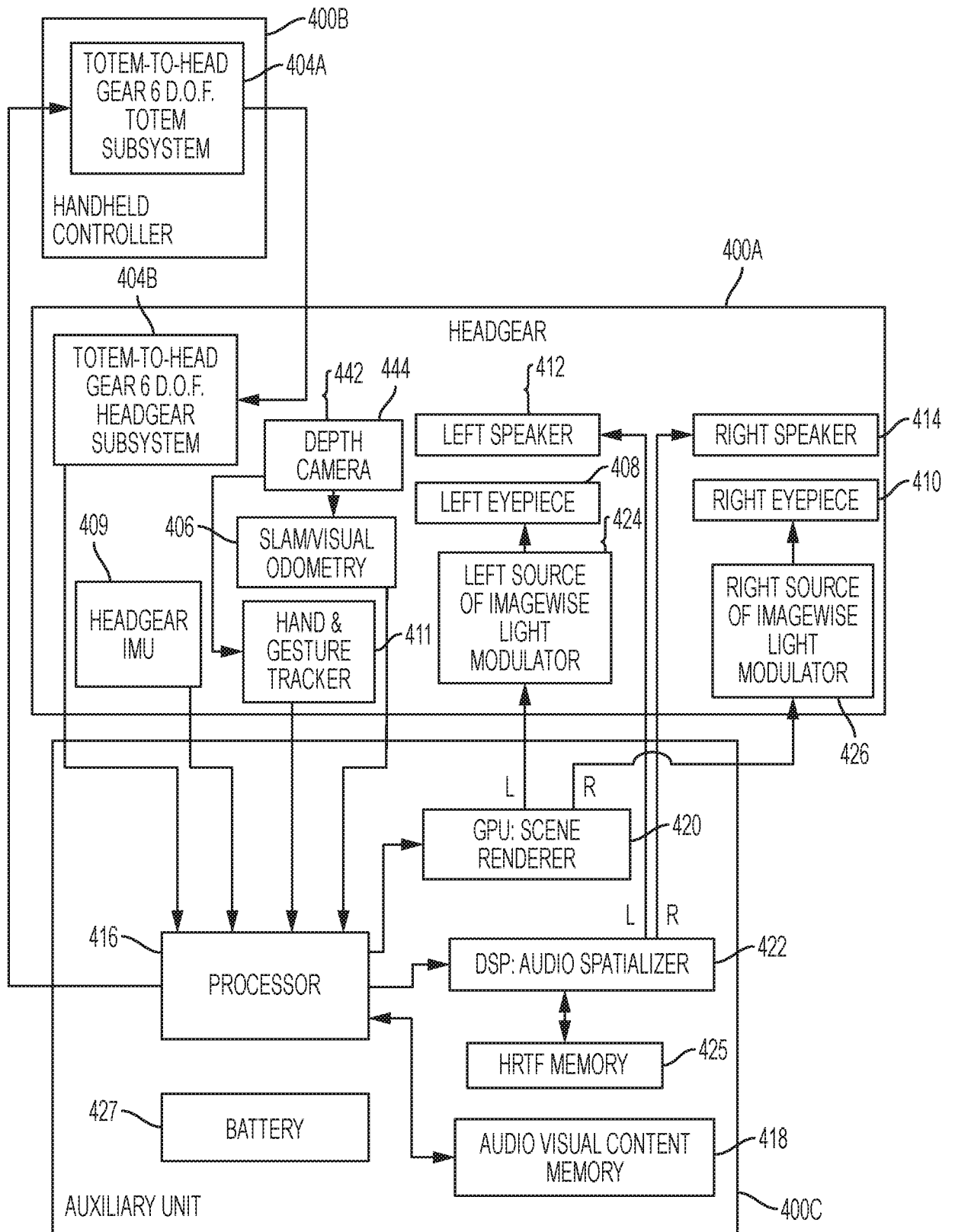


FIG. 4

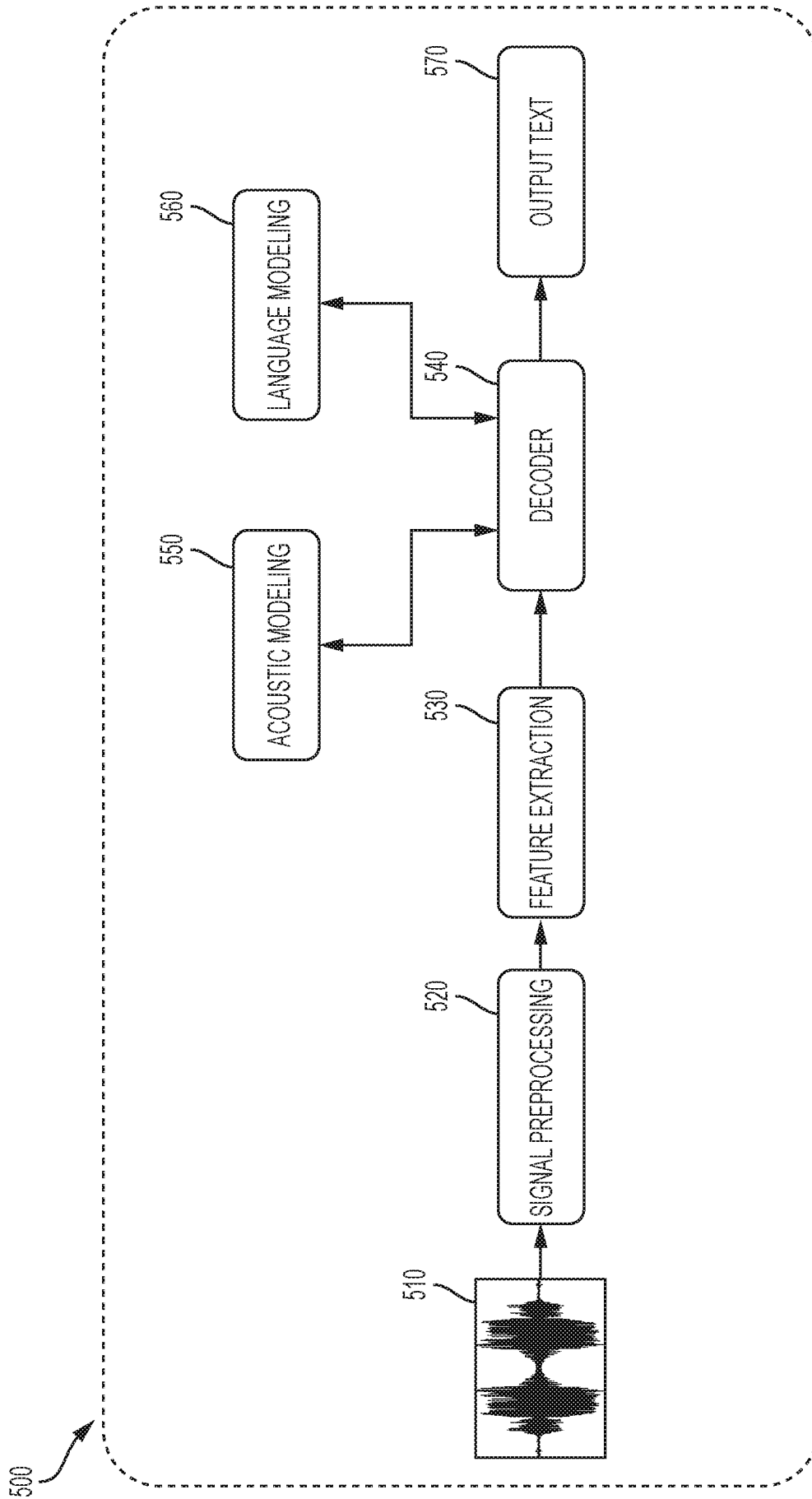


FIG. 5

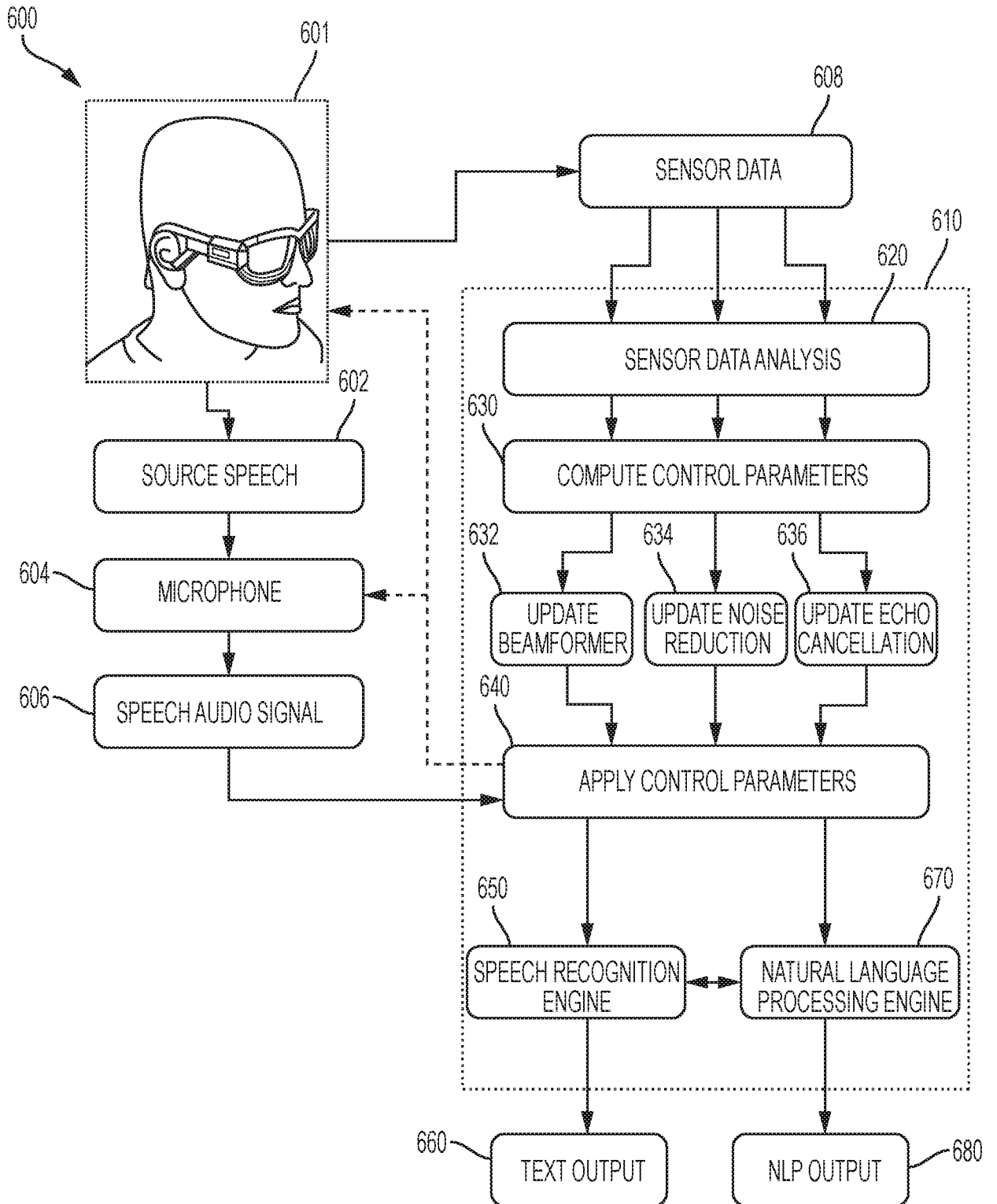


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2019/038546

A. CLASSIFICATION OF SUBJECT MATTER
 IPC(8) - G10L 21/02; G10L 17/22; H04R 3/00 (2019.01)
 CPC - G10L 15/1822; G10L 17/22; G06F 15/16; G10L 15/16; G06F 17/27; H04R 1/1091; H04R 5/027;
 H04R 5/04; H04R 2460/07 (2019.08)

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)
 See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
 USPC - 348/61; 381/92; 704/226 (keyword delimited)

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HARMA et al. "Augmented Reality Audio for Mobile and Wearable Appliances," J. Audio Eng. Soc., Vol. 52, No. 6, 2004 June, [retrieved on 2019-08-20]. Retrieved from the Internet: <URL: https://pdfs.semanticscholar.org/ae54/82c6a8d4add3e9707d780dfb5ce03d8e0120.pdf> pp 1-22	1, 4, 7, 9-15
---		---
Y	US 2011/0238407 A1 (KENT) 29 September 2011 (29.09.2011) entire document	2, 3, 5, 6, 8, 16-30
Y	TONGES. "An augmented Acoustics Demonstrator with Realtime stereo up-mixing and Binaural Auralization," Technische Universität Berlin, Audio Communication Group, [retrieved on 2019-08-22]. Retrieved from the Internet: <URL: https://www2.ak.tu-berlin.de/~akgroup/ak_pub/abschlussarbeiten/2015/ToengesRaffael_MasA.pdf> pp 1-100	2, 3, 8, 17, 18, 23
Y	WO 2014/113891 A1 (HU) 31 July 2014 (31.07.2014) entire document	5, 6, 20, 21
Y	US 2017/0091169 A1 (APPLE INC.) 30 March 2017 (30.03.2017) entire document	16-30
A	US 2014/0379336 A1 (BHATNAGAR) 25 December 2014 (25.12.2014) entire document	1-30
A	US 8,154,588 B2 (BURNS et al) 10 April 2012 (10.04.2012) entire document	1-30
A	WO 2015/169618 A1 (FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. et al) 12 November 2015 (12.11.2015) entire document	1-30
A	US 2012/0209601 A1 (JING) 16 August 2012 (16.08.2012) entire document	1-30

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
 21 August 2019

Date of mailing of the international search report

17 SEP 2019

Name and mailing address of the ISA/US
 Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
 P.O. Box 1450, Alexandria, VA 22313-1450
 Facsimile No. 571-273-8300

Authorized officer
 Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300
 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/038546

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2017/017591 A1 (VOCALZOOM SYSTEMS LTD.) 02 February 2017 (02.02.2017) entire document	1-30