US Patent & Trademark Office
Patent Public Search | Text View

United States Patent Application Publication	20250266053
Kind Code	A1
Publication Date	August 21, 2025
Inventor(s)	SHEEDER; Anthony Robert et al.

IDENTIFYING INPUT FOR SPEECH RECOGNITION ENGINE

Abstract

A method of presenting a signal to a speech recognition engine is disclosed. According to an example of the method, an audio signal is received from a user. A portion of the audio signal is identified, the portion having a first time and a second time. A pause in the portion of the audio signal, the pause comprising the second time, is identified. It is determined whether the pause indicates the completion of an utterance of the audio signal. In accordance with a determination that the pause indicates the completion of the utterance, the portion of the audio signal is presented as input to the speech recognition engine. In accordance with a determination that the pause does not indicate the completion of the utterance, the portion of the audio signal is not presented as input to the speech recognition engine.

Inventors: **SHEEDER; Anthony Robert (Fort Lauderdale, FL), ARORA; Tushar (Hollywood, FL)**

Applicant: **Magic Leap, Inc.** (Plantation, FL)

Family ID: **1000008575181**

Appl. No.: **19/197282**

Filed: **May 02, 2025**

Related U.S. Application Data

parent US continuation 16850965 20200416 parent-grant-document US 12327573 child US 19197282
us-provisional-application US 62836593 20190419

Publication Classification

Int. Cl.: **G10L25/87** (20130101); **G10L15/22** (20060101)

U.S. Cl.:

CPC **G10L25/87** (20130101); **G10L15/22** (20130101); G10L2015/227 (20130101)

Background/Summary

CROSS REFERENCE TO RELATED APPLICATIONS [0001] This application is a continuation of U.S. application Ser. No. 16/850,965, filed Apr. 16, 2020, which claims the benefit under 35 U.S.C. § 119 (e) of U.S. Provisional Patent Application 62/836,593, filed Apr. 19, 2019, the contents of which are incorporated herein by reference in their entireties for all purposes.

FIELD

[0002] This disclosure relates in general to systems and methods for processing speech signals, and in particular to systems and methods for processing a speech signal for presentation to a speech recognition engine.

BACKGROUND

[0003] Systems for speech recognition are tasked with receiving audio input representing human speech, typically via one or more microphones, and processing the audio input to determine words, logical structures, or other outputs corresponding to that audio input. For example, automatic speech recognition (ASR) systems 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. While some ASR systems may operate on a large corpus of speech recorded in advance—as one example, a system tasked with creating a written transcript of a speech that was recorded by a microphone the previous day—some ASR systems must respond to speech input provided in real-time. Real-time speech processing presents ASR systems with a unique set of challenges. For instance, ASR systems typically process speech not as monolithic blocks of input, but as a series of individual words or phrases that carry meaning (“utterances”).

[0004] Identifying when an utterance begins and ends can be crucial for an ASR system to accurately process a user's input speech and provide a desired result. As an example, consider a real-time “voice assistant” ASR system in communication with a weather reporting service: the ASR system can receive speech input from a user inquiring about the weather (e.g., “What's the current weather?”); convert the speech input into a structured query (e.g., a query for data indicating the past, current, or predicted future weather at a specific date and time, and at a specific location); present the structured query to the weather reporting service; receive the query results from the service; and present the query results to the user. The user expects the ASR system to process his or her complete question (rather than individual fragments of the question), and to promptly provide an accurate response. The user further expects that the ASR system will process naturally spoken commands that need not adhere to a specific, rigid format. In this example system, the onus is on the ASR to, in real-time, identify the user's complete question; and process the question to produce an accurate response in a timely manner-ideally, as soon as the user has finished asking the question.

[0005] In this example system, the accuracy of the response may depend on when the ASR system determines the user's question (an utterance) is complete. For instance, a user may ask, “What's the weather tomorrow?” If the ASR system prematurely determines that the utterance is complete after “What's the weather”, its corresponding query to the weather service would omit the modifier “tomorrow”, and the resulting response would thus be inaccurate (it would not reflect the user's desired date/time). Conversely, if the ASR system takes a more conservative approach, and waits for several seconds to confirm the entire utterance has been completed before processing the utterance, the user may not consider the ASR system to be sufficiently responsive to his or her commands. (Additionally, in some cases, such a long waiting period might create inaccuracies by including unrelated follow-up speech in the utterance.)

[0006] ASR systems struggle with this problem of determining, promptly and in real-time, when a speaker's utterance is complete. In some systems, a fixed time-out period is employed to determine the endpoint of an utterance: if, following speech input, no speech is received for the duration of the time-out period (e.g., 750 ms), the speech input may be considered to be the end of an utterance. However, the fixed time-out period solution is imperfect: for example, in situations where a user pauses to formulate a question; where the user is momentarily interrupted; or where the user's speech is otherwise disfluent (e.g., due to anxiety, a speech impediment, environmental distractions, cognitive load, etc.), the time-out period can expire before the user's utterance is complete. And conversely, once the user's utterance is complete, the response is delayed by at least the duration of the time-out period (in which the ASR system confirms no further input is received), and the user cannot provide additional speech input (e.g., belonging to a new utterance) for that duration. Such interactions limit the usefulness of ASR systems, and may highlight, unhelpfully, that the user is communicating with a machine, rather than another human being.

[0007] It is desirable for an ASR system to adopt a more intuitive approach to determining when a user has finished providing an utterance. In ordinary face-to-face interactions—and, to a lesser extent, telephonic interactions—people use a variety of contextual cues to understand when another person has finished talking. For example, when a speaker pauses, people evaluate the speaker's prosody, facial expression, eye gaze, mannerisms, gestures, and posture for indications whether the speaker is finished speaking, or has merely paused in the middle of a single thought. ASR systems may use similar cues to identify where a user's utterance begins and ends. As described below, in some examples, an ASR system can identify such contextual cues from microphone input; and in some examples, an ASR system in communication with one or more sensors (e.g., as part of a wearable system) can glean additional speech cues about the speaker from the outputs of those sensors, and use such cues to identify utterance boundaries without the problems, such as described above, associated with conventional solutions.

BRIEF SUMMARY

[0008] A method of presenting a signal to a speech recognition engine is disclosed. According to an example of the method, an audio signal is received from a user. A portion of the audio signal is identified, the portion having a first time and a second time. A pause in the portion of the audio signal, the pause comprising the second time, is identified. It is determined whether the pause indicates the completion of an utterance of the audio signal. In accordance with a determination that the pause indicates the completion of the utterance, the portion of the audio signal is presented as input to the speech recognition engine. In accordance with a determination that the pause does not indicate the completion of the utterance, the portion of the audio signal is not presented as input to the speech recognition engine.

Description

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] FIG. **1** illustrates an example wearable system according to examples of the disclosure.

[0010] FIG. **2** illustrates an example handheld controller that can be used in conjunction with an example wearable system according to examples of the disclosure.

[0011] FIG. **3** illustrates an example auxiliary unit that can be used in conjunction with an example wearable system according to examples of the disclosure.

[0012] FIG. **4** illustrates an example functional block diagram for an example wearable system according to one or more examples of the disclosure.

[0013] FIG. **5** illustrates an example audio waveform for input to an example ASR system according to one or more examples of the disclosure.

[0014] FIG. **6** illustrates a flow chart of an example of processing an acoustic speech signal using an ASR system according to one or more examples of the disclosure.

[0015] FIG. **7** illustrates a flow chart of an example of processing an acoustic speech signal using an ASR system according to one or more examples of the disclosure.

[0016] FIGS. **8**A-**8**B illustrate flow charts of examples of detecting a pause in an input speech signal according to one or more examples of the disclosure.

[0017] FIGS. **9**A-**9**B illustrate flow charts of examples of determining whether an input utterance has concluded according to one or more examples of the disclosure.

[0018] FIG. **10** illustrates a flow chart of an example of classifying input data to determine a probability associated with that input data according to one or more examples of the disclosure.

DETAILED DESCRIPTION

[0019] 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.

Example Wearable System

[0020] 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 **110**A and **110**B (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 **112**A/**112**B and exit pupil expansion (EPE) grating sets **114**A/**114**B); left and right acoustic structures, such as speakers **120**A and **120**B (which may be mounted on temple arms **122**A and **122**B, 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 **122**A); left and right cameras (e.g., depth (time-of-flight) cameras **130**A and **130**B) 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 **128**B). 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.

[0021] 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**).

[0022] 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.

[0023] 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 **400**B, referred to here as a “totem” (and which may correspond to handheld controller **200** described above); the handheld controller **400**B can include a totem-to-headgear six degree of freedom (6DOF) totem subsystem **404**A. Wearable system **400** can also include example wearable head device **400**A (which may correspond to wearable headgear device **100** described above); the wearable head device **400**A includes a totem-to-headgear 6DOF headgear subsystem **404**B. In the example, the 6DOF totem subsystem **404**A and the 6DOF headgear subsystem **404**B cooperate to determine six coordinates (e.g., offsets in three translation directions and rotation along three axes) of the handheld controller **400**B relative to the wearable head device **400**A. The six degrees of freedom may be expressed relative to a coordinate system of the wearable head device **400**A. 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 **400**A; 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 **400**B can include a camera, as described above; and the headgear **400**A can include an optical target for optical tracking in conjunction with the camera. In some examples, the wearable head device **400**A and the handheld controller **400**B 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 **400**B relative to the wearable head device **400**A may be determined. In some examples, 6DOF totem subsystem **404**A 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 **400**B.

[0024] 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 **400**A) 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 **400**A 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 **400**A), rather than at a fixed position and orientation on the display (e.g., at the same position in the display of wearable head device **400**A). 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 **400**A 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 **400**A 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 **400**A. 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.

[0025] 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 **400**A. 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.

[0026] In some examples, one or more processors **416** may be configured to receive data from headgear subsystem **404**B, 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 **404**A. The processor **416** may be coupled to the 6DOF totem system **404**A wirelessly, such as in examples where the handheld controller **400**B 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 **400**B). 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.

[0027] 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 **400**C (which may correspond to auxiliary unit **300** described above). The auxiliary unit **400**C may include a battery **427** to power its components and/or to supply power to wearable head device **400**A and/or handheld controller **400**B. 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 **400**A, which can in turn reduce fatigue of a user's head and neck.

[0028] 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 **400**C could instead be associated with wearable head device **400**A or handheld controller **400**B. Furthermore, some wearable systems may forgo entirely a handheld controller **400**B or auxiliary unit **400**C. Such changes and modifications are to be understood as being included within the scope of the disclosed examples.

Speech Recognition Systems

[0029] Speech recognition systems in general comprise a speech recognition engine that can 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. In the case of automatic speech recognition (ASR) systems, for example, the output of a speech recognition 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), including those that do not necessarily “recognize” speech, are contemplated and are within the scope of the disclosure. Further, as used herein, a speech recognition engine can include one or more of an automated speech recognition engine, a natural language understanding engine, and other suitable components.

[0030] ASR 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 ASR 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 recognition 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 engines can heighten feelings of immersion. As such, ASR 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.

Identifying Input Speech Boundaries

[0031] The effectiveness of an ASR system may be limited by its ability to promptly present accurate input data to a speech recognition engine. Presenting accurate input data may require correctly identifying when individual sequences of input start and end. Some ASR systems struggle to determine, promptly and in real-time, when a speaker's utterance is complete. The present disclosure is directed to systems and methods for improving the accuracy of a speech processing system by accurately identifying the endpoints of utterances presented as input to the speech processing system. Quickly and accurately determining where an utterance ends enables a speech processing system to promptly deliver correct results in real-time—that is, in response to a live stream of input audio, where the entire input audio signal cannot be known in advance.

[0032] FIG. **5** illustrates an example audio waveform **500**, such as may be detected by one or more microphones and presented as input to an ASR system. Waveform **500** represents a user speaking the example utterance “What's the weather tomorrow in Moscow?”, with the intention that an ASR system receiving that utterance as input will query a weather service, and respond with tomorrow's weather forecast for Moscow. The speed and accuracy of the ASR system's response will depend on what the ASR system determines to be the endpoint of the user's utterance. Example waveform **500** begins at an initial time t.sub.0. If the ASR system determines, prematurely, that the utterance ends at time t.sub.1, which falls after the phrase “what's the weather” and before the word “tomorrow”, the input utterance (i.e., the input speech falling between time t.sub.0 and time t.sub.1) will be determined to be “what's the weather”. Because this input utterance would lack the qualifiers added by the user (i.e., “tomorrow” and “in Moscow”), the ASR system's response to this utterance will not match the user's expectations. For example, the system might return the current weather (not tomorrow's weather) in the user's current location (not in Moscow). Similarly, if the ASR system determines that the utterance ends at time t.sub.2, which falls after the phrase “what's the weather tomorrow” and before the phrase “in Moscow”, the input utterance (i.e., the input speech falling between time to and time t.sub.2) will be determined to be “what's the weather tomorrow”, and the resulting response will again not match the user's expectations (tomorrow's weather in Moscow). An ideal ASR system may determine the end of the utterance to be at time t.sub.3, which falls immediately after the conclusion of the entire input query, and would correctly identify the input utterance as “what's the weather in Moscow”.

[0033] The ideal ASR system would also not include trailing portions of the input signal that do not belong to the input utterance. For example, if an ASR system determines that the utterance ends at time t.sub.4, the input utterance would include all of the correct input speech (i.e., “what's the weather in Moscow”) but would also include extraneous information (the portion of the input speech signal between t.sub.3 and t.sub.4). This extraneous information could introduce errors into the input utterance, and further, will delay the ASR system's response (i.e., by at least the span of processing time of the signal between t.sub.3 and t.sub.4), resulting in a perceived lack of responsiveness by the user.

[0034] Some ASR systems may incorrectly identify the endpoint of an input utterance. For instance, when presented with example waveform **500** as input, some ASR systems may incorrectly identify the end of the utterance as t.sub.1, t.sub.2, or t.sub.4, rather than t.sub.3.

[0035] FIG. **6** illustrates an example process **600** that may be executed by an ASR system. Example process **600** uses a time-out interval to determine when an input utterance has concluded; that is, when input speech has not been received for an amount of time exceeding the time-out interval, the utterance is deemed complete and is presented to an ASR engine for processing. As shown in the figure, input speech is detected in real-time at stage **610** from one or more microphones **602**. At stage **620**, process **600** determines whether input speech is still being received; if so, the input utterance is deemed ongoing, and process **600** returns to stage **610** to continue detecting the input speech. If not, process **600** queries whether a time-out interval **632** has elapsed. If not, process **600** returns to stage **610**; if so, the utterance is deemed complete (stage **640**). At stage **650**, process **600** then presents the utterance (or some representation of the utterance) to an ASR engine **660** for processing. The ASR engine **660** can generate a speech-to-text representation of the input utterance. A natural language understanding (NLU) engine **665** can perform additional processing based on the representation. For example, the NLU engine **665** can generate a semantic representation of the speech-to-text representation output from the ASR engine **660**; determine that the input utterance represents a request for a weather report; query a weather reporting service using a structured query generated from the input utterance; and receive a response from the weather reporting service. In some embodiments, process **600** may stream the utterance (or some representation of the utterance) for processing (e.g., by the ASR engine **660**). In some embodiments, the ASR engine **660** may return or output incremental results. In some embodiments, the NLU engine **665** may receive the output from the ASR engine **660** for processing after a conclusion of the utterance. At stage **670**, process **600** can receive the response from the NLU engine **665**, and at stage **680** present the response to the user (e.g., via a text-to-speech engine coupled to an output speaker).

[0036] The above process **600** may be prone to error because, by concluding an input utterance at stages **630** and **640** using a simple time-out interval, process **600** can prematurely conclude an utterance before the user has completed speaking the utterance. With reference to waveform **500** described above, this may result in the input utterance terminating at time t.sub.1 or t.sub.2 rather than t.sub.3. This may happen when the user inadvertently inserts gaps of non-speech between two words of a single utterance (e.g., pauses between “weather” and “tomorrow” or between “tomorrow” and “in Moscow” in the example waveform **500**). If these gaps exceed the length of the time-out interval **632**, process **600** may prematurely determine that the input utterance has completed, even though the user is still completing that utterance. (This situation may be especially common with complex input queries, where the user may need additional time to formulate their question; or among users with speech impediments, or those who experience anxiety when interacting with microphones or ASR systems.)

[0037] The problem may not be completely solvable simply by increasing the length of the time-out interval **632**, because there is a tradeoff between the duration of this interval and the perceived responsiveness of the ASR system. That is, even if the time-out interval **632** can be increased such that it exceeds any possible intra-utterance input gap—preventing process **600** from prematurely cutting off an input utterance—the ASR system waits for the duration of that extended time-out interval before determining that the utterance has concluded. This delay can annoy the user, who may perceive the delay as non-responsiveness—particularly in comparison to face-to-face human interaction, in which listeners quickly and intuitively understand when a speaker is finished speaking. In some embodiments, the delay may lead to cross-talk—when a user perceives the ASR system to be unresponsive and begins speaking again (e.g., to reiterate the initial input)—which may result in a cascade of errors.

[0038] FIG. **7** illustrates an example process **700** (e.g., which may be executed by an ASR system) that can identify an input utterance more promptly and accurately than can process **600**. In process **700**, as described below, a pause in input speech can be detected, and then analyzed for contextual cues to indicate whether the pause likely represents the completion of the current utterance (in which case the utterance can be concluded and presented to an ASR engine and a NLU engine), or whether the pause indicates that the current utterance is ongoing (in which case the ASR system should continue detecting the current utterance).

[0039] In process **700**, audio input presented by a user is detected at stage **710** from one or more microphones **602**. (In some examples, audio input can be received as streaming data, or as one or more data files, instead of or in addition to microphone output.) This audio input can be stored in an input buffer, or other memory, for access by process **700**. At stage **720**, process **700** can determine (for example, based on the input buffer and/or sensor data as described in more detail below) whether the user has paused while presenting the input speech. If no pause is detected, indicating that the user's current utterance is ongoing, the process can return to stage **710** to continue detecting the audio input. If a pause is detected at stage **720**, process **700** can determine at stage **730** the likelihood that the pause indicates the completion of the current utterance (rather than the continuation of the current utterance). For example, stage **720** can determine a numeric confidence value, representing the likelihood that the pause indicates the current utterance is complete. This determination can be made based on the contents of the input buffer and/or sensor data, as described in more detail below.

[0040] At stage **732**, process **700** can evaluate the determination, at stage **730**, whether the detected pause indicates the completion of the current utterance. If it has been determined with a sufficient confidence (e.g., with a confidence level that exceeds a threshold) that the pause indicates the completion of the current utterance, process **700** can proceed to conclude the utterance (stage **740**); present the utterance (stage **750**) to an ASR engine (**760**); receive a response (stage **770**) from a NLU engine (**765**); and present the response to the user (stage **780**). These steps may correspond to stage **640**, stage **650**, ASR engine **660**, NLU engine **665**, stage **670**, and stage **680**, respectively, described above with respect to process **600**.

[0041] If process **700** determines that the pause does not likely indicate the current utterance has been completed (e.g., that a determined confidence level does not meet a threshold value), process **700** at stage **732** can take various actions in response. In some examples, process **700** can adjust or reset (stage **734**) a parameter used to determine whether a pause is detected, such as described herein with respect to stage **720**. For instance, process **700** at stage **734** can increase, or reset, a time-out interval used at stage **720** to detect a pause in input speech. This may be beneficial if process **700** determines that more time is needed to determine whether the user intends to complete the current utterance. In some examples, process **700** can present the user with a prompt (stage **736**), such as a prompt for additional input (e.g., a visual and/or audible prompt that asks the user to indicate whether they are done speaking). This can be beneficial in situations where it is unclear whether the current utterance is completed—for example, where process **700** returns a confidence value less than, but close to, the threshold value. In some examples, upon detecting that the pause does not indicate the current utterance is completed, process **700** can combine the current utterance with a second utterance (stage **738**); for instance, an utterance preceding the pause could be concatenated with a second utterance following the pause, for presentation of the combined utterances to a speech recognition engine (e.g., the ASR engine and/or NLU engine). In some examples, process **700** may return to stage **710** to continue detecting input speech, without taking any additional action such as described with respect to stages **734**, **736**, or **738**; this behavior may be preferred where stage **730** returns a confidence value that is far below the threshold required to conclude that the current utterance is complete.

[0042] FIG. **8**A illustrates a flow chart of an example process for implementing stage **720** of process **700**, such as described above with respect to FIG. **7**. In the figure, audio input data **810** (e.g., speech input signals stored in an input buffer) can be evaluated to determine the presence of a pause in the input. At stage **820**, the process can determine whether a value of one or more properties of an input signal is above or below a threshold for a period of time exceeding a time-out interval. In some examples, whether an amplitude of the input signal has been below a threshold amplitude level for a period of time exceeding the time-out interval may be determined. If so, this can indicate a pause in the user's input speech (stage **860**), such as described above with respect to FIG. **6**. In some embodiments, the length of the time-out interval may vary depending on the property. In some examples, spectral analysis can identify a speech signal as distinct from other ambient or incidental sounds, and whether an output of the analysis has been above or below one or more thresholds for a period of time exceeding the time-out interval may be determined.

[0043] However, even if it is determined at stage **820** that the time-out interval **822** has not elapsed, the process can examine audio input data **810** (stage **830**) to determine whether the speech data includes verbal cues (other than relative silence) that indicate a pause in the input speech. These verbal cues can include characteristics of the user's prosody (e.g., rhythm, intonation, timbre, volume), the presence of trailing words, the presence of terminating words or phrases (e.g., “thank you” when completing a verbal request), and the like. These verbal cues can indicate that the current utterance is complete, even if the time-out interval has not yet elapsed. At stage **840**, the process can evaluate whether any such verbal cues exist, and if so, whether they indicate that the input speech has paused (stage **860**) or not (stage **850**). In some cases, stage **840** can make this determination by comparing a confidence level generated at stage **830** against a threshold value. By evaluating the presence of verbal cues to indicate that an utterance has completed, even before the expiration of a time-out interval, the process can avoid perceptions of non-responsiveness, such as described above, that can result from waiting for the conclusion of the time-out interval before presenting the utterance for processing (e.g., by an ASR engine and/or NLU engine).

[0044] FIG. **8**B illustrates a flow chart of an example process for implementing stage **720** of process **700** in which sensor input data **844** is used, instead of or in conjunction with audio input data **810** such as described above. In some examples, as described above, sensor input data **844** can correspond to data from sensors such as described above with respect to example wearable head device **100** in FIG. **1**. As described above, such 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 (e.g., electrooculography (EOG) sensors) to measure a user's eye movements or eye gaze; 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 and can provide sensor input data **844**. Moreover, in some examples, sensors other than those of a wearable system can be utilized as appropriate. For instance, sensors associated with a microphone of a speech recognition system (e.g., GPS, IMU) could be used to in conjunction with sensors of a wearable system to determine a relative distance and orientation between the user and the speech recognition system.

[0045] At stage **842**, process **800** can examine sensor input data **844** to determine whether the sensor data includes non-verbal cues that indicate a pause in the input speech. These non-verbal cues can include, for example, characteristics of the user's eye gaze; head pose; gestures; vital signs (e.g., breathing patterns, heart rate); and facial expression. These non-verbal cues can indicate that the current utterance is complete, even if the time-out interval has not yet elapsed, and even in the absence of verbal cues such as described above with respect to FIG. **8**A. For instance, a pause in a user's speech might correspond with a change in the user's eye gaze target, a change in the user's head pose, a gesture performed by the user, a change in the user's vital signs (e.g., breathing pattern, heart rate), a change in the user's facial expression, a change in movement or rotation away from a microphone, a change in the user's posture or other physical characteristics indicated by sensor input data **844**, a change in orientation, and/or a rate of change of any one or more of the aforementioned characteristics. At stage **846**, the process can evaluate whether any such non-verbal cues exist, and if so, whether they indicate that the input speech has paused (stage **860**) or not (stage **850**). In some cases, stage **846** can make this determination by comparing a confidence level generated at stage **842** against a threshold value. As above, by evaluating the presence of non-verbal cues to indicate that an utterance has completed, even before the expiration of a time-out interval, the process can avoid perceptions of non-responsiveness, such as described above, that can result from waiting for the conclusion of the time-out interval before presenting an utterance to a speech recognition engine. While FIG. **8**B shows verbal cues and non-verbal cues being analyzed in separate stages (i.e., stages **840** and **846**), some examples can analyze verbal cues and non-verbal cues in combination, in a single stage.

[0046] FIG. **9**A illustrates a flow chart of an example process for implementing stage **730** of process **700**, such as described above with respect to FIG. **7**. At stage **730**, process **700** determines, for a pause identified at stage **720** such as described above, whether or not the pause likely corresponds to the conclusion of the current utterance. In FIG. **9**A, audio input data **910** (e.g., speech signals stored in an input buffer, which may correspond to **810** described above) can be evaluated at stage **920** to determine the presence of interstitial sounds in the audio input data. Interstitial sounds may be words, phrases, syllables, or other vocalizations present in the input audio that can indicate that the current utterance is not yet complete (such as where the user is mid-thought). For example, interstitial sounds can include hesitation sounds (e.g., “um,” “uh”); elongated syllables (e.g., an elongated “to(ooo)” at the end of the phrase “I'm going to”); repetitions (e.g., “and, and, and . . . ”); trailing filler words (e.g., “like,” “I mean”), and/or other indications that the user is likely to provide additional input audio belonging to the current utterance. Such interstitial sounds can be specific to individual users, to particular languages, or types of verbal input (e.g., questions, declarative statements). As described below, various classifiers can be employed to identify interstitial sounds.

[0047] At stage **930**, the process can determine whether any such interstitial sounds were detected at stage **920**. If not, the process can conclude (stage **970**) that the current utterance is completed. If interstitial sounds are present, the process at stage **940** can evaluate whether the interstitial sounds indicate that the current utterance is ongoing. For example, the presence of hesitation sounds can indicate that the user is in the process of formulating a complete utterance (as in, for instance, “What's the weather . . . uh . . . tomorrow”). Similarly, elongated syllables, repetitions, filler words, and other interstitial sounds can indicate that the current utterance is not yet complete. In some examples, stage **940** can generate a confidence value, indicating the likelihood that interstitial sounds are present and indicate whether the current utterance is or is not complete.

[0048] At stage **950**, if it is determined at stage **940** that the current utterance is ongoing, the process can conclude (stage **960**) that the current utterance is not completed. As described above with respect to FIG. **7**, this can result in the process performing various actions in response: for example, process **700** can extend a time-out interval (e.g., **822**) in which to detect a pause, prompt the user for additional input indicating whether the current utterance is complete, and/or combine the current utterance with a second utterance; or take no action at all. In some examples, which action (if any) is performed can depend on a confidence value generated at stage **940**; for instance, in response to a high confidence value that the utterance is not yet completed, process **700** may merely return to stage **710** to continue detecting audio input, without taking any further action; and in response to a medium confidence value (indicating, e.g., uncertainty regarding whether the current utterance is complete), process **700** may expressly prompt the user for additional input (stage **736**). Similarly, if it is determined at stage **940** that the current utterance is completed, the process can indicate so (stage **970**), and the process can proceed to present the utterance to a speech recognition system, such as described above.

[0049] FIG. **9**B illustrates a flow chart of an example process for implementing stage **730** of example process **700** in which sensor input data **942** is used, instead of or in conjunction with audio input data **910** such as described above. Sensor input data **942** can correspond to sensor input data **844** described above: for example, sensor input data **942** can be output by sensors such as described above with respect to example wearable head device **100** in FIG. **1**. As described above, such sensors can include one or more cameras (e.g., RGB cameras, depth cameras); LIDAR units; radar units; acoustic sensors; GPS units; accelerometers; gyroscopes; IMUs; environmental sensors; biometric sensors (e.g., iris cameras, fingerprint sensors, eye tracking sensors, and/or sensors to measure a user's vital signs). Other suitable sensors can be included and can provide sensor input data **942**. Moreover, in some examples, sensors other than those of a wearable system can be utilized as appropriate. For instance, as described above, sensors associated with a microphone of a speech recognition system (e.g., GPS, IMU) could be used to in conjunction with sensors of a wearable system to determine a relative distance and orientation between the user and the speech recognition system.

[0050] With respect to FIG. **9**B, sensor input data **942** can be evaluated at stage **944** to determine whether the sensor data indicates that the current utterance is ongoing, or whether the current utterance is completed. For instance, the completion (or non-completion) of an utterance might correspond with a change in the user's eye gaze target, a change in the user's head pose, a gesture performed by the user, a change in the user's vital signs (e.g., breathing pattern, heart rate), a change in the user's facial expression, a change in movement or rotation away from a microphone, a change in the user's posture or other physical characteristics indicated by sensor input data **944**, a change in orientation, and/or a rate of change of any one or more of the aforementioned characteristics. In some examples, stage **944** may generate a confidence level, indicating a likelihood with which the current utterance is completed. Based on the determination made at stage **944** (e.g., by comparison of the confidence level to a threshold value), the process at stage **950** can indicate that the utterance either is (stage **970**) or is not (stage **960**) completed.

[0051] In process **700** described above, input data (e.g., audio data, sensor data) can be evaluated at one or more stages for its significance with respect to how data should be presented to a speech recognition engine (e.g., the ASR engine and/or the NLU engine). For instance, at stage **830** of process **720**, as described above, audio input data can be evaluated to determine whether the data includes verbal cues that the current utterance is complete. At stage **842**, as described above, sensor data can be evaluated for non-verbal cues (e.g. changes in facial expression) that the current utterance is complete. At stage **920**, as described above, audio input data can be evaluated to identify the presence of interstitial sounds; and at stage **940**, it can be evaluated whether those interstitial sounds indicate that the current utterance is ongoing. And at stage **944**, as described above, sensor input data can be evaluated to determine whether the sensor input data indicates that the current utterance is ongoing.

[0052] In some examples, audio input data and/or sensor input data used as described above can be classified according to one or more parameters-resulting in one or more classifiers representing the data. These classifiers can be used (e.g., by example process **700**) to evaluate the significance of that data (e.g., a probability associated with the data). FIG. **10** illustrates an example process **1000** for classifying input data **1010** to determine a probability of interest associated with that input data. As used herein, a probability of interest can correspond to a probability described above with respect to example process **700**: a probability that audio input data and/or sensor input data indicates a pause in input speech; a probability that a pause indicates the completion of an utterance; and/or a probability that the presence of interstitial sounds indicates that an utterance is ongoing; or another suitable probability. With respect to FIG. **10**, this determination can be performed using audio input data **1016**, alone or in combination with sensor input data **1020**. Determining a probability value for input data **1010** can be referred to as “classifying” the input data **1010**, and a module or process for performing this determination (e.g., **1074**) can be referred to as a “classifier.”

[0053] In the example process shown in FIG. **10**, input data **1010** (e.g., audio input data **1016** and/or sensor input data **1020**) can be used in conjunction with speech/sensor data **1029** (e.g., from a database) to determine one or more probabilities of interest for input data **1010**. In some examples, audio input data **1016** and/or sensor input data **1020** can be parameterized/characterized according to one or more parameters at stage **1075** in order to facilitate classifying the speech segment based on speech/sensor data **1029**. A Fourier transform of the input data **1010** can be performed in order to provide a spectral representation of the input data **1010** (e.g., a function of frequency indicating the relative prevalence of various frequency parameters in audio input data **1016** and/or sensor input data **1020**). For instance, this process can identify levels of (or changes in) amplitude or component frequencies of a user's speech, position, eye gaze, and/or body movements; these values can be indicative of a pause in the user's speech, the presence of interstitial sounds, or the conclusion of an utterance of the user, such as described above. In some examples, characteristics of the user—for example, the user's age, sex, and/or native language—can be used as parameters to characterize the input data **1010**. Other ways in which input data **1010** can be parameterized, with such parameters used to determine a probability of interest of the input data, will be apparent to those skilled in the art.

[0054] At stage **1076** of the example, a probability value **1078** is determined for a probability of interest of input data **1010**. In some examples, probability value **1078** can be determined using speech/sensor data **1029**, such as where a database including speech/sensor data **1029** identifies, for elements of speech and/or sensor data in the database, whether those elements correspond to input speech. In some examples, audio input data **1016** can include a set of audio waveforms corresponding to speech segments; and can indicate, for each waveform, whether the corresponding speech segment indicates a pause or an interstitial sound. In some examples, instead of or in addition to audio waveforms, audio input data **1016** can include audio parameters that correspond to the speech segments. Audio input data **1016** can be compared with the speech segments of speech/sensor data **1029**—for example, by comparing an audio waveform of audio input data **1016** with analogous waveforms of speech/sensor data **1029**, or by comparing parameters of audio input data **1016** (such as may be characterized at stage **1075**) with analogous parameters of speech/sensor data **1029**. Based on such comparisons, probability value **1078** can be determined for audio input data **1016**.

[0055] Analogous techniques can be applied with respect to sensor input data **1020**. For example, sensor input data **1020** can include sequences of raw sensor data; and can indicate, for the raw sensor data, whether that data indicates a pause, or the completion or continuation of an utterance. Similarly, sensor input data **1020** can include sensor input parameters that correspond to the sensor data. Sensor input data **1020** can be compared with elements of speech/sensor data **1029** such as described above with respect to audio input data **1016**.

[0056] Techniques for determining probability **1078** based on input data **1010** will be familiar to those skilled in the art. For instance, in some examples, nearest neighbor interpolation can be used at stage **1076** to compare elements of input data **1010** to similar data elements in an N-dimensional space (in which the N dimensions can comprise, for example, audio parameters, audio waveform data, sensor parameters, or raw sensor data described above); and to determine probability value **1078** based on the relative distances between an element of input data **1010** and its neighbors in the N-dimensional space. As another example, support vector machines can be used at stage **1076** to determine, based on speech/sensor database **1029**, a basis for classifying an element of input data **1010** as either indicating an utterance is complete or indicating the utterance is not complete; and for classifying input data **1010** (e.g., determining a probability value **1078** that input data **1010** indicates a completed utterance, a pause, or the presence of interstitial sounds) according to that basis. Other suitable techniques for analyzing input data **1010** and/or speech/sensor data **1029**, comparing input data **1010** to speech/sensor data **1029**, and/or classifying input data **1010** based on speech/sensor data **1029** in order to determine probability **1078** will be apparent; the disclosure is not limited to any particular technique or combination of techniques.

[0057] In some examples, machine learning techniques can be used, alone or in combination with other techniques described herein, to determine probability value **1078**. For example, a neural network could be trained on speech/sensor data **1029**, and applied to input data **1010** to determine probability value **1078** for that input data. As another example, a genetic algorithm can be used to determine a function, based on speech/sensor data **1029**, for determining probability value **1078** corresponding to input data **1010**. Other suitable machine learning techniques, which will be familiar to those skilled in the art, will be apparent; the disclosure is not limited to any particular technique or combination of techniques.

[0058] In some examples, speech/sensor data **1029** can be generated by recording a set of speech data and/or sensor data for various users, and identifying, for elements of that data, whether the user has completed an utterance; has paused his or her speech; or is providing interstitial sounds. For instance, a user could be observed interacting with a group of people, with a speech recognition system present in the same room, as the user's speech is recorded; sensor data for the user (e.g., output by a wearable system worn by the user) can also be recorded. The observer can identify, for each region of the recorded data, whether that region of data corresponds to pausing, providing interstitial sounds, or completing an utterance. This information can be apparent to the observer by observing the context in which the user is speaking—commonly, it is easy and intuitive for humans (unlike machines) to determine, based on an observation of a user, whether the user has completed an utterance. This process can be repeated for multiple users until a sufficiently large and diverse set of speech/sensor data is generated.

[0059] With respect to the systems and methods described above, elements of the systems and methods can be implemented by one or more computer processors (e.g., CPUs or DSPs) as appropriate. The disclosure is not limited to any particular configuration of computer hardware, including computer processors, used to implement these elements. In some cases, multiple computer systems can be employed to implement the systems and methods described above. For example, a first computer processor (e.g., a processor of a wearable device coupled to a microphone) can be utilized to receive input microphone signals, and perform initial processing of those signals (e.g., signal conditioning and/or segmentation, such as described above). A second (and perhaps more computationally powerful) processor can then be utilized to perform more computationally intensive processing, such as determining probability values associated with speech segments of those signals. Another computer device, such as a cloud server, can host a speech recognition engine, to which input signals are ultimately provided. Other suitable configurations will be apparent and are within the scope of the disclosure.

[0060] 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

1. A method comprising: receiving, via a microphone, an audio signal, wherein the audio signal comprises voice activity; receiving, via one or more sensors of a head-wearable device, non-verbal sensor data corresponding to a user of the head-wearable device, wherein: the one or more sensors comprise a camera, and the receiving the non-verbal sensor data comprises receiving information determined based on one or more of Simultaneous Localization and Mapping (SLAM) and visual odometry performed via visual data from the camera; classifying audio data corresponding to the audio signal; determining whether the audio signal comprises a pause in the voice activity; responsive to determining that the audio signal comprises the pause in the voice activity, determining, based on the classifying of the audio data and based further on the non-verbal sensor data, whether the pause in the voice activity corresponds to an end point of the voice activity; and responsive to determining that the pause in the voice activity corresponds to the end point of the voice activity, presenting a response to the user based on the voice activity, wherein: the determining whether the audio signal comprises the pause in the voice activity comprises: determining, based on the information determined based on the one or more of SLAM and visual odometry, head poses of the user, and determining whether the head poses comprise a head pose change, wherein the audio signal comprises the pause in accordance with a determination that the head poses comprise the head pose change, and the determining whether the pause in the voice activity corresponds to the end point of the voice activity comprises: determining a probability of interest based on the classifying of the audio data and based further on the non-verbal sensor data, the probability determined based on relative distances between the non-verbal sensor data and its neighbors in an N-dimensional space, and in accordance with a determination that the probability of interest exceeds a threshold, determining that the pause in the voice activity corresponds to the end point of the voice activity.
2. The method of claim 1, further comprising: in accordance with a determination that the probability of interest does not exceed the threshold: determining that the pause in the voice activity does not correspond to the end point of the voice activity, and forgoing presenting the response to the user based on the voice activity.
3. The method of claim 1, wherein the determining whether the audio signal comprises the pause in the voice activity further comprises determining whether an amplitude of the audio signal falls below a second threshold.
4. The method of claim 1 further comprising: in accordance with a determination that the probability of interest does not exceed the threshold: determining that the pause in the voice activity does not correspond to the end point of the voice activity, and determining whether the audio signal comprises a second pause corresponding to the end point of the voice activity.
5. The method of claim 1, wherein the determining whether the audio signal comprises the pause in the voice activity further comprises determining whether the audio signal comprises one or more verbal cues corresponding to the pause in the voice activity.
6. The method of claim 5, wherein the one or more verbal cues comprise a characteristic of the user's prosody.
7. The method of claim 5, wherein the one or more verbal cues comprise a terminating phrase.
8. The method of claim 5, wherein the one or more verbal cues further correspond to the end point of the voice activity.
9. The method of claim 1, wherein the non-verbal sensor data further comprises data indicative of the user's gaze.
10. The method of claim 1, wherein the non-verbal sensor data further comprises data indicative of the user's facial expression.
11. The method of claim 1, wherein the non-verbal sensor data further comprises data indicative of the user's heart rate.
12. The method of claim 1, wherein the determining whether the pause in the voice activity

corresponds to the end point of the voice activity comprises identifying one or more interstitial sounds.

13. The method of claim 1, wherein the one or more sensors comprise the microphone.

14. The method of claim 1, wherein the determining whether the audio signal comprises the pause in the voice activity further comprises determining that a frequency component of the audio signal is indicative of the pause in the voice activity.

15. The method of claim 1, wherein the determining whether the audio signal comprises the pause in the voice activity further comprises determining whether an audio segment of the audio signal comprises the pause in the voice activity.

16. The method of claim 1, wherein the determining whether the pause in the voice activity corresponds to the end point of the voice activity further comprises determining whether an audio segment of the audio signal comprises the end point of the voice activity.

17. A system, comprising: a microphone of a head-wearable device; one or more sensors of the head-wearable device; and one or more processors configured to perform a method comprising: receiving, via the microphone, an audio signal, wherein the audio signal comprises voice activity; receiving, via the one or more sensors, non-verbal sensor data corresponding to a user of the head-wearable device, wherein: the one or more sensors comprise a camera, and the receiving the non-verbal sensor data comprises receiving information determined based on one or more of Simultaneous Localization and Mapping (SLAM) and visual odometry performed via visual data from the camera; classifying audio data corresponding to the audio signal; determining whether the audio signal comprises a pause in the voice activity; responsive to determining that the audio signal comprises the pause in the voice activity, determining, based on the classifying of the audio data and based further on the non-verbal sensor data, whether the pause in the voice activity corresponds to an end point of the voice activity; and responsive to determining that the pause in the voice activity corresponds to the end point of the voice activity, presenting a response to the user based on the voice activity, wherein: the determining whether the audio signal comprises the pause in the voice activity comprises: determining, based on the information determined based on the one or more of SLAM and visual odometry, head poses of the user, and determining whether the head poses comprise a head pose change, wherein the audio signal comprises the pause in accordance with a determination that the head poses comprise the head pose change, and the determining whether the pause in the voice activity corresponds to the end point of the voice activity comprises: determining a probability of interest based on the classifying of the audio data and based further on the non-verbal sensor data, the probability determined based on relative distances between the non-verbal sensor data and its neighbors in an N-dimensional space, and in accordance with a determination that the probability of interest exceeds a threshold, determining that the pause in the voice activity corresponds to the end point of the voice activity.

18. The system of claim 17, further comprising a transmissive display of the head-wearable device, wherein response to the user is presented via the transmissive display.

19. The system of claim 17, wherein the non-verbal sensor data further comprises data indicative of the user's facial expression.

20. A non-transitory computer-readable medium storing instructions, which, when executed by one or more processors, cause the one or more processors to perform a method comprising: receiving, via a microphone, an audio signal, wherein the audio signal comprises voice activity; receiving, via one or more sensors of a head-wearable device, non-verbal sensor data corresponding to a user of the head-wearable device, wherein: the one or more sensors comprise a camera, and the receiving the non-verbal sensor data comprises receiving information determined based on one or more of Simultaneous Localization and Mapping (SLAM) and visual odometry performed via visual data from the camera; classifying audio data corresponding to the audio signal; determining whether the audio signal comprises a pause in the voice activity; responsive to determining that the audio signal comprises the pause in the voice activity, determining, based on the classifying of the audio data and based further on the non-verbal sensor data, whether the pause in the voice activity corresponds to an end point of the voice activity; and responsive to determining that the pause in the voice activity corresponds to the end point of the voice activity, presenting a response to the user based on the voice activity, wherein: the determining whether the audio signal comprises the pause in the voice activity comprises: determining, based on the information determined based on the one or more of SLAM and visual odometry, head poses of the user, and determining whether the head poses comprise a head pose change, wherein the audio signal comprises the pause in accordance with a determination that the head poses comprise the head pose change, and the determining whether the pause in the voice activity corresponds to the end point of the voice activity comprises: determining a probability of interest based on the classifying of the audio data and based further on the non-verbal sensor data, the probability determined based on relative distances between the non-verbal sensor data and its neighbors in an N-dimensional space, and in accordance with a determination that the probability of interest exceeds a threshold, determining that the pause in the voice activity corresponds to the end point of the voice activity.