

SERVER APPARATUS, SYSTEM, AND METHOD FOR PROCESSING DATA

TECHNICAL FIELD

[0001] Various aspects of this disclosure relate to a server apparatus, system, and method for processing data. In some embodiments, the data includes audio data and/or text data.

BACKGROUND

[0002] The following discussion of the background is intended to facilitate an understanding of the present disclosure only. It should be appreciated that the discussion is not an acknowledgment or admission that any of the material referred to was published, known, or is part of the common general knowledge of the person skilled in the art in any jurisdiction as of the priority date of the disclosure.

[0003] An intelligent outbound calling process/system may include one or more cycles of dialogue between a robot and a customer, the customer may be a human customer. The working principle of an intelligent outbound calling robot is based on simulating a sales or customer service expert, interacting with customers through multiple rounds of communication to address a customer's questions. The intelligent outbound calling system may have the advantage of automatic dialing, automatic answering, automatic classification, simultaneous recording of voice and text, and convenient tracking.

[0004] In an ideal situation, the intelligent outbound calling process involves a question-and-answer format, with an answer immediately following each question. However, it may be difficult to achieve such an ideal state in reality. Often, abnormal scenarios such as answering a question before it is asked, providing multiple answers to one question, or not answering a question at all, may occur.

[0005] In addition, existing intelligent outbound calling systems may face a technical problem in that the interruption control of robots, typically initiated by an Interactive Voice Response (IVR) system, automated dialler, or call centre software during the intelligent outbound calling process, may not be well-integrated or well-suited to the specific context or scenario, which may result in insufficient intelligence and accuracy in interruptions and weak adaptability to scenarios.

[0006] Accordingly, the disclosure seeks to provide a server apparatus, system, and method for processing data, to ameliorate the aforementioned drawback at least in part.

SUMMARY

[0007] The disclosure seeks to provide a system, server apparatus, and method for processing data. In some embodiments, the system may form at least part of an intelligent outbound calling robot/platform to provide output with improved accuracy and improved data efficiency.

[0008] This present disclosure seeks to provide an intelligent outbound calling interruption method and system to solve a technical problem of the conventional art where the interruption control of robots during the intelligent outbound calling process may not be well integrated with the scenario, resulting in insufficient intelligence and accuracy in interruptions and weak adaptability to scenarios.

[0009] According to another aspect of the present disclosure, there is provided a server apparatus for processing data, the server apparatus comprises a processor, the processor configured to: obtain, from a voice activity detection (VAD) module, an input data stream over a time period, the input data stream comprising audio data; identify, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet; calculate a first consecutive cumulative number of the at least one non-voice data packet over the time period; calculate a second consecutive cumulative number of the at least one voice data packet over the time period; and compare the first consecutive cumulative number with a first threshold; wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.

[0010] In one embodiment, the processor is further configured to compare the second consecutive cumulative number with a second threshold, and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, the processor is configured to output a status indicative of a speech start event.

[0011] In one embodiment, the processor is further configured to extract the VAD module from a Web Real-Time Communication (WebRTC) module, and transplant the VAD module to a Media Resource Control Protocol (MRCP) module.

[0012] In one embodiment, the input data stream comprises (or may be divided into) a plurality of frames, wherein each of the plurality of frames is associated with a single package data, and wherein the classification algorithm comprises a single package data detection and

analysis module, the single package data detection and analysis module configured to perform one or more of the following: a spectral sub-band division, a Gaussian Mixture Model (GMM), a hypothesis testing, and a maximum likelihood estimation, and wherein the identification of the at least one voice data packet and at least one non-voice data packet is based on a ratio of noise probability to signal probability, or a ratio of signal probability to noise probability.

[0013] In one embodiment, the processor is further configured to provide interruption control of the VAD module based on the status indicative of a speech start event or the status indicative of a silent event.

[0014] In one embodiment, the interruption control comprises switching the VAD module between a plurality of states.

[0015] In one embodiment, the plurality of states comprise an activity state, an inactivity state, an activity transition state, and an inactivity transition state.

[0016] In one embodiment, the status is indicative of the silent event, the processor is configured to execute/perform automatic speech recognition to further determine whether the output of the ASR is indicative of non-speech.

[0017] In one embodiment, where (in a positive determination) the status is indicative of the speech start event, the processor is configured to determine whether the VAD module is in an initial (first) round of dialogue.

[0018] In one embodiment, in a positive determination that the VAD module is not in the initial (first) round of dialogue, the processor is configured to obtain a non-interruptible time parameter.

[0019] In one embodiment, the processor is configured to compare the non-interruptible time parameter with a time node parameter of the speech start event, and wherein where the time node parameter exceeds the non-interruptible time parameter, the processor is configured to provide interruption control of the VAD module based on the status indicative of a speech start event.

[0020] In one embodiment, the interruption control is provided by a software or hardware switch, and wherein the switch comprises a softphone and a soft switch to provide product drivers for voice and chat.

[0021] According to another aspect of the present disclosure there is provided an automated outbound calling system comprising, a voice activity detection (VAD) module, the VAD module operable to receive an input data stream over a time period, the input data stream comprising audio data; an audio data processing module configured to: obtain, from the voice

activity detection (VAD) module, the audio data; identify, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet; calculate a first consecutive cumulative number of the at least one non-voice data packet over the time period; and calculate a second consecutive cumulative number of the at least one voice data packet over the time period; and compare the first consecutive cumulative number with a first threshold; wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.

[0022] In one embodiment, the audio data processing module is configured to compare the second consecutive cumulative number with a second threshold, and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, output a status indicative of a speech start event.

[0023] In one embodiment, the automated outbound calling system further comprises an interruption module, the interruption module configured to provide interruption control to the VAD module based on the status indicative of a speech start event or the status indicative of a silent event.

[0024] In one embodiment, the automated outbound calling system further comprises an extraction module, the extraction module configured to extract the VAD module from a Web Real-Time Communication (WebRTC) module, and transplant the VAD module to a Media Resource Control Protocol (MRCP) module.

[0025] In one embodiment, the interruption module is configured to switch the VAD module between a plurality of states, and wherein the plurality of states comprise an activity state, an inactivity state, an activity transition state, and an inactivity transition state.

[0026] According to another aspect of the present disclosure there is provided a method for processing data using a processor, comprising the steps of: obtaining, from a voice activity detection (VAD) module, an input data stream over a time period, the input data stream comprising audio data; identifying, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet; calculating a first consecutive cumulative number of the at least one non-voice data packet over the time period; calculating a second consecutive cumulative number of the at least one voice data packet over the time period; and comparing the first consecutive cumulative number with a first threshold; wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.

[0027] In one embodiment, the method further comprises comparing the second consecutive cumulative number with a second threshold and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, output a status indicative of a speech start event.

[0028] According to another aspect of the present disclosure there is provided a non-transitory computer-readable medium storing computer executable code comprising instructions for processing data according to any one of the described methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] The disclosure will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

- FIG. 1A is a schematic block diagram of a server apparatus for processing data.
- FIG. 1B is a schematic block diagram comprising the server apparatus of FIG. 1A, and one or more terminal devices, implemented as an automated outbound calling system.
- FIG. 2 is a flow chart depicting an intelligent outbound calling interruption method according to one embodiment.
- FIG. 3A is a flow chart illustrating an embodiment of performing an internal state transition of the intelligent outbound calling interruption method according to an embodiment.
- FIG. 3B is a flow chart illustrating an embodiment of performing an internal state transition of the intelligent outbound calling interruption method according to another embodiment.
- FIG. 4A is a flow chart illustrating an embodiment of controlling interruption through an open-source telephony software for real-time communication protocols according to an embodiment.
- FIG. 4B is a flow chart illustrating an embodiment of controlling interruption through an open-source telephony software for real-time communication protocols according to another embodiment.
- FIG. 5 is a general flow chart of a method for processing data according to the present disclosure.

DETAILED DESCRIPTION

[0030] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the disclosure may be practiced. Other embodiments may be utilized and structural, and logical changes may be made without departing from the scope of the disclosure. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.

[0031] Embodiments, which are non-limiting examples, described in the context of one of the enclosure systems, server devices, or methods are analogously valid for the other systems, devices, or methods. Similarly, embodiments described in the context of a system are analogously valid for a device or a method, and vice-versa.

[0032] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0033] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements. Furthermore, as used in the present disclosure and the appended claims, the term “by” may also mean “from”, depending on the context. Furthermore, as used in the present disclosure and the appended claims, the term “if” may also mean “when” or “upon”, depending on the context. Furthermore, as used in the present disclosure and the appended claims, the words “and/or” may refer to and encompass any and all possible combinations of one or more of the associated listed items.

[0034] As used herein, the term “data” may be understood to include information in any suitable analogue or digital form, for example, provided as a file, a portion of a file, a set of files, a signal or stream, a portion of a signal or stream, a set of signals or streams, waveforms, and the like. The term data, however, is not limited to the aforementioned examples and may take various forms and represent any information as understood in the art.

[0035] As used herein, the term “first”, “second”, etc. are used to distinguish one element/feature from another, and, unless otherwise stated, may not denote order, priority or sequence.

[0036] As used herein, the term “module” refers to, forms part of, or includes an application Specific Integrated Circuit (ASIC); an electronic circuit; a combinational logic circuit; a field programmable gate array (FPGA); a processor (shared, dedicated, or group) that executes code; other suitable hardware components that provide the described functionality; or a combination of some or all of the above, such as in a system-on-chip. The term module may include memory (shared, dedicated, or group) that stores code executed by the processor. A single module or a combination of modules may be regarded as a device. A processor may include one or more modules. For example, multiple modules described in this disclosure may form a processor.

[0037] As used herein, the term “associate”, “associated”, and “associating” indicate a defined relationship (or cross-reference) between two items.

[0038] As used herein, the term “obtain”, in the context of obtaining data, broadly include pull technology used any time a transfer of data is initiated by a request sent from a client to a server. Push technology, on the other hand, is implemented any time a transfer of information is initiated by a server without waiting on a request from a client. In some embodiments, the term obtain may include receive.

[0039] As used herein, “memory” may be understood as a non-transitory computer-readable medium in which data or information can be stored. References to “memory” included herein may thus be understood as referring to volatile or non-volatile memory, including random access memory (“RAM”), read-only memory (“ROM”), flash memory, solid-state storage, magnetic tape, hard disk drive, optical drive, etc., or any combination thereof. Furthermore, it is appreciated that registers, shift registers, processor registers, data buffers, etc., are also embraced herein by the term memory. It is appreciated that a single component referred to as “memory” or “a memory” may be composed of more than one different type of memory, and thus may refer to a collective component including one or more types of memory. It is readily understood that any single memory component may be separated into multiple collectively equivalent memory components, and vice versa. Furthermore, while memory may be depicted as separate from one or more other components (such as in the drawings), it is understood that memory may be integrated within another component, such as on a common integrated chip.

[0040] As used herein, the term “voice activity detection (VAD)” broadly refers to any process or method of distinguishing packets or segments of an audio signal, packet, and/or data that contain human voice from those other segments that consist of silence, noise, or other non-voice or non-speech sounds. The VAD seeks to enable more efficient and accurate processing in applications where recognizing or coding speech specifically is important, often enhancing

the performance of systems that rely on clean speech signals. An example of a VAD module may include one or more of the following steps: receiving an audio signal and dividing it into frames of a specific time duration; extracting features from each frame, which may include the energy, spectral properties, or a combination of any of energy and spectral properties; applying a decision rule or a machine learning model to determine if the frame contains speech or non-speech elements; employing noise suppression techniques to improve the accuracy of the detection.

[0041] As used herein, the term “interruption”, in the context of an intelligent outbound calling system, refers to an event or condition that disrupts or pauses the ongoing automated or semi-automated call process. In some embodiments, an interruption may be initiated by an Interactive Voice Response (IVR) system, automated dialler, or call centre software. The interruption can occur for various reasons and may be designed to make a call-handling process more adaptive, user-friendly, and efficient. In some embodiments, machine learning or sentiment analysis interruption may be used to detect changes in a human user’s tone, emotional state, or word choice that suggest frustration, confusion, or a need for additional support. Such changes may trigger an interruption to transition the call to a live agent or to modify the call flow.

[0042] As used herein, the term “Web Real-Time Communication WebRTC”, refers to a technology that supports real-time voice or video conversations in web browsers, which may be via one or more application programming interfaces. Currently, WebRTC technology may be embedded in browsers, allowing users to engage in real-time communication through the browser without the need for any plugins or software, which is very convenient and flexible.

[0043] As used herein, the term “UniMRCP Single Package Data” is used in the context of real-time audio processing systems like UniMRCP, where audio data is typically handled in small, discrete chunks or frames. These frames can be referred to as "single packages" of data, each representing a short duration of the audio signal (e.g., 10 milliseconds (ms) or 20 milliseconds). The term "single package" refers that each individual frame or packet, or segment of audio data may be processed and analysed separately, rather than handling the entire audio stream at once.

[0044] As used herein, the term “detection and analysis” broadly includes various techniques, algorithms, and/or modules for identifying non-speech and speech data from an audio data. In some example or instance, a WebRTC VAD module may process each single package (or frame) of audio data to detect whether it contains speech or non-speech, the non-

speech may include silence or noise. The analysis part may include extracting features from the audio frame (such as energy levels, spectral properties, etc.) and making a determination about the presence of speech. This decision may be binary, indicating either the presence or absence of speech activity.

[0045] As used herein, the term “transplant” refers to an integration of a WebRTC VAD module into an UniMRCP module to enable frame-based detection and analysis of audio. Each audio frame that UniMRCP handles may be passed to the VAD module for analysis. The VAD module will then return a result (e.g., "speech detected" or "no speech detected") for that single frame. Such transplant may facilitate real-time processing of an input data stream, frame by frame, for the system to make fast decisions about when to engage or disengage certain audio processing functions, such as starting or stopping speech recognition.

[0046] In one example, an UniMRCP module may receive a continuous input data stream, the input data stream comprising audio data from a microphone. The audio stream may be divided into frames (i.e. single packages) of 20 ms each. As each frame arrives, the frame may be sent to the WebRTC VAD module. The VAD module analyses the frame and outputs whether it contains speech. The UniMRCP uses this output to decide whether to continue processing the audio for speech recognition or to pause and wait for more relevant frames. In the context of the present disclosure, a single package data detection and analysis module may refer to the system's ability to detect and analyse speech in real-time, one frame at a time, making the integration efficient and responsive.

[0047] As used herein, the term “configured to” broadly refers to how an apparatus, a system, or a component may be arranged, designed, or programmed to perform a specific function. For example, if a processor is "configured to transmit data," it means the processor comprises the hardware (e.g. circuitry) and software (e.g. code) for performing the data transmission function.

[0048] According to an aspect of the disclosure and with reference to FIG. 1, there is provided a server apparatus for processing data. The server apparatus may be part of an intelligent outbound calling system or robot. In some embodiments, the server apparatus may be configured to approximate a real-life conversation with a human user involving a question-and-answer format, with an answer immediately following each question. In some embodiments, interruption control may be used to control the response of the intelligent outbound calling system or robot to questions raised by a human user.

[0049] The server apparatus may comprise a processor and a memory, the processor is capable of being configured to execute instructions stored in the memory to receive a request. In the embodiment illustrated in FIG. 1A and FIG. 1B, the server apparatus may be a communications server apparatus. The communications server apparatus may be in the form of a server computer 100, the server computer 100 may be a single server as illustrated schematically in FIG. 1A, or have the functionality performed distributed across multiple server components.

[0050] In some embodiments, the server computer 100 includes a communication interface 102. The communication interface 102 may be configured to receive input data stream over a period of time, where the input data stream includes audio data). The communication interface 102 may include a transmitter module and/or a receiver module allowing the server apparatus to communicate over a communications network. The communication interface 102 may include one or more user-interfaces configured to provide users for user control and may include, for example, one or more computing peripheral devices such as display monitors, computer keyboards and the like.

[0051] The server computer 100 may further include a processor in the form of processing unit 104 and a memory 106. The memory 106 may be used by the processing unit 104 to store, for example, the input data stream to be processed.

[0052] As shown in FIG. 1B, the processing unit 104 may be configured to receive, from a terminal device 110, an input data stream 111 for processing. The input data stream 111 may include audio data. In some embodiments, the input data stream 111 may include a multimedia video data file. The server apparatus 100 and one or more terminal devices 110 may be connected via a network 120 to form a system 150 for processing the data. The network 120 may be an Internet or Intranet network. In some embodiments, the system 150 may comprise one or more information or data sources. The data sources may be integrated with the server apparatus/server computer 100.

[0053] In some embodiments, the system 150 may comprise a voice activity detection (VAD) module 170, the VAD module 170 operable to receive the input data stream 111 over a time period, the input data stream comprising audio data. The audio data may be pre-processed by the VAD module 170. In some embodiments, the server apparatus 100 may be or may include an audio data processing module 180, the audio data processing module 180 configured to: obtain, from the VAD module 170, the audio data; identify, from the audio data, using a classification algorithm, at least one voice data packet 112 and/or at least one non-voice

data packet 113; calculate a first consecutive cumulative number 114 of the at least one non-voice data packet 113 over the time period; and compare the first consecutive cumulative number 114 with a first threshold; wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event 115.

[0054] In some embodiments, the audio data processing module 180 may be configured to calculate a second consecutive cumulative number 116 of the at least one voice data packet 112 over the time period; compare the second consecutive cumulative number 116 with a second threshold; and wherein in a positive determination that the second consecutive cumulative number 116 is greater than or equal to the second threshold, output a status indicative of a speech start event 117.

[0055] In some embodiments, the system 150 may further comprise an interruption module 190, the interruption module 190 configured to provide interruption control to the VAD module 170 based on the status indicative of a speech start event 117 or the status indicative of a silent event 115.

[0056] In some embodiments, the VAD module 170 may be extracted from a Web Real-Time Communications (WebRTC) module via an extraction module (not shown). The WebRTC module may be an open-source project that enables real-time voice, text, and video communication capabilities between web browsers and devices. The WebRTC module may be configured to provide software developers with application programming interfaces (APIs), the APIs may be written in JavaScript. The extracted VAD module 170 may be transplanted to a Media Resource Control Protocol (MRCP) module, such as a UniMRCP, to identify the at least one voice data packet and the at least one non-voice data packet.

[0057] In some embodiments, the input data stream may be divided into a plurality of frames, wherein each of the plurality of frames is associated with a single package data. A single package data detection and analysis module 185 may implement the classification algorithm, the single package data detection and analysis module 185 configured to perform one or more of the following classification algorithm: a spectral sub-band division, a Gaussian Mixture Model (GMM), a hypothesis testing, and a maximum likelihood estimation, and wherein the identification of the at least one voice data packet and at least one non-voice data packet is based on a ratio of noise probability to signal probability.

[0058] In some embodiments, the interruption module 190 may be further configured to provide interruption control of the VAD module 170 based on the status indicative of a speech

start event or the status indicative of a silent event. The interruption control may comprise switching the VAD module 170 between a plurality of states 118. The plurality of states 118 may comprise an activity state, an inactivity state, an activity transition state, and an inactivity transition state.

[0059] In some embodiments, where the status is indicative of a silent event, the audio data processing module 180 is configured to execute or perform automatic speech recognition to further determine whether the output of the ASR is indicative of non-speech.

[0060] In some embodiments, where the status is indicative of the speech start event, the audio data processing module 180 is configured to determine whether the VAD module 170 is in an initial (first) round of dialogue. In a positive determination that the VAD module 170 is not in the initial (first) round of dialogue, the audio data processing module 180 is configured to obtain a non-interruptible time parameter.

[0061] In some embodiments, the audio data processing module 180 is configured to compare the non-interruptible time parameter with a time node parameter of the speech start event, and where the time node parameter exceeds the non-interruptible time parameter, the audio data processing module 180 is configured to provide interruption control of the VAD module based on the status indicative of a speech start event.

[0062] In some embodiments, the interruption control may be provided by a software or hardware switch, and wherein the switch comprises a softphone and a soft switch to provide product drivers for voice and chat. In some embodiments, the software or hardware switch may be an open-source telephony software for real-time communication protocols using audio, video, text and other forms of media, such as FreeSwitch™.

[0063] In some embodiments, two or more of the VAD module 170, the audio data processing module 180, the single package data detection and analysis module 185, the interruption module 190, and the extraction module may be integrated with the computer server 100.

[0064] In some embodiments, the voice data may be a human voice data.

[0065] In some embodiments, the interruption control based on the open-source telephony software may be configured to achieve interruption effects, the interruption process endowed with a strong scenario adaptability. The single package data detection and analysis module 185 may be configured to automatically adjust the mean and variance parameters of the Gaussian Mixture Model according to the scenario.

[0066] FIG. 2 shows a flow chart depicting an intelligent outbound calling interruption method 200 according to an embodiment of using the server apparatus 100.

[0067] The method 200 may be performed iteratively on an input data stream over a time period, and includes the steps of:

[0068] Step S201: Extracting the VAD module of WebRTC and transplanting it to UniMRCP to obtain a transplantation result. The UniMRCP is an open-source, cross-platform implementation of Media Resource Control Protocol MRCP clients and servers. The MRCP is a communication protocol used by media resource servers to provide various speech services to clients. Currently defined media resource services may include speech recognition, speech synthesis, recording, speaker identification, and verification. By replacing UniMRCP's built-in VAD module with WebRTC's VAD module, a technical effect of eliminating environmental noise interference, accurately identifying speech packets and non-speech packets, and enhancing adaptability may be achieved.

[0069] The transplantation result may be based on the result obtained from identifying, from the audio data, using the classification algorithm, at least one voice data packet and/or at least one non-voice data packet, that is performed after the transplantation. In other words, the transplantation result is the classification result obtained by transplanting the VAD module of WebRTC to UniMRCP. The VAD module 170 is used in the audio signal processing, such as speech enhancement and speech recognition. The VAD module 170 may be configured to identify speech data packets (or speech segments) and non-speech data packets (or non-speech segments) from an audio data, which may include audio signals. In some embodiments, the output may be a binary 1 when speech is detected, and may be a binary 0 otherwise.

[0070] Step S202: Performing single package data detection and analysis based on the transplantation result to obtain a detection and analysis result.

[0071] Based on the obtained transplantation result, which may be in the form of the binary 1 or binary 0 as described, further single package data detection and analysis can be performed to obtain the detection and analysis result. In some embodiments, a single package data refers to an energy of 160 frames (based on the VAD module 170) with a duration of 20 milliseconds (ms). A classification algorithm comprising a Gaussian normal distribution may be used to analyze the data of each package. The techniques involved for the analysis may include spectral sub band division, Gaussian Mixture Model (GMM), hypothesis testing, and maximum likelihood estimation. Based on the ratio of noise probability to signal probability, a conclusion of a speech based on the audio data and identification of voice data or non-voice data, such as

human voice or non-human voice, is drawn. The single package data detection and analysis module may output noise as 0 or human voice as 1, which forms the detection and analysis result. In some embodiments, the single package data detection and analysis module 185 can automatically adjust the mean and variance parameters of the Gaussian Mixture Model according to the scenario, demonstrating strong scenario adaptability. This may achieve the technical effect of further processing voice signals in package units and clearly identifying whether the single package data represents human voice or non-human voice (noise), laying the foundation for subsequently obtaining a cumulative counting rule.

[0072] Step S203: obtaining or determining a cumulative counting rule.

[0073] Step S204: performing cumulative counting of the detection and analysis result according to the cumulative counting rule to calculate the first consecutive cumulative number of at least one non-voice data packet over the time period, and calculating a second consecutive cumulative number of at least one non-voice data packets over the time period, i.e. to obtain a human voice cumulative result and a non-human voice cumulative result.

[0074] Based on the detection and analysis result obtained from the single package data detection and analysis module 185, the 20ms data of a single package may not be sufficient to determine whether to deploy an interruption control. For example, it may be normal for a 20 ms human voice to suddenly appear within a 1000 ms noise signal. Therefore, cumulative counting needs to be performed separately for human voice or non-human voice signals. The cumulative counting rule involves continuously (or iteratively) accumulating the number of packages for human voice or non-human voice separately. Under the cumulative counting rule, the detection and analysis result may be accumulated over the time period to obtain the human voice cumulative result and the non-human voice cumulative result. The human voice cumulative result may represent a number of single package data where the sound signal is a human voice. The non-human voice cumulative result represents the number of single package data where the sound signal is non-human voice. This achieves the technical effect of clarifying the number of single package data for human voice or non-human voice, providing data support for subsequent internal state transitions.

[0075] Step S205: Performing internal state transition based on the human voice cumulative result and the non-human voice cumulative result to obtain an event judgment result. In some embodiments, the event judgement result may be a binary 1, indicating human voice, and 0, indicating non-human voice.

[0076] The method of performing internal state transition may be made with reference to FIG. 3A and FIG. 3B.

[0077] FIG. 3A is a flow chart illustrating an embodiment of performing an internal state transition of the intelligent outbound calling interruption method according to an embodiment.

[0078] According to the embodiment illustrated in FIG. 3A, the method of determining internal state transition 300A may include:

[0079] Step S301: obtaining the first threshold in the form of a non-human voice cumulative threshold.

[0080] Step S302: determining whether the non-human voice cumulative result meets the non-human voice cumulative threshold. This may be based on a comparison between the non-human voice cumulative result with the first threshold.

[0081] Step S303: wherein when the non-human voice cumulative result meets (i.e. greater than or equal to) the non-human voice cumulative threshold, outputting a status indicative of a silent event, i.e. an event judgment result as a silent event.

[0082] In some embodiments, the obtaining of an event judgement result based on the internal state transition determined by the human voice cumulative result and the non-human voice cumulative result may include using the obtained first threshold, i.e. the non-human voice cumulative threshold, to determine whether the non-human voice cumulative result meets the first threshold. Here, the first non-human voice cumulative threshold refers to the minimum number of consecutive non-human voice package data detected when outputting a status indicating the silent event. For example, at least 50 consecutive packages may be needed for the input stream to be detected as non-human voice to output a silent event, and these 50 non-human voice packages may be the first threshold. Furthermore, when the non-human voice cumulative result meets the first threshold, i.e., when the number of individual non-human voice package data is greater than or equal to the first threshold, the event judgement result is output as a silent event (or the status indicative of a silent event). This achieves the technical effect of scientifically and quickly identifying one or more silent events, this improving work efficiency, and providing data support for starting a next round of dialogue based on a recognition result.

[0083] FIG. 3B is a flow chart illustrating an embodiment of performing an internal state transition of the intelligent outbound calling interruption method according to another embodiment.

[0084] According to the embodiment illustrated in FIG. 3B, the method of determining internal state transition 300B may include:

[0085] Step S311: Obtaining the second threshold, i.e. a human voice cumulative threshold.

[0086] Step S312: Determining whether the human voice cumulative result meets the second threshold.

[0087] Step S313: wherein when the human voice cumulative result meets (i.e. greater or equal) the first human voice cumulative threshold, output the event judgement result as a start-of-speech event or speech start event.

[0088] Step S314: obtaining a state buffer parameter.

[0089] Step S315: performing sensitivity buffering for the internal state transition based on the state buffer parameter to obtain an internal state transition result.

[0090] Step S316: obtaining the event judgement result based on the internal state transition result.

[0091] when obtaining the event judgement result based on the internal state transition determined by the human voice cumulative result and/or the non-human voice cumulative result, the obtained second threshold is used to determine whether the human voice cumulative result meets the second threshold. The second threshold may refer to the minimum number of consecutive human voice packages detected when determining whether the state should be indicative of a start-of-speech of speech start event. For example, at least 15 consecutive data packages need to be detected as human voice to output the speech start event, and these 15 human voice packages will be the second threshold. Furthermore, when the human voice cumulative result meets the first human voice cumulative threshold, i.e., the number of individual human voice package data is greater than or equal to the second threshold, the event judgement result is output as a speech start event and/or a status indicative of the speech start event. This achieves the technical effect of automatically and quickly identifying start-of-speech events, improving work efficiency, and laying the foundation for subsequent interruption processing.

[0092] In obtaining the event judgement result through an internal state transition based on the human voice cumulative result and the non-human voice cumulative result, the state buffer parameter can be obtained. Sensitivity buffering for the internal state transition may be performed through the state buffer parameter to obtain an internal state transition result, and then the event judgement result may be obtained.

[0093] In some embodiments, the internal state transition may refer to the transition between four internal states: an Activity state, an Inactivity state, an Activity transition state, and an Inactivity transition state.

[0094] The activity state may refer to a state that indicates whether the intelligent voice customer service system is actively processing a user's request (the user's request may include the input data stream) or engaging in a conversation. This includes the execution of core functions such as speech recognition, natural language understanding, and response generation.

[0095] The inactivity state indicates that the intelligent voice customer service system is not currently processing a user's request or that the conversation between the user's request and the system is paused. This may occur when there is no user input or when the system is waiting for further instructions from the user.

[0096] The activity transition state represents a process in which the intelligent voice customer service system is transitioning from the activity state to another state, such as the inactivity state. This may happen when a conversation ends, a user's request has been processed, or the system needs to temporarily pause processing.

[0097] The inactivity transition state represents a process in which the intelligent voice customer service system is transitioning from the inactivity state to another state, such as the activity state. This may occur when a user provides a new input data, the system detects unprocessed requests, or the system needs to restart a conversation.

[0098] Transition from the activity state to the inactivity state may occur when a user's request has been processed and the system detects no new user input within a certain defined period. The intelligent voice customer service system may then transition from the activity state to the inactivity state. It is contemplated that the system may enter a standby mode, reducing the utilization of processing resources to save energy and computational power.

[0099] Transition from the inactivity state to the activity state may occur when a user provides new input in the form of an input data stream comprising audio data, or the system detects unprocessed requests. The intelligent voice customer service system then transitions from the inactivity state to the activity state. In this case, the system may be configured to reactivate processing resources to handle the new user request or continue processing unfinished requests.

[00100] During this activity transition state, the system may perform cleanup tasks such as saving the current conversation progress, releasing unnecessary resources, etc. This helps ensure a smooth transition from the current conversation to the next state.

[00101] During this inactivity transition state, the system may perform preparatory work such as loading necessary processing modules, initializing conversation states, etc. This helps ensure a prompt response to new user requests or the continuation of unfinished requests.

[00102] In some embodiments, state buffering may be configured in cases where it was found that the state transition was too sensitive. For example, a lack of human voice signals can lead to a state transition to the activity transition state. A state buffer parameter refers to adding a continuous buffer period between state transitions. When performing an internal state transition, the system may be configured to include a state buffer parameter to accumulate a set duration. If the duration is met, the transition will occur; otherwise, the transition will not occur, ensuring state reliability. The state buffer parameter is added to prevent unreliable states and the generation of meaningless events. The state buffer parameter is related to factors such as interruption effects and working environments. It achieves a good real interruption effect. During the testing process, parameters are continuously adjusted, effects are compared, and working modes are determined. There are slight differences in the parameters corresponding to these modes. When frontend business scenarios change, the technical effect of further enhancing environmental adaptability is achieved through changes in working modes.

[00103] FIG. 4A is a flow chart illustrating an embodiment of a method for controlling interruption through an open-source telephony software for real-time communication protocols according to an embodiment.

[00104] According to the embodiment illustrated in FIG. 4A, the method 400A for controlling interruption through an open-source telephony software for real-time communication protocols may include:

[00105] Step S401: Controlling interruption through FreeSwitch™ based on the event judgement result.

[00106] Step S402: wherein when the event judgement result received by FreeSwitch™ is the silence event, performing an automatic speech recognition (ASR) recognition, using an ASR module, on the detection and analysis result to obtain a recognition result. The recognition result may be a text output of the ASR recognition module.

[00107] Step S403: Starting the next round of dialogue based on the recognition result.

[00108] Based on the obtained human voice cumulative result and the non-human voice cumulative result, the internal state may then be switched. In different states, different internal events may be triggered, leading to the event judgement result. Here, the internal state transition refers to the transition between four internal states of the activity state, the inactivity state, the

activity transition state, and the inactivity transition state. The event judgement result indicates whether the event is a silent event or a speech start event. Furthermore, interruption control is performed through FreeSwitch™ according to the event judgement result. In some embodiments, the FreeSwitch™ provide product drivers for voice and chat. In some embodiments, FreeSwitch™ can be configured for use as a switching engine, Private Branch Exchange, user-level switch (PBX), multimedia gateway, and multimedia server, etc. In some embodiments, FreeSwitch™ may be configured to provide a cross-platform open-source telephone switching platform with strong scalability, designed to provide routing and interconnection communication protocols for audio, video, text, or any other form of media. During a testing process, FreeSwitch™ is used to separately control interruptions for silence events and speech start events, and to compare the interruption effects. When changes occur in the frontend business scenarios, such an approach may further enhance environmental adaptability and achieves good technical effects for real interruption control.

[00109] When FreeSwitch™ performs interruption control based on the event judgement result, and when the event judgement result received by FreeSwitch™ is the silent event, the recognition result can be obtained by performing ASR recognition on the detection and analysis result. Here, FreeSwitch™ is a software-based telephone switching solution that includes a softphone and a soft switch to provide product drivers for voice and chat. The detection and analysis result refers to the use of Gaussian normal distribution to detect the data of each single package, and based on the ratio of noise probability to signal probability, a conclusion of human voice or non-human voice is given, with individual Package data outputting noise as 0 or human voice as 1. ASR recognition refers to Automatic Speech Recognition technology, which is a technology that converts sound into text. The recognition result refers to the text information obtained from the detection and analysis result. Furthermore, the next round of dialogue is initiated based on the recognition result, i.e., when the event judgement result received by FreeSwitch™ is the silence event, and the recognition result obtained is blank text information, thus confirming that it is indeed a silent event, the next round of dialogue may be initiated. This achieves the technical effect of further accurately judging silent events through ASR recognition, thereby quickly initiating the next round of dialogue.

[00110] FIG. 4B is a flow chart illustrating an embodiment of controlling interruption through an open-source telephony software for real-time communication protocols according to another embodiment.

[00111] According to the embodiment illustrated in FIG. 4B, the method 400B may include:

[00112] Step S411: When the event judgment result received by FreeSwitch™ is the speech start event, determining whether the current dialogue of the intelligent outbound calling system or robot is the first round of dialogue.

[00113] Step S412: When the current dialogue of the robot is not the first round of dialogue, obtain a non-interruptible time parameter.

[00114] Step S413: When the time node parameter of the start-of-speech event does not satisfy the non-interruptible time parameter, perform interruption processing on the current playback.

[00115] Step S414: If the current dialogue of the robot is the first round of dialogue, then do not perform interruption processing.

[00116] When controlling interruption based on the event judgement result through FreeSwitch™, when the event judgement result received by FreeSwitch™ is the start-of-speech event, it is determined whether the current dialogue of the robot is the first round of dialogue. If the current dialogue of the robot is the first round of dialogue, interruption processing is not performed, meaning that the first round cannot be interrupted. For example, assuming that the intelligent outbound calling system or robot requires 10 seconds to play the first round of dialogue, the non-interruptible time for the first round is 10 seconds. Before the robot completes playing the first round of dialogue, single Package data detection and analysis are not performed. In actual operations, the intelligent outbound calling system or robot may express its complete intent at the first opportunity, and the first round of dialogue cannot be interrupted. This achieves the technical effect of expressing a complete intent to the user and enhancing the user experience.

[00117] When the event judgement result received by FreeSwitch™ indicates a speech start event, and it is determined whether the current dialogue of the robot is the first round of dialogue, if the current dialogue of the robot is not the first round of dialogue, the non-interruptible time parameter can be obtained. In some embodiments, the non-interruptible time parameter may refer to the time range within which interruption processing cannot be performed in other rounds after the first round of dialogue. For example, interruption is not allowed within the first 3 seconds of subsequent rounds of dialogue between a user and the intelligent outbound calling system or robot. That is, for other rounds, in the first 3 seconds before the intelligent outbound calling system or robot starts playing the dialogue, interruption processing and single package data detection and analysis cannot be performed on it. Furthermore, when the time node parameter of the start-of-speech event does not satisfy the

non-interruptible time parameter, meaning that the time node parameter of the start-of-speech event is not within the first non-interruptible time, interruption processing is performed. The non-interruptible time can be controlled at the dialogue flow node, with the first node of the flow set as the playback duration, and other nodes can be flexibly set. This achieves the technical effect of flexibly setting non-interruptible time, preventing frequent interruptions, and enhancing user service satisfaction.

[00118] FIG. 5 is a general flow chart illustrating a method 500 for processing data. The method 500 may be stored in a non-transitory computer-readable medium as computer executable code comprising instructions for processing data.

[00119] The method 500 may comprise the steps of:

[00120] Step S501: obtaining, from a voice activity detection (VAD) module, an input data stream over a time period, the input data stream comprising audio data.

[00121] Step S502: identifying, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet.

[00122] Step S503: calculating a first consecutive cumulative number of the at least one non-voice data packet over the time period.

[00123] Step S504: comparing the first consecutive cumulative number with a first threshold.

[00124] wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold;

[00125] Step S505: outputting a status indicative of a silent event.

[00126] In some embodiments, the method may further comprise the steps of:

[00127] Step S506: calculating a second consecutive cumulative number of the at least one voice data packet over the time period; and

[00128] Step S507: comparing the second consecutive cumulative number with a second threshold.

[00129] and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold,

[00130] Step S508: outputting a status indicative of a speech start event.

[00131] It is appreciable that steps S503 to S508 may be performed in any order of simultaneously.

[00132] The steps of the disclosed method are not restricted to the sequence set forth in the claims unless explicitly required. Steps may be performed in a different order, concurrently, or omitted entirely, depending on the embodiment.

[00133] This present disclosure addresses the technical issues in the prior art where interruption control in intelligent outbound call processes cannot be well integrated with scenarios, resulting in insufficient intelligence and accuracy in interruptions and weak scenario adaptability. By extracting the VAD module from WebRTC, transplanting it to UniMRCP, and combining it with FreeSwitchTM to achieve interruption effects, the interruption process is endowed with strong scenario adaptability. The single-package detection module can automatically adjust the mean and variance parameters of a probabilistic model, such as a Gaussian Mixture Model, according to application. At the same time, it achieves good real interruption effects. During the testing process, parameters may be continuously adjusted, effects are compared, and working modes may be determined. There are slight differences in the parameters corresponding to these modes. When frontend business scenarios change, the technical effect of further enhancing environmental adaptability is achieved through changes in working modes.

[00134] It may be appreciable that the server apparatus may be used in various applications, potentially in any applications involving speech detection and analysis, such as customer service robots, intelligent dialogue robot systems, etc.

[00135] The methods described herein may be performed and the various processing or computation units and the devices and computing entities described herein may be implemented by one or more circuits. In an embodiment, a "circuit" may be understood as any kind of a logic implementing entity, which may be hardware, software, firmware, or any combination thereof. Thus, in an embodiment, a "circuit" may be a hard-wired logic circuit or a programmable logic circuit such as a programmable processor, e.g. a microprocessor. A "circuit" may also be software being implemented or executed by a processor, e.g. any kind of computer program, e.g. a computer program using a virtual machine code. Any other kind of implementation of the respective functions which are described herein may also be understood as a "circuit" in accordance with an alternative embodiment.

[00136] While the disclosure has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the disclosure as defined by the appended claims. The scope of the disclosure is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

CLAIMS

1. A server apparatus for processing data, the server apparatus comprises a processor, the processor configured to:
 - obtain, from a voice activity detection (VAD) module, an input data stream over a time period, the input data stream comprising audio data;
 - identify, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet;
 - calculate a first consecutive cumulative number of the at least one non-voice data packet over the time period;
 - calculate a second consecutive cumulative number of the at least one voice data packet over the time period; and
 - compare the first consecutive cumulative number with a first threshold;wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.
2. The server apparatus of claim 1, wherein the processor is further configured to compare the second consecutive cumulative number with a second threshold, and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, the processor is configured to output a status indicative of a speech start event.
3. The server apparatus of claim 1 or 2, wherein the processor is further configured to extract the VAD module from a Web Real-Time Communication (WebRTC) module, and transplant the VAD module to a Media Resource Control Protocol (MRCP) module.
4. The server apparatus of claim 3, wherein the input data stream comprises a plurality of frames, wherein each of the plurality of frames is associated with a single package data, and wherein the classification algorithm comprises a single package data detection and analysis module, the single package data detection and analysis module configured to perform one or more of the following: a spectral sub-band division, a Gaussian Mixture Model (GMM), a hypothesis testing, and a maximum likelihood estimation, and wherein the identification of the at least one voice data packet and at least one non-voice data packet is based on a ratio of noise probability to signal probability.

5. The server apparatus of any one of claims 2 to 4, wherein the processor is further configured to provide interruption control of the VAD module based on the status indicative of the speech start event or the status indicative of the silent event.
6. The server apparatus of claim 5, wherein the interruption control comprises switching the VAD module between a plurality of states.
7. The server apparatus of claim 6, wherein the plurality of states comprise an activity state, an inactivity state, an activity transition state, and an inactivity transition state.
8. The server apparatus of any one of the preceding claims, wherein where the status is indicative of the silent event, the processor is configured to perform automatic speech recognition (ASR) to further determine whether the output of the ASR is indicative of non-speech.
9. The server apparatus of any one of the preceding claims, wherein where the status is indicative of the speech start event, the processor is configured to determine whether the VAD module is in an initial round of dialogue.
10. The server apparatus of claim 9, wherein in a positive determination that the VAD module is not in the initial round of dialogue, the processor is configured to obtain a non-interruptible time parameter.
11. The server apparatus of claim 10, wherein the processor is configured to compare the non-interruptible time parameter with a time node parameter of the speech start event, and wherein where the time node parameter exceeds the non-interruptible time parameter, the processor is configured to provide interruption control of the VAD module based on the status indicative of a speech start event.
12. The server apparatus of claim 5, wherein where the interruption control is provided by a switch, and wherein the switch comprises a softphone and a soft switch to provide product drivers for voice and chat.

13. An automated outbound calling system comprising,
 - a voice activity detection (VAD) module, the VAD module operable to receive an input data stream over a time period, the input data stream comprising audio data;
 - an audio data processing module configured to:
 - obtain, from the voice activity detection (VAD) module, the audio data;
 - identify, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet;
 - calculate a first consecutive cumulative number of the at least one non-voice data packet over the time period; and
 - calculate a second consecutive cumulative number of the at least one voice data packet over the time period; and
 - compare the first consecutive cumulative number with a first threshold;
 - wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.
14. The automated outbound calling system of claim 13, wherein the audio data processing module is configured to compare the second consecutive cumulative number with a second threshold, and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, output a status indicative of a speech start event.
15. The automated outbound calling system of claim 13 or 14, further comprising an interruption module, the interruption module configured to provide interruption control to the VAD module based on the status indicative of a speech start event or the status indicative of a silent event.
16. The automated outbound calling system of any one of claims 13 to 15, further comprising an extraction module, the extraction module configured to extract the VAD module from a Web Real-Time Communication (WebRTC) module, and transplant the VAD module to a Media Resource Control Protocol (MRCP) module.

17. The automated outbound calling system of claim 15 or 16, wherein the interruption module is configured to switch the VAD module between a plurality of states, and wherein the plurality of states comprise an activity state, an inactivity state, an activity transition state, and an inactivity transition state.
18. A method for processing data using a processor, comprising the steps of:
 - obtaining, from a voice activity detection (VAD) module, an input data stream over a time period, the input data stream comprising audio data;
 - identifying, from the audio data, using a classification algorithm, at least one voice data packet and/or at least one non-voice data packet;
 - calculating a first consecutive cumulative number of the at least one non-voice data packet over the time period;
 - calculating a second consecutive cumulative number of the at least one voice data packet over the time period; and
 - comparing the first consecutive cumulative number with a first threshold;wherein in a positive determination that the first consecutive cumulative number is greater than or equal to the first threshold, output a status indicative of a silent event.
19. The method of claim 18, further comprises comparing the second consecutive cumulative number with a second threshold and wherein in a positive determination that the second consecutive cumulative number is greater than or equal to the second threshold, output a status indicative of a speech start event.
20. A non-transitory computer-readable medium storing computer executable code comprising instructions for processing data according to the method of claim 18 or 19.