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

(11) **EP 3 504 861 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
07.10.2020 Bulletin 2020/41

(21) Application number: **17762298.2**

(22) Date of filing: **25.08.2017**

(51) Int Cl.:
G10L 21/043 *(2013.01)* ***H04L 29/08*** *(2006.01)*
G10L 25/78 *(2013.01)*

(86) International application number:
PCT/US2017/048584

(87) International publication number:
WO 2018/039547 (01.03.2018 Gazette 2018/09)

(54) AUDIO TRANSMISSION WITH COMPENSATION FOR SPEECH DETECTION PERIOD DURATION

AUDIOÜBERTRAGUNG MIT KOMPENSATION DER ZEITDAUER DER SPRACHERKENNUNG

TRANSMISSION AUDIO À COMPENSATION POUR LA DURÉE DE LA PÉRIODE DE DÉTECTION DE VOIX

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

(30) Priority: **25.08.2016 US 201615246950**

(43) Date of publication of application:
03.07.2019 Bulletin 2019/27

(60) Divisional application:
20192093.1

(73) Proprietor: **Google LLC**
Mountain View, CA 94043 (US)

(72) Inventors:
- **KAY, Erik**
  Mountain View, California 94043 (US)
- **LINDBERG, Jonas Erik**
  111 22 Stockholm (SE)
- **LACHAPELLE, Serge**
  111 22 Stockholm (SE)
- **LUNDIN, Henrik**
  111 22 Stockholm (SE)

(74) Representative: **Betten & Resch**
Patent- und Rechtsanwälte PartGmbB
Maximiliansplatz 14
80333 München (DE)

(56) References cited:
EP-A1- 1 750 397
US-A1- 2008 281 586
US-B1- 7 016 850

EP 3 504 861 B1

Description

BACKGROUND

[0001] Audio communication sessions, such as voice over Internet protocol (VoIP) sessions, can involve two or more users providing audio inputs to their computing devices and the devices exchanging encoded audio packets indicative of the audio inputs via a network. Upon receipt, the audio packets are decoded to obtain an audio signal, which can be output by the receiving computing device via a speaker. In some cases, the playback of received audio can be captured by a microphone of the listening computing device, such as during a period when the listening user is not actively speaking. This captured playback can then be transmitted and output at the other computing device, which is also known as echo.

[0002] Document US 7 016 850 B1 refers to incoming speech frames that are buffered, wherein a pitch period for a current portion of the signal is estimated, and then a pitch period of the signal is cut from that portion. This is continued until the original access delay, as estimated from the time lag between the commencement of voice input for the talkspurt, and notification that a channel is available, is eliminated. The remainder of the talkspurt is then transmitted without such compression.

[0003] Document US 2008/281586 A1 refers to a "speech onset detector" that provides a variable length frame buffer in combination with either variable transmission rate or temporal speech compression for buffered signal frames. The variable length buffer buffers frames that are not clearly identified as either speech or non-speech frames during an initial analysis. Buffering of signal frames continues until a current frame is identified as either speech or non-speech. If the current frame is identified as non-speech, buffered frames are encoded as non-speech frames. However, if the current frame is identified as a speech frame, buffered frames are searched for the actual onset point of the speech. Once that onset point is identified, the signal is either transmitted in a burst, or a time-scale modification of the buffered signal is applied for compressing buffered frames beginning with the frame in which onset point is detected.; The compressed frames are then encoded as one or more speech frames.

[0004] Document EP 1 750 397 A1 refers to preventing a receiving buffer from becoming empty by: storing received packets in the receiving buffer; detecting the largest arrival delay jitter of the packets and the buffer level of the receiving buffer by a state detecting part; obtaining an optimum buffer level for the largest delay jitter using a predetermined table by a control part; determining, based on the detected buffer level and the optimum buffer level, the level of urgency about the need to adjust the buffer level; expanding or reducing the waveform of a decoded audio data stream of the current frame decoded from a packet read out of the receiving buffer by a consumption adjusting part to adjust the consumption of reproduction frames on the basis of the urgency level, the detected buffer level, and the optimum buffer level.

SUMMARY

[0005] The present invention is defined in the independent claims. Preferred embodiments are defined in the dependent claims. A computer-implemented method, a first computing device, and a computer-readable medium are presented. The first computing device can include one or more processors and a non-transitory memory storing a set of instructions that, when executed by the one or more processors, causes the first computing device to perform operations. The computer-readable medium can also have the set of instructions stored thereon that, when executed by the one or more processors of the first computing device, causes the first computing device to perform the operations.

[0006] The method and the operations can include obtaining, by the first computing device, an audio input signal for an audio communication session with a second computing device using audio data captured by a microphone of the first computing device; analyzing, by the first computing device, the audio input signal to detect a speech input by a first user associated with the first computing device; determining, by the first computing device, a duration of a detection period from when the audio input signal was obtained until the analyzing has completed; transmitting, from the first computing device and to the second computing device, (i) a portion of the audio input signal beginning at a start of the speech input and (ii) the detection period duration, wherein receipt of the portion of the audio input signal and the detection period duration causes the second computing device to accelerate playback of the portion of the audio input signal to compensate for the detection period duration; analyzing, by the first computing device, the audio input signal to detect an end of the speech input by the first user; and terminating transmission, from the first computing device to the second computing device, of the portion of the audio input signal at a point corresponding to the detected end of the speech input by the first user.

[0007] In some embodiments, the method and the operations can further include encoding, by the first computing device, the portion of the audio input signal to obtain a set of audio packets, wherein the transmitting includes transmitting, to the second computing device, (i) the set of audio packets and (ii) the detection period duration.

[0008] In some embodiments multiple pitch periods are removed, in particular multiple pitch periods having a length of less than 15 milliseconds, in particular less than 7.5 milliseconds. This limiting can prevent buffering problems.

[0009] In some embodiments, receipt of the set of audio packets and the detection period duration causes the second computing device to: decode the set of audio packets to obtain an audio output signal; remove a re-

dundant portion of the audio output signal corresponding to one or more pitch periods to obtain the modified audio output signal, wherein the modified output signal has a shorter length than the audio output signal; and output, by a speaker of the second computing device, the modified audio output signal. In some embodiments, a quantity of the one or more removed pitch periods corresponds to the detection period duration. In some embodiments, receipt of the set of audio packets and the detection period duration causes the second computing device to remove the redundant portion of the audio output signal by: cross-correlating the audio output signal with itself to obtain an autocorrelation signal; and detecting one or more peaks of the autocorrelation signal that exceed a threshold indicative of the one or more pitch periods of the audio output signal.

[0010] In a further embodiment the threshold is in the range of 0.9 to 0.3, in particular in the range of 0.6 and 0.45, more particular 0.5, wherein lower thresholds can yield increased speed.

[0011] In some embodiments, analyzing the audio input signal to detect the speech input includes applying a voice activity detection (VAD) technique to the audio input signal, the VAD technique having an aggressiveness or accuracy that corresponds to the detection period duration. In some embodiments, applying the voice detection technique to the audio input signal includes distinguishing the speech input by the first user from speech by the second user within the audio input signal.

[0012] Further areas of applicability of the present disclosure will become apparent from the detailed description provided hereinafter. It should be understood that the detailed description and specific examples are intended for purposes of illustration only and are not intended to limit the scope of the invention, as defined by the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The present disclosure will become more fully understood from the detailed description and the accompanying drawings, wherein:

FIG. 1 is a diagram of a computing system including an example computing device according to some implementations of the present disclosure;

FIG. 2 is a functional block diagram of the example computing device of FIG. 1; and

FIG. 3 is a flow diagram of an example technique for decreasing echo and transmission periods for audio communication sessions according to some implementations of the present disclosure.

DETAILED DESCRIPTION

[0014] During audio communication sessions, such as voice over Internet protocol (VoIP) sessions, audio packets are received and decoded to obtain an audio signal, which is output by the receiving computing device via a speaker. In some cases, the playback of received audio can be captured by a microphone of the computing device, such as during a period when the listening user is not actively speaking. This captured audio playback can then be transmitted to and output by the other computing device, which is also known as echo. To minimize echo, echo suppression or echo cancellation techniques can be used. Echo cancellation, for example, can involve identifying the output audio signal that is output by the speaker and then detecting and removing the output audio signal from the input audio signal captured by the microphone. These techniques, however, do not work in certain environments, e.g., noisy areas. As a result, users may have to manually mute the microphones of their computing devices while they are not talking. Other techniques aim to detect local speech and only transmit audio when the user is actively speaking, but this can result in clipping of the user's speech and/or the audio becoming out-of-sync with corresponding video, e.g., for a video chat session.

[0015] Accordingly, improved techniques are presented for decreasing echo and transmission periods for audio communication sessions. The techniques begin by detecting when a user is speaking (e.g., using a voice activity detection, or VAD technique). For accuracy purposes, there may be a slight lag (e.g., one hundred milliseconds or more) associated with the VAD technique. This analysis of this audio input can also involve distinguishing between the local user's speech and filtering out or ignoring speech by the other user (e.g., captured echo). Once speech is detected, the computing device can begin transmission. This transmission can include a portion of the speech input back to the point when speech was initially detected. The other computing device can receive the transmitted audio, but there will be a synchronization gap, e.g., due to the VAD technique.

[0016] The transmitting computing device, therefore, can also calculate and transmit information indicative of a duration of the delay period. This can be used by the receiving device to rapidly regain sync without clipping any of the audio. Instead of speeding up the audio playback, which is done by conventional techniques and can undesirably affect the pitch, the receiving computing device can detect and remove a redundant portion of the audio output signal corresponding to one or more pitch periods before playback. Removing one or more pitch periods results in faster playback, because the length of the audio signal is shortened, but without any undesirable pitch modification.

[0017] One technical problem being solved is echo prevention. As mentioned above, this echo can occur due to audio playback being captured by the listening computing device and transmitted back to the originating computing device. The technical advantages of these techniques include not requiring the user to actively control the microphone/speakers to avoid echo. Another technical problem being solved is audio synchronization

after a delay without affecting the audio pitch. As mentioned above, conventional techniques accelerate audio playback, which affects the pitch and is undesirable to the listening user. The technical advantages of these techniques, therefore, include fast audio playback synchronization after a delay without affecting the audio pitch.

[0018] Referring now to FIG. 1, a diagram of an example computing network 100 is illustrated. The computing network 100 can include a first computing device 104 that can communicate with a second computing device 108 via a network 112. While mobile phone configurations of the computing devices 104, 108 are illustrated, it will be appreciated that the first and second computing devices 104, 108 can be any suitable computing devices configured for communication via the network 112 (desktop computers, laptop computers, tablet computers, etc.). The network 112 can be a cellular network (2G, 3G, 4G long term evolution (LTE), etc.), a computing network (local area network, the Internet, etc.), or some combination thereof. A server computing device 116 can also communicate via the network 112. For example, the server computing device 116 could coordinate the audio communication session (e.g., a voice over Internet protocol (VoIP) session) between the first and second computing devices 104, 108.

[0019] This audio communication session could be established, for example, in response to inputs from users 120, 124 at one or both of the first and second computing devices 104, 108. For example, the second user 124 may provide an input at the second computing device 108 to call the first user 120 (an audio communication session request), which could then be accepted by the first user 120 via another input at the first computing device 104, thereby establishing the audio communication session. During the audio communication session, audio packets corresponding to audio inputs (e.g., from users 120, 124) can be exchanged via the server computing device 116 between the first and second computing devices 104, 108. While the first computing device 104 is described as receiving audio data packets from the second computing device 108, it will be appreciated that the first computing device 104 can also transmit audio packets to the second computing device 108.

[0020] The term "audio communication session" as used herein can refer to either an audio-only communication session or an audio/video communication session. Further, while the techniques herein are described as being implemented at one of the first and second computing devices 104, 108 that is receiving the audio packets (the receiving device), it will be appreciated that at least a portion of these techniques could be implemented at the server computing device 116. More particularly, when the server computing device 116 is coordinating the audio communication session, the audio packets can flow through the server computing device 116. For example, the server computing device 116 could have a queue of audio packets and could perform at least a portion of these techniques, such as decoding, compressing, and then re-encoding for transmission to the receiving device, which could then merely decode and playback upon receipt.

[0021] Referring now to FIG. 2, a functional block diagram of an example computing device 200 is illustrated. The computing device 200 can represent the configurations of the first and second computing devices 104, 108. It will be appreciated that the server computing device 116 could also have the same or similar configuration as the computing device 200. The computing device 200 can include a communication device 204 (e.g., a wireless transceiver) configured for communication via the network 112. A processor 208 can be configured to control operation of the computing device 200. The term "processor" as used herein can refer to both a single processor and two or more processors operating in a parallel or distributed architecture. A memory 212 can be any suitable storage medium (flash, hard disk, etc.) configured to store information at the computing device 200. In one implementation, the memory 212 can store instructions executable by the processor 208 to cause the computing device 200 to perform at least a portion of the disclosed techniques.

[0022] The computing device 200 can also include a microphone 216 configured to capture audio input and a speaker 220 configured to generate audio output. The microphone 216 can be any suitable acoustic-to electric transducer or sensor that converts sound into an electrical signal. This can include speech (e.g., by users 120, 124) as well as other noise, such as background noise. The captured audio data (e.g., an analog signal) is then digitized and converted to an audio input signal (e.g., a digital signal). This audio input signal can be encoded into audio packets for transmission via the network 112. Received audio packets can be decoded into an audio output signal. The audio output signal can be provided to the speaker 220, which in turn can produce audible sound corresponding to the audio output signal. The speaker 220 can include a set of electroacoustic transducers that convert an electrical signal into a corresponding sound. While not shown, it will be appreciated that the computing device 200 can include other suitable components, such as a display (a touch display), physical buttons, a camera, and the like.

[0023] Once the audio communication session is established between the first and second computing devices 104, 108, audio information can be exchanged. The first computing device 104 can capture audio information using its microphone 216 to obtain an audio input signal. The first computing device 104 can then analyze the audio input signal to detect a speech input by the first user 120, such as by applying speech detection (e.g., a VAD technique) on the audio input signal. To achieve a desired accuracy, the VAD technique may have a slight delay associated therewith (e.g., a few hundred milliseconds). This delay period, also referred to herein as a detection period, can be described as having a duration that cor-

responds to an aggressiveness or accuracy of the VAD technique. In other words, this period represents a delay from when the audio input signal is obtained to a point where the speech input is detected.

[0024] Once the speech input is detected in the audio input signal, the first computing device 104 can identify a portion of the audio input signal beginning at the point of the detected speech. The first computing device 104 can then encode audio data packets corresponding to this identified portion of the audio input signal. The first computing device 104 can transmit these encoded audio data packets to the second computing device 108, along with information indicative of the detection period duration. This information relating to the detection period duration could also be included in encoded data packets. No audio information, however, is transmitted prior to these encoded audio data packets. By transmitting only the portion of the audio input signal beginning with the speech input, echo can be decreased or eliminated without using an echo canceler or suppresser.

[0025] The first computing device 104 can also analyze the audio input signal to determine an end of the speech input by the first user 120. Once the end of this speech input has been detected, the first computing device 104 can terminate transmission of the portion of the audio input signal to the second computing device 108. The transmission termination point can be a particular point in the audio input signal that corresponds to the detected end of the speech input. The first computing device 104 can then continue analyzing the audio input signal to detect a next occurrence of a speech input by the first user 120, after which transmission to the second computing device 108 can resume according to the techniques herein.

[0026] The second computing device 108 can receive the encoded audio packets and can decode the encoded audio packets to obtain an audio output signal. The second computing device 108 can also receive the information indicative of the detection period duration and can process it accordingly to obtain the detection period duration. The second computing device 108 can then accelerate playback of the audio output signal to compensate for the determined detection period. This acceleration of the audio playback can include compressing (e.g., removing a redundant portion of) the audio output signal and then outputting the modified audio output signal. In some implementations, a quantity of the one or more removed pitch periods corresponds to the detection period duration. After the pitch period(s) are removed, the second computing device 108 has a modified audio output signal having a shorter duration than the original audio output signal, which results in accelerated playback.

[0027] In some implementations, the second computing device 108 can utilize signal correlation to identify one or more pitch periods for removal. More particularly, the second computing device 108 can cross-correlate the audio output signal with itself to obtain an autocorrelation signal. Autocorrelation, cross-autocorrelation, and serial correlation all refer to the process of cross-correlating a signal with itself at different temporal points. The autocorrelation signal represents a similarity of samples as a function of a time gap between them and it can be used for identifying the presence of a periodic signal obscured by noise. Specifically, the second computing device 108 can identify a peak in the autocorrelation signal, which represents a strong periodicity in the audio output signal. This identification can be performed using a threshold. For example only, a threshold of approximately 0.5 can be used. In contrast, a straightforward accelerated playback technique might use a threshold of approximately 0.9. It will be appreciated that any suitable threshold may be used, but lower thresholds will generally provide for increased speed.

[0028] Specifically, the lower threshold of approximately 0.5 increases speed (e.g., up to -15%) while making little if any difference on the quality of the modified audio output signal. The location of this peak can also represent a pitch period of the audio input signal (i.e., a pitch period of the speech). The second computing device 108 can then remove at least one of the pitch periods from the audio output signal to obtain a modified audio output signal. In some implementations, multiple pitch periods could be removed, but the length of the multiple pitch periods could be limited to a certain size (e.g., less than 7.5 milliseconds) to avoid potential buffering problems. Various combinations of the above could also be implemented: lower correlation threshold only, removal of multiple pitch periods, or both. The results can include up to 25% increased speed compared to straightforward playback acceleration techniques, while not having a negative effect on audio output pitch. The effective accelerate rate is increased to between 50% and 90%, depending on the audio input signal, which translated to reducing buffer delay by 500ms to 900ms.

[0029] Referring now to FIG. 3, a flow diagram of an example technique 300 for decreasing echo and transmission periods for audio communication sessions is shown. At 304, an audio communication session (VoIP, video chat, etc.) can be established (e.g., by the server computing device 116) between the first computing device 104 and the second computing device 108. At 308, the first computing device 104 can obtain an audio input signal for the audio communication session based on audio data captured by its microphone 216. At 312, the first computing device 104 can analyze the audio input signal to detect a speech input by the first user 120. At 316, the first computing device 104 can determine a duration of a detection period from when the audio input signal is obtained to when the analyzing has completed. At 320, the first computing device 104 can transmit, to the second computing device 108, the portion of the audio input signal (e.g., encoded audio packets) and the detection period duration.

[0030] At 324, the first computing device 104 can analyze the audio input signal to detect an end of the speech input by the first user 120. If the end is not detected, the

technique 300 can return to 324. If the end is detected, however, the technique 300 can proceed to 328 where the first computing device 104 can terminate transmission of the portion of the audio input signal at an appropriate point. The technique 300 can then end or return to 304. As previously discussed herein, receipt of the portion of the audio input signal and the detection period duration causes the second computing device 108 to accelerate playback of the portion of the audio input signal to compensate for the detection period duration, e.g., by removing a redundant portion of the audio output signal corresponding to one or more pitch periods to obtain a modified audio output signal for output by its speaker 220.

[0031] One or more systems and methods discussed herein do not require collection or usage of user personal information. In situations in which certain implementations discussed herein may collect or use personal information about users (e.g., user data, information about a user's social network, user's location and time, user's biometric information, user's activities and demographic information), users are provided with one or more opportunities to control whether the personal information is collected, whether the personal information is stored, whether the personal information is used, and how the information is collected about the user, stored and used. That is, the systems and methods discussed herein collect, store and/or use user personal information only upon receiving explicit authorization from the relevant users to do so. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. As one example, a user's identity may be treated so that no personally identifiable information can be determined. As another example, a user's geographic location may be generalized to a larger region so that the user's particular location cannot be determined.

[0032] Example embodiments are provided so that this disclosure will be thorough. Numerous specific details are set forth such as examples of specific components, devices, and methods, to provide a thorough understanding of embodiments of the present disclosure. It will be apparent to those skilled in the art that specific details need not be employed, that example embodiments may be embodied in many different forms and that neither should be construed to limit the scope of the invention, as defined by the appended claims. In some example embodiments, well-known procedures, well-known device structures, and well-known technologies are not described in detail.

[0033] The terminology used herein is for the purpose of describing particular example embodiments only and is not intended to be limiting. As used herein, the singular forms "a," "an," and "the" may be intended to include the plural forms as well, unless the context clearly indicates otherwise. The term "and/or" includes any and all combinations of one or more of the associated listed items. The terms "comprises," "comprising," "including," and "having," are inclusive and therefore specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. The method steps, processes, and operations described herein are not to be construed as necessarily requiring their performance in the particular order discussed or illustrated, unless specifically identified as an order of performance. It is also to be understood that additional or alternative steps may be employed.

[0034] Although the terms first, second, third, etc. may be used herein to describe various elements, components, regions, layers and/or sections, these elements, components, regions, layers and/or sections should not be limited by these terms. These terms may be only used to distinguish one element, component, region, layer or section from another region, layer or section. Terms such as "first," "second," and other numerical terms when used herein do not imply a sequence or order unless clearly indicated by the context. Thus, a first element, component, region, layer or section discussed below could be termed a second element, component, region, layer or section without departing from the teachings of the example embodiments.

[0035] As used herein, the term module may refer to, be part of, or include: an Application Specific Integrated Circuit (ASIC); an electronic circuit; a combinational logic circuit; a field programmable gate array (FPGA); a processor or a distributed network of processors (shared, dedicated, or grouped) and storage in networked clusters or datacenters that executes code or a process; other suitable 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 also include memory (shared, dedicated, or grouped) that stores code executed by the one or more processors.

[0036] The term code, as used above, may include software, firmware, byte-code and/or microcode, and may refer to programs, routines, functions, classes, and/or objects. The term shared, as used above, means that some or all code from multiple modules may be executed using a single (shared) processor. In addition, some or all code from multiple modules may be stored by a single (shared) memory. The term group, as used above, means that some or all code from a single module may be executed using a group of processors. In addition, some or all code from a single module may be stored using a group of memories.

[0037] The techniques described herein may be implemented by one or more computer programs executed by one or more processors. The computer programs include processor-executable instructions that are stored on a non-transitory tangible computer readable medium. The computer programs may also include stored data. Non-limiting examples of the non-transitory tangible computer readable medium are nonvolatile memory, magnetic storage, and optical storage.

[0038] Some portions of the above description present

the techniques described herein in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. These operations, while described functionally or logically, are understood to be implemented by computer programs. Furthermore, it has also proven convenient at times to refer to these arrangements of operations as modules or by functional names, without loss of generality.

[0039] Unless specifically stated otherwise as apparent from the above discussion, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system memories or registers or other such information storage, transmission or display devices.

[0040] Certain aspects of the described techniques include process steps and instructions described herein in the form of an algorithm. It should be noted that the described process steps and instructions could be embodied in software, firmware or hardware, and when embodied in software, could be downloaded to reside on and be operated from different platforms used by real time network operating systems.

[0041] The present disclosure also relates to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general-purpose computer selectively activated or reconfigured by a computer program stored on a computer readable medium that can be accessed by the computer. Such a computer program may be stored in a tangible computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, application specific integrated circuits (ASICs), or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus. Furthermore, the computers referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[0042] The algorithms and operations presented herein are not inherently related to any particular computer or other apparatus. Various general-purpose systems may also be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatuses to perform the required method steps. The required structure for a variety of these systems will be apparent to those of skill in the art, along with equivalent variations. In addition, the present disclosure is not described with reference to any particular programming language. It is appreciated that a variety of programming languages may be used to implement the teachings of the present disclosure as described herein, and any references to specific languages are provided for disclosure of enablement and best mode of the present invention.

[0043] The present disclosure is well suited to a wide variety of computer network systems over numerous topologies. Within this field, the configuration and management of large networks comprise storage devices and computers that are communicatively coupled to dissimilar computers and storage devices over a network, such as the Internet.

[0044] The foregoing description of the embodiments has been provided for purposes of illustration and description. It is not intended to be exhaustive or to limit the disclosure. Individual elements or features of a particular embodiment are generally not limited to that particular embodiment, but, where applicable, are interchangeable and can be used in a selected embodiment, even if not specifically shown or described. The same may also be varied in many ways.

Claims

1. A computer-implemented method (300), comprising:

obtaining (308), by a first computing device (104), an audio input signal for an audio communication session with a second computing device (108) using audio data captured by a microphone (216) of the first computing device (104);

analyzing (312), by the first computing device (104), the audio input signal to detect a speech input by a first user associated with the first computing device (104);

determining (316), by the first computing device (104), a detection period duration from when the audio input signal was obtained until the analyzing has completed;

transmitting (320), from the first computing device (104) and to the second computing device (108), (i) a portion of the audio input signal beginning at a start of the speech input and (ii) the detection period duration;

in response to receiving, at the second computing device (108), the portion of the audio input signal and the detection period duration, accelerating playback, by the second computing device (108), of the portion of the audio input signal to compensate for the detection period duration;

analyzing (324), by the first computing device (104), the audio input signal to detect an end of the speech input by the first user; and

terminating (328) transmission, from the first computing device (104) to the second comput-

ing device (108), of the portion of the audio input signal at a point corresponding to the detected end of the speech input by the first user.

2. The computer-implemented method (300) of claim 1, further comprising encoding, by the first computing device (104), the portion of the audio input signal to obtain a set of audio packets, wherein the transmitting includes transmitting, to the second computing device (108), (i) the set of audio packets and (ii) the detection period duration.

3. The computer-implemented method (300) of claim 2, wherein receipt of the set of audio packets and the detection period duration causes the second computing device (108) to:

decode the set of audio packets to obtain an audio output signal;
remove a redundant portion of the audio output signal corresponding to one or more pitch periods to obtain a modified audio output signal, wherein the modified audio output signal has a shorter length than the audio output signal; and
output, by a speaker of the second computing device (108), the modified audio output signal, wherein a quantity of the one or more removed pitch periods preferably corresponds to the detection period duration.

4. The computer-implemented method (300) according to claim 3, wherein multiple pitch periods are removed, in particular multiple pitch periods having a length of less than 15 milliseconds, in particular less than 7.5 milliseconds.

5. The computer-implemented method (300) of claim 3, wherein receipt of the set of audio packets and the detection period duration causes the second computing device (108) to remove the redundant portion of the audio output signal by:
cross-correlating the audio output signal with itself to obtain an autocorrelation signal; and
detecting one or more peaks of the autocorrelation signal that exceed a threshold indicative of the one or more pitch periods of the audio output signal, wherein the threshold is preferably in a range of 0.9 to 0.3, in particular in the range of 0.6 and 0.45, more particular 0.5.

6. The computer-implemented method (300) of claim 1, wherein analyzing the audio input signal to detect the speech input includes applying a voice activity detection, VAD, technique to the audio input signal, the VAD technique having an aggressiveness or accuracy that corresponds to the detection period duration, wherein applying the voice detection technique to the audio input signal preferably includes distinguishing the speech input by the first user from speech by a second user within the audio input signal.

7. A computing network (100) including a first computing device (104) and a second computing device (108), wherein the first and second computing devices (104, 108) comprise one or more processors (208) and a non-transitory memory (212) storing a set of instructions that, when executed by the one or more processors (208), causes the first and the second computing device (104, 108) to perform operations comprising:

obtaining, by the first computing device (104), an audio input signal for an audio communication session with the second computing device (108) using audio data captured by a microphone (216) of the first computing device (104);
analyzing, by the first computing device (104), the audio input signal to detect a speech input by a first user associated with the first computing device (104);
determining, by the first computing device (104), a detection period duration from when the audio input signal was obtained until the analyzing has completed;
transmitting, by the first computing device (104), to the second computing device (108), (i) a portion of the audio input signal beginning at a start of the speech input and (ii) the detection period duration;
in response to receiving, at the second computing device (108), the portion of the audio input signal and the detection period duration, accelerating playback, by the second computing device (108), of the portion of the audio input signal to compensate for the detection period duration;
analyzing, by the first computing device (104), the audio input signal to detect an end of the speech input by the first user; and
terminating, by the first computing device (104), transmission, to the second computing device (108), of the portion of the audio input signal at a point corresponding to the detected end of the speech input by the first user.

8. The computing network (100) of claim 7, wherein the operations further comprise encoding, by the first computing device (104), the portion of the audio input signal to obtain a set of audio packets, wherein the transmitting includes transmitting, to the second computing device (108), (i) the set of audio packets and (ii) the detection period duration.

9. The computing network (100) of claim 8, wherein receipt of the set of audio packets, by the first computing device (104), and the detection period dura-

tion preferably causes the second computing device (108) to:

decode the set of audio packets to obtain an audio output signal;
remove a redundant portion of the audio output signal corresponding to one or more pitch periods to obtain the modified audio output signal, wherein the modified audio output signal has a shorter length than the audio output signal; and
output, by a speaker of the second computing device (108), the modified audio output signal.

10. The computing network (100) of claim 9, wherein a quantity of the one or more removed pitch periods preferably corresponds to the detection period duration.

11. The computing network (100) of claim 9 or 10, wherein multiple pitch periods are removed, in particular multiple pitch periods having a length of less than 15 milliseconds, in particular less than 7.5 milliseconds.

12. The computing network (100) of claim 9, wherein receipt of the set of audio packets, by the first computing device (104), and the detection period duration causes the second computing device (108) to remove the redundant portion of the audio output signal by:

cross-correlating the audio output signal with itself to obtain an autocorrelation signal; and
detecting one or more peaks of the autocorrelation signal that exceed a threshold indicative of the one or more pitch periods of the audio output signal.

13. The computing network (100) of claim 12, wherein the threshold is in a range of 0.9 to 0.3, in particular in the range of 0.6 and 0.45, more particular 0.5.

14. The computing network (100) of claim 7, wherein analyzing the audio input signal to detect the speech input, by the first computing device, includes applying a voice activity detection (VAD) technique to the audio input signal, the VAD technique having an aggressiveness or accuracy that corresponds to the detection period duration, wherein applying the VAD to the audio input signal, by the first computing device (104), preferably includes distinguishing the speech input by the first user from speech by a second user within the audio input signal.

15. A non-transitory computer-readable medium having a set of instructions stored thereon that, when executed by one or more processors (208) of a first and a second computing device (104, 108), causes the first and second computing device (104, 108) to perform the method steps of any of claims 1 to 6.

Patentansprüche

1. Ein computer-implementiertes Verfahren (300), wobei das Verfahren folgende Schritte aufweist:

Erhalten (308), durch eine erste Rechenvorrichtung (104), eines Audio-Eingangssignals für eine Audio-Kommunikationssitzung mit einer zweiten Rechenvorrichtung (108) unter Verwendung von Audiodaten, die durch ein Mikrofon (216) der ersten Rechenvorrichtung (104) erfasst wurden;
Analysieren (312), durch die erste Rechenvorrichtung (104), des Audio-Eingangssignals, um eine Spracheingabe durch einen ersten Benutzer zu erfassen, der mit der ersten Rechenvorrichtung (104) verbunden ist;
Bestimmen (316), durch die erste Rechenvorrichtung (104), einer Erfassungsperiodendauer von dem Zeitpunkt an, zu dem das Audio-Eingangssignal erhalten wurde, bis die Analyse abgeschlossen ist;
Übertragen (320), von der ersten Rechenvorrichtung (104) und zu der zweiten Rechenvorrichtung (108), (i) eines Teils des Audio-Eingangssignals beginnend mit einem Beginn der Spracheingabe und (ii) der Erfassungsperiodendauer;
ansprechend auf den Empfang des Teils des Audio-Eingangssignals und der Erfassungsperiodendauer an der zweiten Rechenvorrichtung (108), Beschleunigung der Wiedergabe des Teils des Audio-Eingangssignals durch die zweite Rechenvorrichtung (108), um die Erfassungsperiodendauer zu kompensieren;
Analysieren (324) des Audio-Eingangssignals durch die erste Rechenvorrichtung (104), um ein Ende der Spracheingabe durch den ersten Benutzer zu erfassen; und
Beenden (328) der Übertragung, von der ersten Rechenvorrichtung (104) zur zweiten Rechenvorrichtung (108), des Teils des Audio-Eingangssignals an einem Punkt, der dem erfassten Ende der Spracheingabe durch den ersten Benutzer entspricht.

2. Computer-implementiertes Verfahren (300) nach Anspruch 1, das ferner das Codieren des Teils des Audio-Eingangssignals durch die erste Rechenvorrichtung (104) umfasst, um einen Satz von Audiopaketen zu erhalten, wobei das Übertragen das Übertragen (i) des Satzes von Audiopaketen und (ii) der Erfassungsperiodendauer an die zweite Rechenvorrichtung (108) umfasst.

3. Computer-implementiertes Verfahren (300) nach Anspruch 2, wobei der Empfang des Satzes von Audiopaketen und der Erfassungsperiodendauer die zweite Rechenvorrichtung (108) dazu veranlasst, (i) den Satz von Audiopaketen und (ii) die Erfassungsperiodendauer zu übertragen:

den Satz von Audiopaketen zu dekodieren, um ein Audio-Ausgangssignal zu erhalten;
einen redundanten Teil des Audioausgangssignals zu entfernen, der einer oder mehreren Tonhöhenperioden entspricht, um ein modifiziertes Audioausgangssignal zu erhalten, wobei das modifizierte Audioausgangssignal eine kürzere Länge als das Audioausgangssignal hat; und
das modifizierte Audio-Ausgangssignals durch einen Lautsprecher der zweiten Rechenvorrichtung (108) auszugeben, wobei eine Menge der einen oder mehreren entfernten Tonhöhenperioden vorzugsweise der Erfassungsperiodendauer entspricht.

4. Computer-implementiertes Verfahren (300) nach Anspruch 3, bei dem mehrere Tonhöhenperioden entfernt werden, insbesondere mehrere Tonhöhenperioden mit einer Länge von weniger als 15 Millisekunden, insbesondere weniger als 7,5 Millisekunden.

5. Computer-implementiertes Verfahren (300) nach Anspruch 3, wobei der Empfang des Satzes von Audiopaketen und die Dauer der Erfassungsperiode die zweite Rechenvorrichtung (108) veranlasst, den redundanten Teil des Audioausgangssignals zu entfernen durch:

Kreuzkorrelieren des Audio-Ausgangssignals mit sich selbst, um ein Autokorrelationssignal zu erhalten; und
Detektieren einer oder mehrerer Spitzen des Autokorrelationssignals, die einen Schwellenwert überschreiten, der eine oder mehrere Tonhöhenperioden des Audioausgangssignals anzeigt, wobei der Schwellenwert vorzugsweise in einem Bereich von 0,9 bis 0,3, insbesondere im Bereich von 0,6 und 0,45, insbesondere 0,5, liegt.

6. Computer-implementiertes Verfahren (300) nach Anspruch 1, wobei das Analysieren des Audio-Eingangssignals zum Erfassen der Spracheingabe das Anwenden einer Sprachaktivitäts-Erfassungstechnik, VAD, auf das Audio-Eingangssignal einschließt, wobei die VAD-Technik eine Aggressivität oder Genauigkeit aufweist, die der Dauer der Erfassungsperiode entspricht, wobei das Anwenden der Spracherfassungstechnik auf das Audio-Eingangssignal vorzugsweise das Unterscheiden der Spracheingabe durch den ersten Benutzer von der Sprache durch einen zweiten Benutzer innerhalb des Audio-Eingangssignals einschließt.

7. Ein Computernetzwerk (100), das eine erste Rechenvorrichtung (104) und eine zweite Rechenvorrichtung (108) enthält, wobei die erste und die zweite Rechenvorrichtung (104, 108) einen oder mehrere Prozessoren (208) und einen nicht-flüchtigen Speicher (212) umfassen, der einen Satz von Befehlen speichert, der, wenn er von dem einen oder den mehreren Prozessoren (208) ausgeführt wird, bewirkt, dass die erste und die zweite Rechenvorrichtung (104, 108) Operationen ausführen, die Folgendes umfassen
Erhalten, durch die erste Rechenvorrichtung (104), eines Audio-Eingangssignals für eine Audio-Kommunikationssitzung mit der zweiten Rechenvorrichtung (108) unter Verwendung von Audiodaten, die durch ein Mikrofon (216) der ersten Rechenvorrichtung (104) erfasst wurden;
Analysieren, durch die erste Rechenvorrichtung (104), des Audio-Eingangssignals, um eine Spracheingabe durch einen ersten Benutzer zu erfassen, der der ersten Rechenvorrichtung (104) zugeordnet ist;
Bestimmen, durch die erste Rechenvorrichtung (104), einer Erfassungsperiodendauer von dem Zeitpunkt an, an dem das Audioeingangssignal erhalten wurde, bis die Analyse abgeschlossen ist;
Übertragen, durch die erste Rechenvorrichtung (104), (i) eines Teils des Audio-Eingangssignals beginnend mit einem Beginn der Spracheingabe und (ii) der Erfassungsperiodendauer an die zweite Rechenvorrichtung (108);
ansprechend auf den Empfang des Teils des Audio-Eingangssignals und der Erfassungsperiodendauer an der zweiten Rechenvorrichtung (108), Beschleunigen der Wiedergabe des Teils des Audio-Eingangssignals durch die zweite Rechenvorrichtung (108), um die Erfassungsperiodendauer zu kompensieren;
Analysieren des Audio-Eingangssignals durch die erste Rechenvorrichtung (104), um ein Ende der Spracheingabe durch den ersten Benutzer zu erfassen; und
Beenden der Übertragung des Teils des Audio-Eingangssignals durch die erste Rechenvorrichtung (104) an die zweite Rechenvorrichtung (108) an einem Punkt, der dem erfassten Ende der Spracheingabe durch den ersten Benutzer entspricht.

8. Computernetzwerk (100) nach Anspruch 7, wobei die Operationen ferner das Codieren des Teils des Audio-Eingangssignals durch die erste Rechenvorrichtung (104) umfassen, um einen Satz von Audiopaketen zu erhalten, wobei das Übertragen das Übertragen (i) des Satzes von Audiopaketen und (ii)

der Erfassungsperiodendauer an die zweite Rechenvorrichtung (108) umfasst.

9. Computernetzwerk (100) nach Anspruch 8, wobei der Empfang des Satzes von Audiopaketen durch die erste Berechnungsvorrichtung (104) und die Erfassungsperiodendauer vorzugsweise die zweite Berechnungsvorrichtung (108) dazu veranlasst:

den Satz von Audiopaketen zu dekodieren, um ein Audio-Ausgangssignal zu erhalten;
einen redundanten Teil des Audioausgangssignals zu entfernen, der einer oder mehreren Tonhöhenperioden entspricht, um das modifizierte Audioausgangssignal zu erhalten, wobei das modifizierte Audioausgangssignal eine kürzere Länge als das Audioausgangssignal hat; und
das modifizierte Audio-Ausgangssignals durch einen Lautsprecher der zweiten Rechenvorrichtung (108) auszugeben.

10. Computernetzwerk (100) nach Anspruch 9, wobei eine Menge der einen oder mehreren entfernten Tonhöhenperioden vorzugsweise der Dauer der Erfassungsperiode entspricht.

11. Computernetzwerk (100) nach Anspruch 9 oder 10, bei dem mehrere Tonhöhenperioden entfernt werden, insbesondere mehrere Tonhöhenperioden mit einer Länge von weniger als 15 Millisekunden, insbesondere weniger als 7,5 Millisekunden.

12. Computernetzwerk (100) nach Anspruch 9, wobei der Empfang des Satzes von Audiopaketen und der Dauer der Erfassungsperiode durch die erste Berechnungsvorrichtung (104), die zweite Berechnungsvorrichtung (108) veranlasst, den redundanten Teil des Audioausgangssignals zu entfernen durch:

Kreuzkorrelieren des Audio-Ausgangssignals mit sich selbst, um ein Autokorrelationssignal zu erhalten; und
Detektieren einer oder mehrerer Spitzen des Autokorrelationssignals, die einen Schwellenwert überschreiten, der eine oder mehrere Tonhöhenperioden des Audioausgangssignals anzeigt.

13. Computernetzwerk (100) nach Anspruch 12, wobei der Schwellenwert in einem Bereich von 0,9 bis 0,3, insbesondere im Bereich von 0,6 und 0,45, insbesondere 0,5, liegt.

14. Computernetzwerk (100) nach Anspruch 7, wobei das Analysieren des Audio-Eingangssignals zum Erfassen der Spracheingabe durch die erste Rechenvorrichtung das Anwenden einer Sprachaktivitäts-Erfassungstechnik (VAD) auf das Audio-Eingangssignal einschließt, wobei die VAD-Technik eine Aggressivität oder Genauigkeit aufweist, die der Dauer der Erfassungsperiode entspricht, wobei das Anwenden der VAD auf das Audio-Eingangssignal durch die erste Rechenvorrichtung (104) vorzugsweise das Unterscheiden der Spracheingabe durch den ersten Benutzer von der Sprache durch einen zweiten Benutzer innerhalb des Audio-Eingangssignals einschließt.

15. Ein nicht-flüchtiges computerlesbares Medium mit einem darauf gespeicherten Satz von Befehlen, der, wenn er von einem oder mehreren Prozessoren (208) einer ersten und einer zweiten Rechenvorrichtung (104, 108) ausgeführt wird, die erste und die zweite Rechenvorrichtung (104, 108) dazu veranlasst, die Verfahrensschritte nach einem der Ansprüche 1 bis 6 auszuführen.

Revendications

1. Un procédé (300) mis en oeuvre par ordinateur, comprenant :
le fait (308) d'obtenir, par un premier dispositif informatique (104), un signal d'entrée audio pour une session de communication audio avec un deuxième dispositif informatique (108), en utilisant des données audio capturées par un microphone (216) du premier dispositif informatique (104) :

le fait (312) d'analyser, par le premier dispositif informatique (104), le signal d'entrée audio pour détecter une entrée vocale par un premier utilisateur associé au premier dispositif informatique (104) ;
le fait (316) de déterminer, par le premier dispositif informatique (104), une durée de période de détection depuis le moment où le signal d'entrée audio a été obtenu jusqu'à la fin de l'analyse ;
le fait (320) de transmettre, depuis le premier dispositif informatique (104) et vers le deuxième dispositif informatique (108), (i) une partie du signal d'entrée audio commençant à un début de l'entrée vocale et (ii) la durée de la période de détection ;
en réponse à la réception, au niveau du deuxième dispositif informatique (108), de la partie du signal d'entrée audio et de la durée de la période de détection, le fait d'accélérer la lecture, par le deuxième dispositif informatique (108), de la partie du signal d'entrée audio pour compenser la durée de la période de détection ;
le fait (324) d'analyser, par le premier dispositif informatique (104), le signal d'entrée audio pour détecter une fin de l'entrée vocale par le premier utilisateur ; et

le fait (328) de terminer la transmission, depuis le premier dispositif informatique (104) vers le deuxième dispositif informatique (108), de la partie du signal d'entrée audio en un point correspondant à la fin détectée de l'entrée vocale par le premier utilisateur.

2. Le procédé informatique (300) selon la revendication 1, comprenant en outre le fait de coder, par le premier dispositif informatique (104), la partie du signal d'entrée audio pour obtenir un ensemble de paquets audio, la transmission comprenant le fait de transmettre, au deuxième dispositif informatique (108), (i) l'ensemble des paquets audio et (ii) la durée de la période de détection.

3. Le procédé (300) mis en œuvre par ordinateur selon la revendication 2, dans lequel la réception de l'ensemble de paquets audio et de la durée de la période de détection amène le deuxième dispositif informatique (108) à :

décoder l'ensemble de paquets audio pour obtenir un signal de sortie audio ;
supprimer une partie redondante du signal de sortie audio correspondant à une ou plusieurs périodes de hauteur tonale, ou "de pitch", pour obtenir un signal de sortie audio modifié, le signal de sortie audio modifié ayant une longueur plus courte que le signal de sortie audio ; et
délivrer, par un haut-parleur du deuxième dispositif informatique (108), le signal de sortie audio modifié, une quantité desdites une ou plusieurs périodes de hauteur tonale supprimées correspondant de préférence à la durée de la période de détection.

4. Le procédé (300) mis en œuvre par ordinateur selon la revendication 3, dans lequel plusieurs périodes de hauteur tonale sont supprimées, en particulier plusieurs périodes de hauteur tonale ayant une durée inférieure à 15 millisecondes, en particulier inférieure à 7,5 millisecondes.

5. Le procédé (300) mis en œuvre par ordinateur selon la revendication 3, dans lequel la réception de l'ensemble de paquets audio et la durée de la période de détection amène le deuxième dispositif informatique (108) à supprimer la partie redondante du signal de sortie audio en :

inter-corrélant le signal de sortie audio avec lui-même pour obtenir un signal d'autocorrélation ; et
détectant un ou plusieurs pics du signal d'autocorrélation qui dépassent un seuil indicatif desdites une ou plusieurs périodes de hauteur tonale du signal de sortie audio, le seuil étant de préférence situé dans une gamme allant de 0,9 à 0,3, en particulier dans la gamme allant de 0,6 et 0,45, et, de façon encore plus en particulier, est de 0,5.

6. Le procédé informatique (300) selon la revendication 1, dans lequel l'analyse du signal d'entrée audio pour détecter l'entrée vocale comprend le fait d'appliquer une technique de détection d'activité vocale, en abrégé VAD, au signal d'entrée audio, la technique VAD ayant une agressivité. ou une précision qui correspond à la durée de la période de détection, le fait d'appliquer la technique de détection vocale au signal d'entrée audio incluant de préférence le fait de distinguer l'entrée vocale par le premier utilisateur de la parole d'un deuxième utilisateur dans le signal d'entrée audio.

7. Un réseau informatique (100) comprenant un premier dispositif informatique (104) et un deuxième dispositif informatique (108), les premier et deuxième dispositifs informatiques (104, 108) comprenant un ou plusieurs processeurs (208) et une mémoire non transitoire (212) stockant un ensemble d'instructions qui, lorsqu'elles sont exécutées par lesdits un ou plusieurs processeurs (208), amènent le premier et le deuxième dispositifs informatiques (104, 108) à effectuer des opérations comprenant :

le fait d'obtenir, par le premier dispositif informatique (104), un signal d'entrée audio pour une session de communication audio avec le deuxième dispositif informatique (108), en utilisant des données audio capturées par un microphone (216) du premier dispositif informatique (104) ;
le fait d'analyser, par le premier dispositif informatique (104), le signal d'entrée audio pour détecter une entrée vocale par un premier utilisateur associé au premier dispositif informatique (104) ;
le fait de déterminer, par le premier dispositif informatique (104), une durée de période de détection depuis le moment auquel le signal d'entrée audio a été obtenu jusqu'à la fin de l'analyse ;
le fait de transmettre, par le premier dispositif informatique (104), au deuxième dispositif informatique (108), (i) une partie du signal d'entrée audio commençant à un début de l'entrée vocale et (ii) la durée de la période de détection ;
en réponse à la réception, au niveau du deuxième dispositif informatique (108), de la partie du signal d'entrée audio et de la durée de la période de détection, le fait d'accélérer la lecture, par le deuxième dispositif informatique (108), de la partie du signal d'entrée audio pour compenser la durée de la période de détection ;

le fait d'analyser, par le premier dispositif informatique (104), le signal d'entrée audio pour détecter une fin de l'entrée vocale par le premier utilisateur ; et
le fait de terminer, par le premier dispositif informatique (104), la transmission, vers le deuxième dispositif informatique (108), de la partie du signal d'entrée audio en un point correspondant à la fin détectée de l'entrée vocale par le premier utilisateur.

8. Le réseau informatique (100) selon la revendication 7, dans lequel les opérations comprennent en outre le fait de coder, par le premier dispositif informatique (104), la partie du signal d'entrée audio pour obtenir un ensemble de paquets audio, la transmission comprenant le fait de transmettre, au deuxième dispositif informatique (108), (i) l'ensemble des paquets audio et (ii) la durée de la période de détection.

9. Le réseau informatique (100) selon la revendication 8, dans lequel la réception de l'ensemble de paquets audio, par le premier dispositif informatique (104), et la durée de la période de détection amènent de préférence le deuxième dispositif informatique (108) à :

décoder l'ensemble de paquets audio pour obtenir un signal de sortie audio ;
supprimer une partie redondante du signal de sortie audio correspondant à une ou plusieurs périodes de hauteur tonale, ou "de pitch", pour obtenir le signal de sortie audio modifié, le signal de sortie audio modifié ayant une longueur plus courte que le signal de sortie audio ; et
délivrer, par un haut-parleur du deuxième dispositif informatique (108), le signal de sortie audio modifié.

10. Le réseau informatique (100) selon la revendication 9, dans lequel une quantité desdites une ou plusieurs périodes de hauteur tonale supprimées correspond de préférence à la durée de la période de détection.

11. Le réseau informatique (100) selon la revendication 9 ou la revendication 10, dans lequel plusieurs périodes de hauteur tonale sont supprimées, en particulier des périodes de hauteur tonale multiples d'une durée inférieure à 15 millisecondes, en particulier inférieure à 7,5 millisecondes.

12. Le réseau informatique (100) selon la revendication 9, dans lequel la réception de l'ensemble de paquets audio, par le premier dispositif informatique (104), et la durée de la période de détection amènent le deuxième dispositif informatique (108) à supprimer la partie redondante du signal de sortie audio en :

inter-corrélant le signal de sortie audio avec lui-même pour obtenir un signal d'autocorrélation ; et
détectant un ou plusieurs pics du signal d'autocorrélation qui dépassent un seuil indicatif desdites une ou plusieurs périodes de hauteur tonale du signal de sortie audio.

13. Le réseau informatique (100) selon la revendication 12, dans lequel le seuil est situé dans une gamme allant de 0,9 à 0,3, en particulier dans la gamme allant de 0,6 à 0,45, et de façon encore plus en particulier, est de 0,5.

14. Le réseau informatique (100) selon la revendication 7, dans lequel l'analyse du signal d'entrée audio pour détecter l'entrée vocale, par le premier dispositif informatique, comprend le fait d'appliquer une technique de détection d'activité vocale (VAD) au signal d'entrée audio, la technique VAD ayant une agressivité ou une précision qui correspond à la durée de la période de détection, le fait d'appliquer la VAD au signal d'entrée audio, par le premier dispositif informatique (104), comprenant de préférence le fait de distinguer l'entrée vocale par le premier utilisateur de la parole d'un deuxième utilisateur dans le signal d'entrée audio.

15. Un support non transitoire, lisible par ordinateur, sur lequel est stocké un ensemble d'instructions qui, lorsqu'elles sont exécutées par un ou plusieurs processeurs (208) d'un premier et d'un deuxième dispositif informatique (104, 108), amènent le premier et le deuxième dispositifs informatiques (104, 108) à mettre en œuvre les étapes du procédé selon l'une quelconque des revendications 1 à 6.

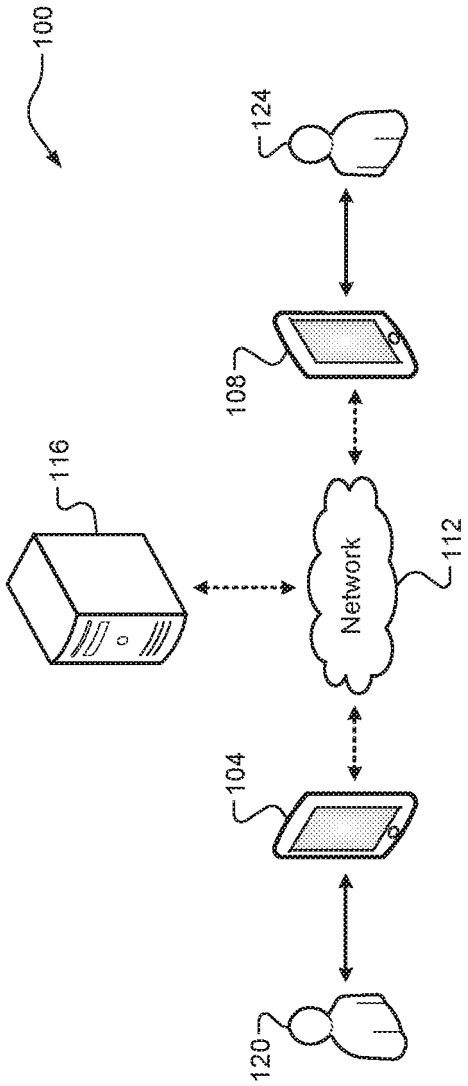
100
116
Network
112
108
124
104
120

FIG. 1

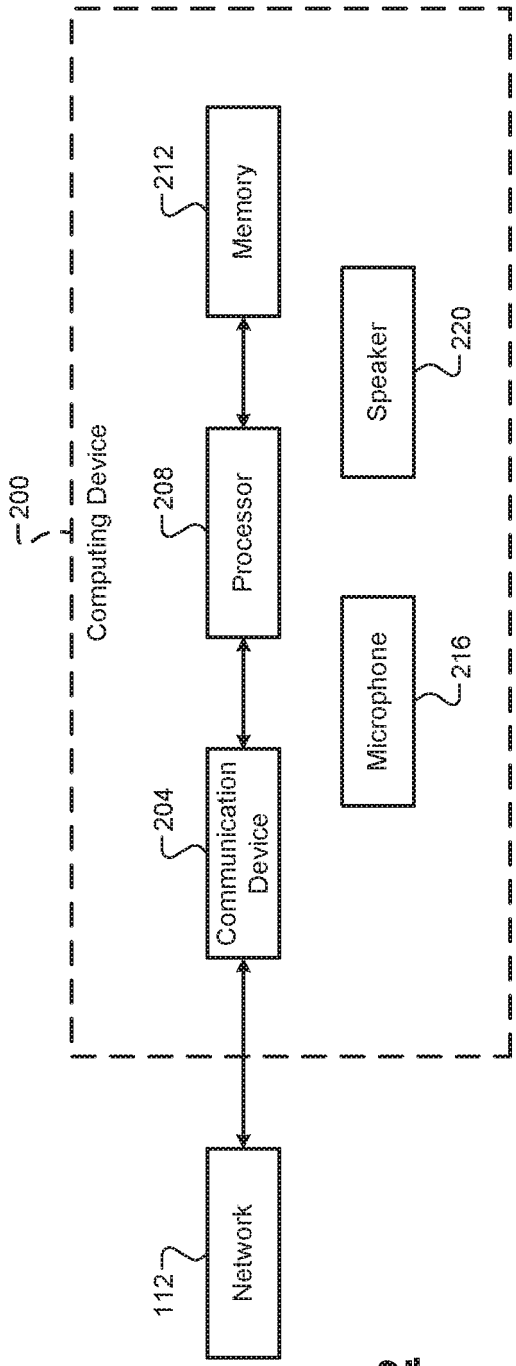
200
Computing Device
212
Memory
208
Processor
204
Communication Device
112
Network
Microphone
216
Speaker
220

FIG. 2

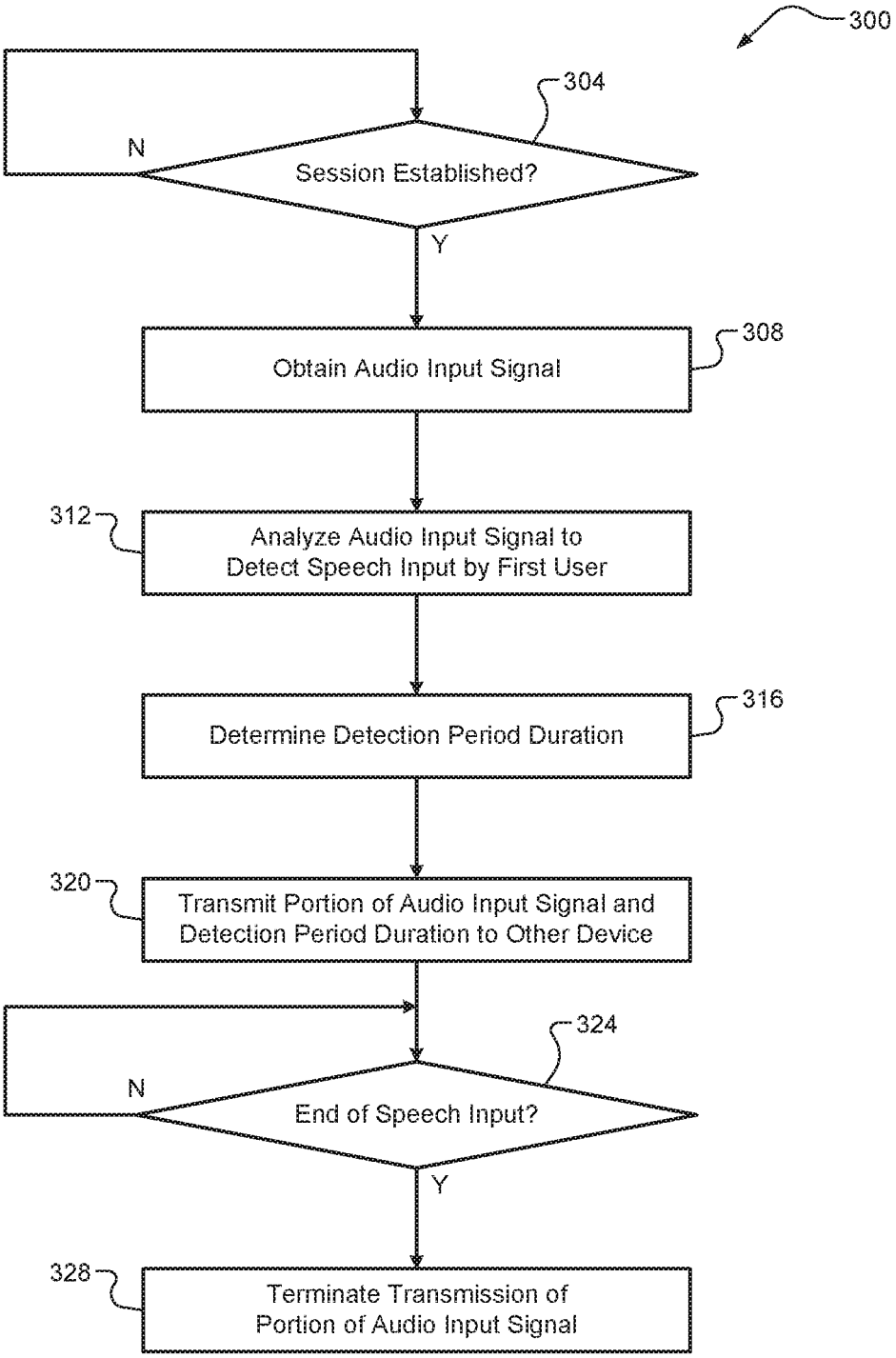
300
304
N
Session Established?
Y
308
Obtain Audio Input Signal
312
Analyze Audio Input Signal to
Detect Speech Input by First User
316
Determine Detection Period Duration
320
Transmit Portion of Audio Input Signal and
Detection Period Duration to Other Device
324
N
End of Speech Input?
Y
328
Terminate Transmission of
Portion of Audio Input Signal

FIG. 3

REFERENCES CITED IN THE DESCRIPTION

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

Patent documents cited in the description

- US 7016850 B1 **[0002]**
- US 2008281586 A1 **[0003]**
- EP 1750397 A1 **[0004]**