

(51) International Patent Classification:
G10L 25/78 (2013.01)(21) International Application Number:
PCT/US2016/012692(22) International Filing Date:
8 January 2016 (08.01.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **NUANCE COMMUNICATIONS, INC.**
[US/US]; One Wayside Road, Burlington, Massachusetts
01803 (US).(72) Inventors: **GRAF, Simon**; Blumenscheinweg 8, 89077
Ulm (DE). **BUCK, Markus**; Geschwister-Rommmer-Weg 4,
88400 Biberach (DE). **HERBIG, Tobias**; Traminerweg 40,
89075 Ulm (DE).(74) Agents: **WHITTENBERGER, Mark H.** et al.; Holland &
Knight LLP, 10. St. James Avenue, Boston, Massachusetts
02116 (US).(81) Designated States (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,
SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.(84) Designated States (*unless otherwise indicated, for every
kind of regional protection available*): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,
TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— with international search report (Art. 21(3))

(54) Title: SYSTEM AND METHOD FOR SPEECH DETECTION ADAPTATION

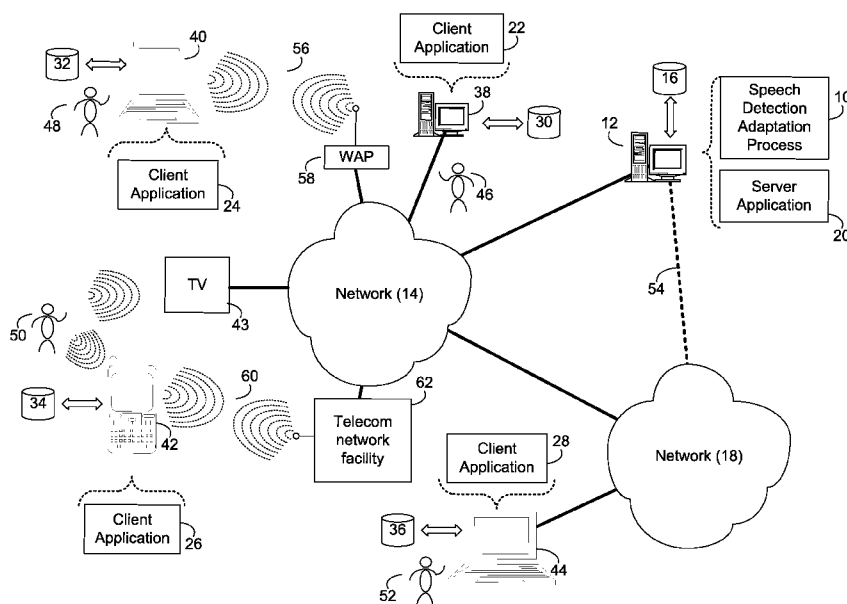


FIG. 1

(57) Abstract: A method for speech detection adaptation is provided. Embodiments may include receiving, at a processor, a speech signal corresponding to a particular environment and estimating one or more acoustical parameters that characterize the environment. In some embodiments, the one or more acoustical parameters are not configured to identify a known scenario. Embodiments may include dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.

System and Method for Speech Detection Adaptation

Technical Field

[001] This disclosure relates to signal processing and, more particularly, to a method for speech detection adaptation.

Background

[002] Several algorithms in speech enhancement applications - such as noise estimation, automatic gain control (AGC), or end-pointing of speech intervals as a preprocessing step for speech recognizers may rely on the results of speech detectors that identify the presence of speech in a noisy signal. The algorithms are applied in many different acoustical environments. As such, robustness against various types of noise interferences is desired.

[003] Most speech detectors do not assume a prior knowledge of the acoustical environment therefore a tradeoff between the rate of correctly detected speech frames and the rate of noise-only frames falsely classified as speech has to be found. These universal detectors are typically tuned by optimizing the average detection results over all relevant environments.

Summary of Disclosure

[004] In one implementation of the present disclosure, a method for speech detection adaptation is provided. The method may include receiving, at a processor, a speech signal corresponding to a particular environment and estimating one or more acoustical parameters that characterize the environment. In some embodiments, the one or more acoustical parameters are not configured to identify a known scenario. The method may include dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.

[005] One or more of the following features may be included. In some embodiments, after dynamically controlling, the method may include identifying at least one known scenario based on the one or more acoustical parameters. The method may further include selecting a fixed configuration tuned for the at least one known scenario instead of the dynamically controlled configuration. The method may also include determining a reliability of a voice activity detection (VAD) feature by rating the VAD feature's performance in at least one previous frame. The method may further include dynamically adapting an amount of context that is employed for each of the one or more acoustical parameters. The method may also include identifying a VAD feature having an insignificant contribution and omitting at least one calculation associated with the VAD feature. In some embodiments, estimating one or more acoustical parameters may be performed for a subset of frames. The one or more acoustical parameters may be derived from non-acoustic information. The non-acoustic information may include at least one of video, acceleration sensors, and geo-localization information. The one or more acoustical parameters may be employed in a non-speech detection algorithm.

[006] In another implementation, a system for speech detection adaptation is provided. The system may include a processor configured to perform one or more operations. Some operations may include receiving, at a processor, a speech signal corresponding to a particular environment and estimating one or more acoustical parameters that characterize the environment. In some embodiments, the one or more acoustical parameters are not configured to identify a known scenario. Operations may include dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.

[007] One or more of the following features may be included. In some embodiments, after dynamically controlling, the method may include identifying at least one known scenario based on the one or more acoustical parameters. Operations

may further include selecting a fixed configuration tuned for the at least one known scenario instead of the dynamically controlled configuration. Operations may also include determining a reliability of a VAD feature by rating the VAD feature's performance in at least one previous frame. Operations may further include dynamically adapting an amount of context that is employed for each of the one or more acoustical parameters. Operations may also include identifying a VAD feature having an insignificant contribution and omitting at least one calculation associated with the VAD feature. In some embodiments, estimating one or more acoustical parameters may be performed for a subset of frames. The one or more acoustical parameters may be derived from non-acoustic information. The non-acoustic information may include at least one of video, acceleration sensors, and geo-localization information. The one or more acoustical parameters may be employed in a non-speech detection algorithm.

[008] The details of one or more implementations are set forth in the accompanying drawings and the description below. Other features and advantages will become apparent from the description, the drawings, and the claims.

Brief Description of the Drawings

[009] FIG. 1 is a diagrammatic view of a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0010] FIG. 2 is a flowchart of a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0011] FIG. 3 shows an example input signal graph and a speech detector output in accordance with an embodiment of the present disclosure;

[0012] FIG. 4 is a diagrammatic view of a universal detection system configured to implement an speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0013] FIG. 5 is a diagrammatic view of a system configured to implement a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0014] FIG. 6 is a diagrammatic view of a system configured to implement a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0015] FIG. 7 is a diagrammatic view of a system configured to implement a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0016] FIG. 8 is diagram of an example of a noisy input signal in accordance with an embodiment of the present disclosure;

[0017] FIG. 9 is a diagrammatic view of a system configured to implement a speech detection adaptation process in accordance with an embodiment of the present disclosure;

[0018] FIG. 10 is a diagrammatic view of a system configured to implement a speech detection adaptation process in accordance with an embodiment of the present disclosure; and

[0019] FIG. 11 shows an example of a computer device and a mobile computer device that can be used to implement embodiments of the present disclosure.

[0020] Like reference symbols in the various drawings may indicate like elements.

Detailed Description of the Embodiments

[0021] Embodiments provided herein are directed towards a system and process for performing speech detection adaptation. Embodiments may include receiving, at a processor, a speech signal corresponding to a particular environment and estimating one or more acoustical parameters that characterize the environment. In some embodiments, the one or more acoustical parameters are not configured to identify a

known scenario. Embodiments may include dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.

[0022] Referring to FIG. 1, there is shown a speech detection adaptation process 10 that may reside on and may be executed by any of the devices shown in FIG. 1, for example, computer 12 (and/or any suitable device that includes a microphone), which may be connected to network 14 (e.g., the Internet or a local area network). Server application 20 may include some or all of the elements of speech detection adaptation process 10 described herein. Examples of computer 12 may include but are not limited to a single server computer, a series of server computers, a single personal computer, a series of personal computers, a mini computer, a mainframe computer, an electronic mail server, a social network server, a text message server, a photo server, a multiprocessor computer, one or more virtual machines running on a computing cloud, and/or a distributed system. The various components of computer 12 may execute one or more operating systems, examples of which may include but are not limited to: Microsoft Windows Server™, Novell Netware™, Redhat Linux™, Unix, or a custom operating system, for example.

[0023] The instruction sets and subroutines of speech detection adaptation process 10, which may be stored on storage device 16 coupled to computer 12, may be executed by one or more processors (not shown) and one or more memory architectures (not shown) included within computer 12. Storage device 16 may include but is not limited to: a hard disk drive; a flash drive, a tape drive; an optical drive; a RAID array; a random access memory (RAM); and a read-only memory (ROM).

[0024] Network 14 may be connected to one or more secondary networks (e.g., network 18), examples of which may include but are not limited to: a local area network; a wide area network; or an intranet, for example.

[0025] In some embodiments, speech detection adaptation process 10 may be accessed and/or activated via client applications 22, 24, 26, 28. Examples of client applications 22, 24, 26, 28 may include but are not limited to a standard web browser, a customized web browser, or a custom application that can display data to a user. The instruction sets and subroutines of client applications 22, 24, 26, 28, which may be stored on storage devices 30, 32, 34, 36 (respectively) coupled to client electronic devices 38, 40, 42, 44 (respectively), may be executed by one or more processors (not shown) and one or more memory architectures (not shown) incorporated into client electronic devices 38, 40, 42, 44 (respectively). The examples of client electronic devices shown in FIG. 1 are merely provided by way of example as any suitable device may be used in accordance with the teachings of the present disclosure (e.g., those having a loudspeaker, microphone, and/or processor).

[0026] Storage devices 30, 32, 34, 36 may include but are not limited to: hard disk drives; flash drives; tape drives; optical drives; RAID arrays; random access memories (RAM); and read-only memories (ROM). Examples of client electronic devices 38, 40, 42, 44 may include, but are not limited to, personal computer 38, laptop computer 40, smart phone 42, television 43, notebook computer 44, a server (not shown), a data-enabled, cellular telephone (not shown), and a dedicated network device (not shown).

[0027] One or more of client applications 22, 24, 26, 28 may be configured to effectuate some or all of the functionality of speech detection adaptation process 10. Accordingly, speech detection adaptation process 10 may be a purely server-side application, a purely client-side application, or a hybrid server-side / client-side application that is cooperatively executed by one or more of client applications 22, 24, 26, 28 and speech detection adaptation process 10.

[0028] Client electronic devices 38, 40, 42, 43, 44 may each execute an operating system, examples of which may include but are not limited to Apple iOS TM, Microsoft Windows TM, Android TM, Redhat Linux TM, or a custom operating system.

Each of client electronic devices 38, 40, 42, 43, and 44 may include one or more microphones and/or speakers configured to implement speech detection adaptation process 10 as is discussed in further detail below.

[0029] Users 46, 48, 50, 52 may access computer 12 and speech detection adaptation process 10 directly through network 14 or through secondary network 18. Further, computer 12 may be connected to network 14 through secondary network 18, as illustrated with phantom link line 54. In some embodiments, users may access speech detection adaptation process 10 through one or more telecommunications network facilities 62.

[0030] The various client electronic devices may be directly or indirectly coupled to network 14 (or network 18). For example, personal computer 38 is shown directly coupled to network 14 via a hardwired network connection. Further, notebook computer 44 is shown directly coupled to network 18 via a hardwired network connection. Laptop computer 40 is shown wirelessly coupled to network 14 via wireless communication channel 56 established between laptop computer 40 and wireless access point (i.e., WAP) 58, which is shown directly coupled to network 14. WAP 58 may be, for example, an IEEE 802.11a, 802.11b, 802.11g, Wi-Fi, and/or Bluetooth device that is capable of establishing wireless communication channel 56 between laptop computer 40 and WAP 58. All of the IEEE 802.11x specifications may use Ethernet protocol and carrier sense multiple access with collision avoidance (i.e., CSMA/CA) for path sharing. The various 802.11x specifications may use phase-shift keying (i.e., PSK) modulation or complementary code keying (i.e., CCK) modulation, for example. Bluetooth is a telecommunications industry specification that allows e.g., mobile phones, computers, and smart phones to be interconnected using a short-range wireless connection.

[0031] Smart phone 42 is shown wirelessly coupled to network 14 via wireless communication channel 60 established between smart phone 42 and

telecommunications network facility 62, which is shown directly coupled to network 14.

[0032] The phrase "telecommunications network facility", as used herein, may refer to a facility configured to transmit, and/or receive transmissions to/from one or more mobile devices (e.g. cellphones, etc). In the example shown in FIG. 1, telecommunications network facility 62 may allow for communication between TV 43, cellphone 42 (or television remote control, etc.) and server computing device 12. Embodiments of speech detection adaptation process 10 may be used with any or all of the devices described herein as well as many others.

[0033] As will be discussed below in greater detail in FIGS. 2-11, speech detection adaptation ("SDA") process 10 may include receiving (202), at a processor, a speech signal corresponding to a particular environment and estimating (204) one or more acoustical parameters that characterize the environment. In some embodiments, the one or more acoustical parameters are not configured to identify a known scenario. Embodiments may include dynamically controlling (206) a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.

[0034] Embodiments of the SDA process included herein may describe an environmentally dependent adaptation of speech detectors. Based on estimated environmental parameters, a combination and tuning of multiple VAD features may be chosen that is expected to cope best with the acoustical environment at hand. Earlier approaches classified the environment based on a finite set of acoustic scenarios (one of N selection). Depending on this classification, a fixed set of VAD features is employed for speech detection. Some examples of VAD features may include, but are not limited to, those described in "Features for Voice Activity Detection: A Comparative Analysis, Journal on Advances in Signal Processing, 2015.

[0035] Embodiments of the SDA process included herein may be configured to characterize the acoustical environment by multiple parameters. Instead of matching one environment from a finite set of known scenarios, different properties of the acoustical environment may be determined. Based on these environmental parameters, the configuration and tuning of VAD features may be dynamically adapted. Accordingly, the speech detector may be configured to adapt to new acoustical environments that were not considered in training by interpolating between multiple configurations. By directly making use of the environmental parameters, a smooth weighting of the influence of different VAD features can be achieved. In this way, SDA process may provide a new dynamic feature control that allows speech detectors and other algorithms to adapt to new acoustical environments. It is more flexible than earlier approaches and reduces the dependency on training data.

[0036] Speech detectors rely on different VAD features that are configured to indicate the presence of speech in a noisy signal. Each feature may reflect a property of speech (e.g., voicing or high SNR) that distinguishes the desired speech from noise interferences. Depending on the acoustical environment, the performance of features varies. As a result, it may be beneficial to select a combination of reasonably tuned features that are adapted to the acoustical environment.

[0037] Embodiments of SDA process 10 may include dynamic feature control that allows speech detection algorithms to adapt to new acoustical environments. The approach may include one or more stages that are discussed in further detail herein below.

[0038] In some embodiments, in a first stage, parameters are determined that describe different properties of the acoustical environment. Some of these parameters may include, but are not limited to, acoustical environmental parameters – such as reverberation time (T60), a priori signal-to-noise ratio (SNR), amount of interfering speech (babble), amount and shape of other non-stationary interferences, spectral distribution of the stationary background noise, and/or distortions (linear or non-

linear) of speech, etc. The environmental parameters may be estimated based on information from a long temporal context. The amount of context may be different for each parameter. The parameters may also include measures that do not explicitly estimate physical properties of the environment (e.g., kurtosis of the signal or flatness of the background noise spectrum).

[0039] In some embodiments, in addition to environmental properties, the confidence of each feature can be taken into account by rating its performance in the past (e.g., plausibility and comparison with other features). Estimation of environmental parameters does not have to be performed for each frame. In this way, the system may distribute calculations to multiple frames in order to reduce computational costs. Important environments may be covered by special cases that provide a complete set of feature parameters fitted to a certain environment. The parameter sets may be determined, for example, in a training. Environmental parameters may be derived from the results of other algorithms, from the VAD features themselves or even from non-acoustic information (e.g., video, acceleration sensors, geo-localization, etc.)

[0040] In some embodiments, in a second stage, speech detection algorithms may be adapted to the environment. The acoustical environmental parameters may be directly employed to control the combination and tuning of VAD features. The adapted speech detector may be expected to cope best with the environment.

[0041] In some embodiments, parameters of the features including, but not limited to, the amount of temporal context, smoothing parameters, decision thresholds, etc. may be adapted based on the estimated environmental parameters. Weighting of the features may be performed based on their expected reliabilities given the estimated environmental parameters and the confidence of earlier results. Features with low contribution to the final VAD may not necessarily have to be evaluated resulting in reduction of computational costs by focusing on relevant features. The estimated environmental parameters may be employed in any suitable algorithms, some of

which may include, but are not limited to, noise estimation, noise reduction, dereverberation, speech recognition, etc.

[0042] Referring now to FIGS. 3-4, embodiments of SDA process 10 may include a universal speech detector configured to determine one or more intervals that contain speech based on a noisy speech signal. In some embodiments, feature and detector configurations may be tuned based on training data. Average performance over different scenarios may be optimized. In some cases, no information on the environment may be employed. As shown in FIG. 3, a noisy signal may contain various different types of interferences. Voice activity detection (VAD) is sometimes falsely triggered or misses speech.

[0043] Referring now to FIG. 5, an embodiment depicting fixed feature configurations for known scenarios is provided. In some embodiments, this may include selecting 1 of N scenarios that are known from training and using a fixed feature configuration for the current scenario.

[0044] Referring also to FIG. 6, an embodiment of SDA process 10 depicting environmentally dependent dynamic feature control is provided. In accordance with SDA process 10 the system 600 may be configured to estimate acoustical parameters that characterize the environment and to dynamically control configuration based on the parameters. As is shown in FIG. 6, the system may also apply the configuration and apply a weighted sum to emphasize reliable features.

[0045] Referring now to FIG. 7, another embodiment of SDA process 10 depicting environmentally dependent dynamic feature control is provided. When a specific scenario is detected, the dynamic configuration may be overwritten with a configuration optimized in a training (e.g., fall back on standard structure). As is shown in FIG. 8, temporal context for parameter estimation is typically much longer than the context employed in VAD features (e.g., long term properties vs. instantaneous VAD).

[0046] In some embodiments, environmental parameters may include acoustical properties of the environment (e.g., amount of non-stationary components, reverberation, a priori signal-to-noise ratio (SNR), etc.), measures that do not explicitly estimate physical properties, e.g., signal kurtosis, confidences of features (e.g., estimated by rating their performance in the past). In some embodiments, environmental parameters may address a long temporal context, hence, temporal synchronicity of results may be less important and even short dropouts can be compensated. Therefore, the parameters may be calculated in independent threads or can be extracted from results of other algorithms.

[0047] In some embodiments, environmental parameters may be employed to control features (e.g., smoothing parameters), detectors (e.g., thresholds), feature combination (e.g., weighting of different features). Soft control may allow for adaptation to environments that were not considered during training thus providing increased flexibility and reduced dependency on training data.

[0048] Referring now to FIG. 9, another embodiment of SDA process 10 depicting environmentally dependent dynamic feature control is provided. As is shown in the Figure, feature control may be configured to softly map environmental parameters to feature configurations.

[0049] Referring now to FIG. 10, another embodiment of SDA process 10 depicting environmentally dependent dynamic feature control is provided. As is shown in the Figure, control of feature configuration may be realized, e.g., with neural networks, etc.

[0050] As discussed above, in some embodiments of SDA process 10 speech detector configuration may be dynamically controlled based on one or more environmental parameters. Dynamic control of speech detector may influence the configuration of feature parameters and detector parameters (e.g., the combination of features).

[0051] In some embodiments, environmental parameters may characterize the acoustic environment but may not necessarily identify a specific known scenario (e.g., characterization vs. selection). Known scenarios may be identified based on the environmental parameters in addition to the dynamic control. When a known scenario is explicitly identified based on the environmental parameters, a fixed configuration tuned for this specific scenario may be selected instead of the dynamically controlled configuration.

[0052] In some embodiments, the reliability of a VAD feature may be determined by rating the feature's performance in previous frames. This reliability may be employed as an environmental parameter to control the speech detector configuration. The amount of context that is employed for environmental parameter estimation may be different for each environmental parameter and may be dynamically adapted.

[0053] In some embodiments, any calculation of VAD features having a low contribution may be omitted. Accordingly, by focusing on relevant features, the computational costs may be reduced.

[0054] In some embodiments, estimation of environmental parameters may not need to be performed for each frame. Instead, calculations may be distributed to multiple frames to reduce computational costs.

[0055] In some embodiments, environmental parameters may be derived from the results of other algorithms or even non-acoustic information (e.g., video, acceleration sensors, geo-localization, etc.). The environmental parameters may be employed in algorithms other than speech detection, such as noise estimation, noise reduction, dereverberation, or speech recognition.

[0056] Referring now to FIG. 11, an example of a generic computer device 1100 and a generic mobile computer device 1150, which may be used with the techniques described here is provided. Computing device 1100 is intended to represent various forms of digital computers, such as tablet computers, laptops, desktops, workstations, personal digital assistants, servers, blade servers, mainframes, and other appropriate

computers. In some embodiments, computing device 1150 can include various forms of mobile devices, such as personal digital assistants, cellular telephones, smartphones, and other similar computing devices. Computing device 1150 and/or computing device 1100 may also include other devices, such as televisions with one or more processors embedded therein or attached thereto as well as any of the microphones, microphone arrays, and/or speakers described herein. The components shown here, their connections and relationships, and their functions, are meant to be exemplary only, and are not meant to limit implementations of the inventions described and/or claimed in this document.

[0057] In some embodiments, computing device 1100 may include processor 1102, memory 1104, a storage device 1106, a high-speed interface 1108 connecting to memory 1104 and high-speed expansion ports 1110, and a low speed interface 1112 connecting to low speed bus 1114 and storage device 1106. Each of the components 1102, 1104, 1106, 1108, 1110, and 1112, may be interconnected using various busses, and may be mounted on a common motherboard or in other manners as appropriate. The processor 1102 can process instructions for execution within the computing device 1100, including instructions stored in the memory 1104 or on the storage device 1106 to display graphical information for a GUI on an external input/output device, such as display 1116 coupled to high speed interface 1108. In other implementations, multiple processors and/or multiple buses may be used, as appropriate, along with multiple memories and types of memory. Also, multiple computing devices 1100 may be connected, with each device providing portions of the necessary operations (e.g., as a server bank, a group of blade servers, or a multi-processor system).

[0058] Memory 1104 may store information within the computing device 1100. In one implementation, the memory 1104 may be a volatile memory unit or units. In another implementation, the memory 1104 may be a non-volatile memory unit or

units. The memory 1104 may also be another form of computer-readable medium, such as a magnetic or optical disk.

[0059] Storage device 1106 may be capable of providing mass storage for the computing device 1100. In one implementation, the storage device 1106 may be or contain a computer-readable medium, such as a floppy disk device, a hard disk device, an optical disk device, or a tape device, a flash memory or other similar solid state memory device, or an array of devices, including devices in a storage area network or other configurations. A computer program product can be tangibly embodied in an information carrier. The computer program product may also contain instructions that, when executed, perform one or more methods, such as those described above. The information carrier is a computer- or machine-readable medium, such as the memory 1104, the storage device 1106, memory on processor 1102, or a propagated signal.

[0060] High speed controller 1108 may manage bandwidth-intensive operations for the computing device 1100, while the low speed controller 1112 may manage lower bandwidth-intensive operations. Such allocation of functions is exemplary only. In one implementation, the high-speed controller 1108 may be coupled to memory 1104, display 1116 (e.g., through a graphics processor or accelerator), and to high-speed expansion ports 1110, which may accept various expansion cards (not shown). In the implementation, low-speed controller 1112 is coupled to storage device 1106 and low-speed expansion port 1114. The low-speed expansion port, which may include various communication ports (e.g., USB, Bluetooth, Ethernet, wireless Ethernet) may be coupled to one or more input/output devices, such as a keyboard, a pointing device, a scanner, or a networking device such as a switch or router, e.g., through a network adapter.

[0061] Computing device 1100 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a standard server 1120, or multiple times in a group of such servers. It may also be implemented

as part of a rack server system 1124. In addition, it may be implemented in a personal computer such as a laptop computer 1122. Alternatively, components from computing device 1100 may be combined with other components in a mobile device (not shown), such as device 1150. Each of such devices may contain one or more of computing device 1100, 1150, and an entire system may be made up of multiple computing devices 1100, 1150 communicating with each other.

[0062] Computing device 1150 may include a processor 1152, memory 1164, an input/output device such as a display 1154, a communication interface 1166, and a transceiver 1168, among other components. The device 1150 may also be provided with a storage device, such as a microdrive or other device, to provide additional storage. Each of the components 1150, 1152, 1164, 1154, 1166, and 1168, may be interconnected using various buses, and several of the components may be mounted on a common motherboard or in other manners as appropriate.

[0063] Processor 1152 may execute instructions within the computing device 1150, including instructions stored in the memory 1164. The processor may be implemented as a chipset of chips that include separate and multiple analog and digital processors. The processor may provide, for example, for coordination of the other components of the device 1150, such as control of user interfaces, applications run by device 1150, and wireless communication by device 1150.

[0064] In some embodiments, processor 1152 may communicate with a user through control interface 1158 and display interface 1156 coupled to a display 1154. The display 1154 may be, for example, a TFT LCD (Thin-Film-Transistor Liquid Crystal Display) or an OLED (Organic Light Emitting Diode) display, or other appropriate display technology. The display interface 1156 may comprise appropriate circuitry for driving the display 1154 to present graphical and other information to a user. The control interface 1158 may receive commands from a user and convert them for submission to the processor 1152. In addition, an external interface 1162 may be provide in communication with processor 1152, so as to enable near area

communication of device 1150 with other devices. External interface 1162 may provide, for example, for wired communication in some implementations, or for wireless communication in other implementations, and multiple interfaces may also be used.

[0065] In some embodiments, memory 1164 may store information within the computing device 1150. The memory 1164 can be implemented as one or more of a computer-readable medium or media, a volatile memory unit or units, or a non-volatile memory unit or units. Expansion memory 1174 may also be provided and connected to device 1150 through expansion interface 1172, which may include, for example, a SIMM (Single In Line Memory Module) card interface. Such expansion memory 1174 may provide extra storage space for device 1150, or may also store applications or other information for device 1150. Specifically, expansion memory 1174 may include instructions to carry out or supplement the processes described above, and may include secure information also. Thus, for example, expansion memory 1174 may be provide as a security module for device 1150, and may be programmed with instructions that permit secure use of device 1150. In addition, secure applications may be provided via the SIMM cards, along with additional information, such as placing identifying information on the SIMM card in a non-hackable manner.

[0066] The memory may include, for example, flash memory and/or NVRAM memory, as discussed below. In one implementation, a computer program product is tangibly embodied in an information carrier. The computer program product may contain instructions that, when executed, perform one or more methods, such as those described above. The information carrier may be a computer- or machine-readable medium, such as the memory 1164, expansion memory 1174, memory on processor 1152, or a propagated signal that may be received, for example, over transceiver 1168 or external interface 1162.

[0067] Device 1150 may communicate wirelessly through communication interface 1166, which may include digital signal processing circuitry where necessary. Communication interface 1166 may provide for communications under various modes or protocols, such as GSM voice calls, SMS, EMS, or MMS speech recognition, CDMA, TDMA, PDC, WCDMA, CDMA2000, or GPRS, among others. Such communication may occur, for example, through radio-frequency transceiver 1168. In addition, short-range communication may occur, such as using a Bluetooth, WiFi, or other such transceiver (not shown). In addition, GPS (Global Positioning System) receiver module 1170 may provide additional navigation- and location-related wireless data to device 1150, which may be used as appropriate by applications running on device 1150.

[0068] Device 1150 may also communicate audibly using audio codec 1160, which may receive spoken information from a user and convert it to usable digital information. Audio codec 1160 may likewise generate audible sound for a user, such as through a speaker, e.g., in a handset of device 1150. Such sound may include sound from voice telephone calls, may include recorded sound (e.g., voice messages, music files, etc.) and may also include sound generated by applications operating on device 1150.

[0069] Computing device 1150 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a cellular telephone 1180. It may also be implemented as part of a smartphone 1182, personal digital assistant, remote control, or other similar mobile device.

[0070] Various implementations of the systems and techniques described here can be realized in digital electronic circuitry, integrated circuitry, specially designed ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof. These various implementations can include implementation in one or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable

processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device, and at least one output device.

[0071] These computer programs (also known as programs, software, software applications or code) include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. As used herein, the terms "machine-readable medium" "computer-readable medium" refers to any computer program product, apparatus and/or device (e.g., magnetic discs, optical disks, memory, Programmable Logic Devices (PLDs)) used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term "machine-readable signal" refers to any signal used to provide machine instructions and/or data to a programmable processor.

[0072] As will be appreciated by one skilled in the art, the present disclosure may be embodied as a method, system, or computer program product. Accordingly, the present disclosure may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "module" or "system." Furthermore, the present disclosure may take the form of a computer program product on a computer-usable storage medium having computer-usable program code embodied in the medium.

[0073] Any suitable computer usable or computer readable medium may be utilized. The computer-usable or computer-readable medium may be, for example but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, device, or propagation medium. More specific examples (a non-exhaustive list) of the computer-readable medium would include the following: an electrical connection having one or more wires, a portable computer

diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a transmission media such as those supporting the Internet or an intranet, or a magnetic storage device. Note that the computer-usable or computer-readable medium could even be paper or another suitable medium upon which the program is printed, as the program can be electronically captured, via, for instance, optical scanning of the paper or other medium, then compiled, interpreted, or otherwise processed in a suitable manner, if necessary, and then stored in a computer memory. In the context of this document, a computer-usable or computer-readable medium may be any medium that can contain, store, communicate, propagate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

[0074] Computer program code for carrying out operations of the present disclosure may be written in an object oriented programming language such as Java, Smalltalk, C++ or the like. However, the computer program code for carrying out operations of the present disclosure may also be written in conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0075] The present disclosure is described below with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the disclosure. It will be understood

that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0076] These computer program instructions may also be stored in a computer-readable memory that can direct a computer or other programmable data processing apparatus to function in a particular manner, such that the instructions stored in the computer-readable memory produce an article of manufacture including instruction means which implement the function/act specified in the flowchart and/or block diagram block or blocks.

[0077] The computer program instructions may also be loaded onto a computer or other programmable data processing apparatus to cause a series of operational steps to be performed on the computer or other programmable apparatus to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide steps for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0078] To provide for interaction with a user, the systems and techniques described here can be implemented on a computer having a display device (e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user and a keyboard and a pointing device (e.g., a mouse or a trackball) by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback (e.g., visual feedback,

auditory feedback, or tactile feedback); and input from the user can be received in any form, including acoustic, speech, or tactile input.

[0079] The systems and techniques described here may be implemented in a computing system that includes a back end component (e.g., as a data server), or that includes a middleware component (e.g., an application server), or that includes a front end component (e.g., a client computer having a graphical user interface or a Web browser through which a user can interact with an implementation of the systems and techniques described here), or any combination of such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication (e.g., a communication network). Examples of communication networks include a local area network ("LAN"), a wide area network ("WAN"), and the Internet.

[0080] The computing system may include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

[0081] The flowchart and block diagrams in the figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart

illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

[0082] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the disclosure. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, 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.

[0083] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description of the present disclosure has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to the disclosure in the form disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the disclosure. The embodiment was chosen and described in order to best explain the principles of the disclosure and the practical application, and to enable others of ordinary skill in the art to understand the disclosure for various embodiments with various modifications as are suited to the particular use contemplated.

[0084] Having thus described the disclosure of the present application in detail and by reference to embodiments thereof, it will be apparent that modifications and variations are possible without departing from the scope of the disclosure defined in the appended claims.

What Is Claimed Is:

1. A method for speech detection adaptation comprising:
 - receiving, at a processor, a speech signal corresponding to a particular environment;
 - estimating one or more acoustical parameters that characterize the environment, wherein the one or more acoustical parameters are not configured to identify a known scenario; and
 - dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.
2. The method for speech detection adaptation of claim 1, further comprising:
 - after dynamically controlling, identifying at least one known scenario based on the one or more acoustical parameters.
3. The method for speech detection adaptation of claim 2, further comprising:
 - selecting a fixed configuration tuned for the at least one known scenario instead of the dynamically controlled configuration.
4. The method for speech detection adaptation of claim 1, further comprising:
 - determining a reliability of a VAD feature by rating the VAD feature's performance in at least one previous frame.
5. The method for speech detection adaptation of claim 1, further comprising:
 - dynamically adapting an amount of context that is employed for each of the one or more acoustical parameters.
6. The method for speech detection adaptation of claim 1, further comprising:

identifying a VAD feature having an insignificant contribution; and
omitting at least one calculation associated with the VAD feature.

7. The method for speech detection adaptation of claim 1, wherein estimating the one or more acoustical parameters is performed for a subset of frames.
8. The method for speech detection adaptation of claim 1, wherein the one or more acoustical parameters are derived from non-acoustic information.
9. The method for speech detection adaptation of claim 8, wherein the non-acoustic information includes at least one of video, acceleration sensors, and geo-localization information.
10. The method for speech detection adaptation of claim 1, wherein the one or more acoustical parameters are employed in a non-speech detection algorithm.
11. A system for speech detection adaptation comprising:
 - a processor configured to perform one or more operations including:
 - receiving, at a processor, a speech signal corresponding to a particular environment;
 - estimating one or more acoustical parameters that characterize the environment, wherein the one or more acoustical parameters are not configured to identify a known scenario; and
 - dynamically controlling a speech detector based upon, at least in part, the one or more acoustical parameters, wherein dynamically controlling includes configuring feature parameters and detector parameters.
12. The system for speech detection adaptation of claim 11, further comprising:

after dynamically controlling, identifying at least one known scenario based on the one or more acoustical parameters.

13. The system for speech detection adaptation of claim 12, further comprising:
selecting a fixed configuration tuned for the at least one known scenario instead of the dynamically controlled configuration.
14. The system for speech detection adaptation of claim 11, further comprising:
determining a reliability of a VAD feature by rating the VAD feature's performance in at least one previous frame.
15. The system for speech detection adaptation of claim 11, further comprising:
dynamically adapting an amount of context that is employed for each of the one or more acoustical parameters.
16. The system for speech detection adaptation of claim 11, further comprising:
identifying a VAD feature having an insignificant contribution; and
omitting at least one calculation associated with the VAD feature.
17. The system for speech detection adaptation of claim 11, wherein estimating one or more acoustical parameters is performed for a subset of frames.
18. The system for speech detection adaptation of claim 11, wherein the one or more acoustical parameters are derived from non-acoustic information.
19. The system for speech detection adaptation of claim 11, wherein the non-acoustic information includes at least one of video, acceleration sensors, and geo-localization information.

20. The system for speech detection adaptation of claim 11, wherein the one or more acoustical parameters are employed in a non-speech detection algorithm.

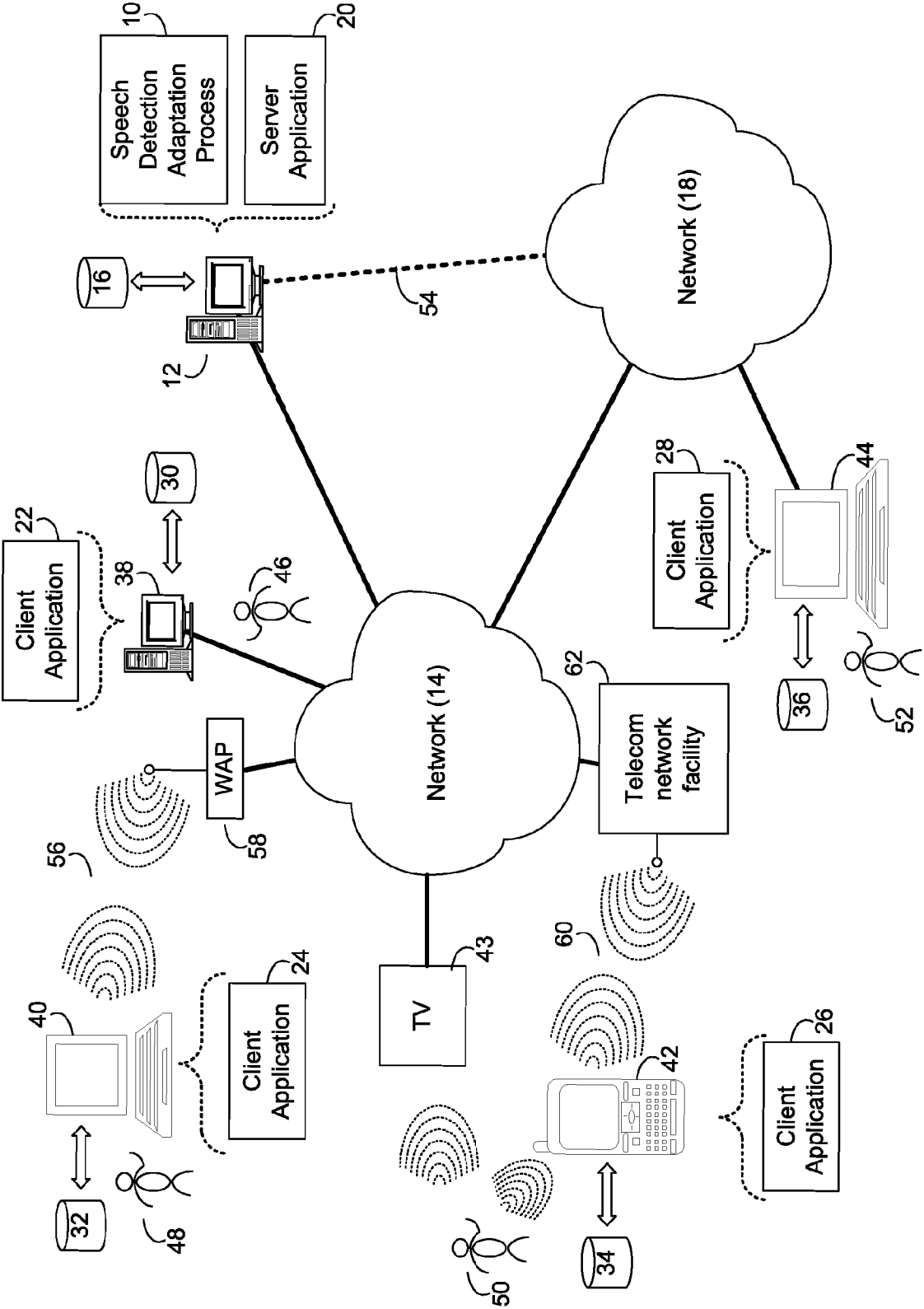


FIG. 1

2/11

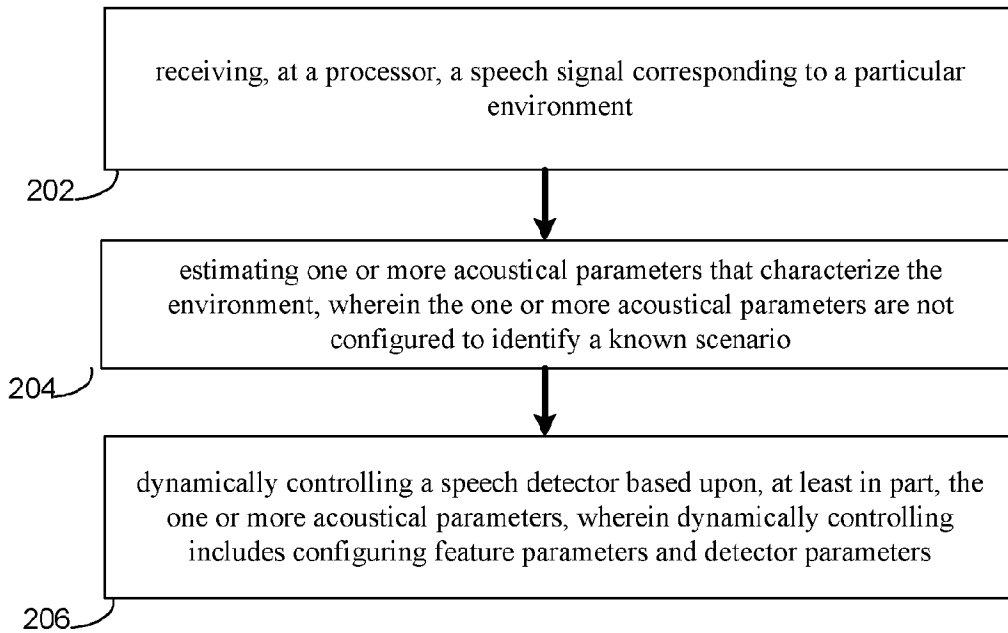
200

FIG. 2

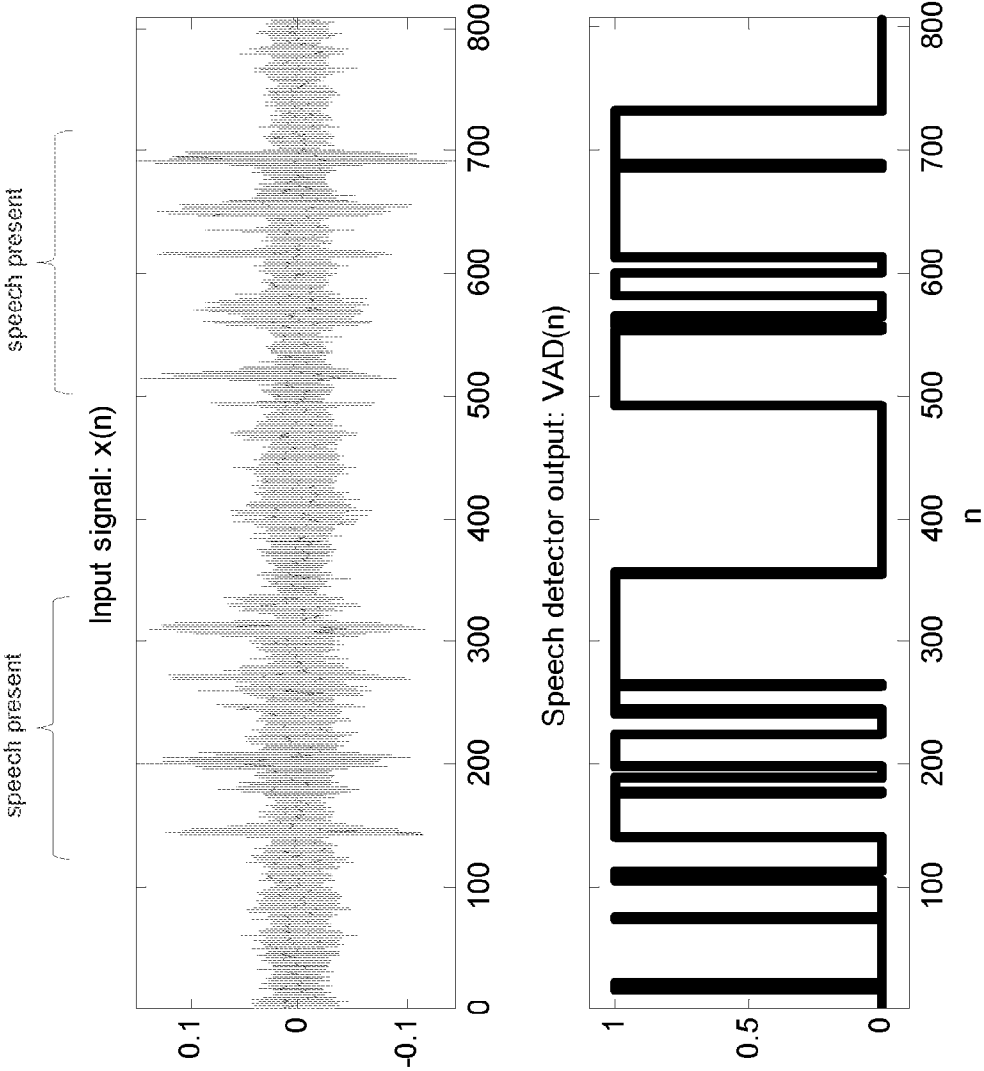


FIG. 3

400

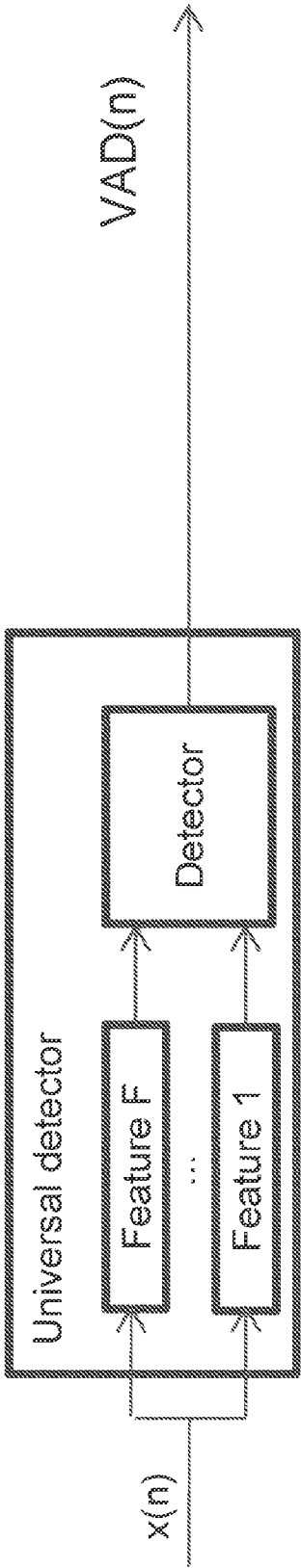


FIG. 4

500

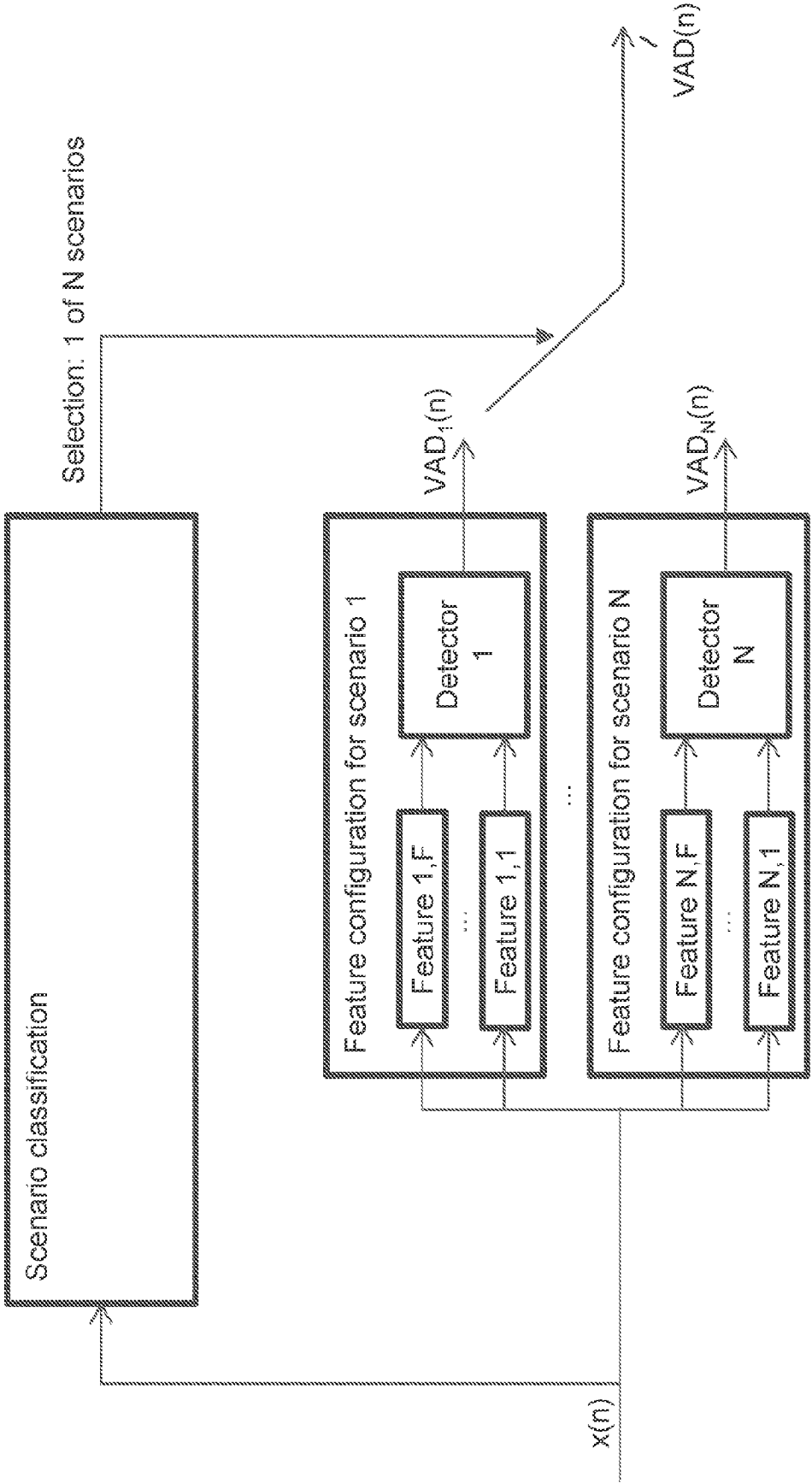
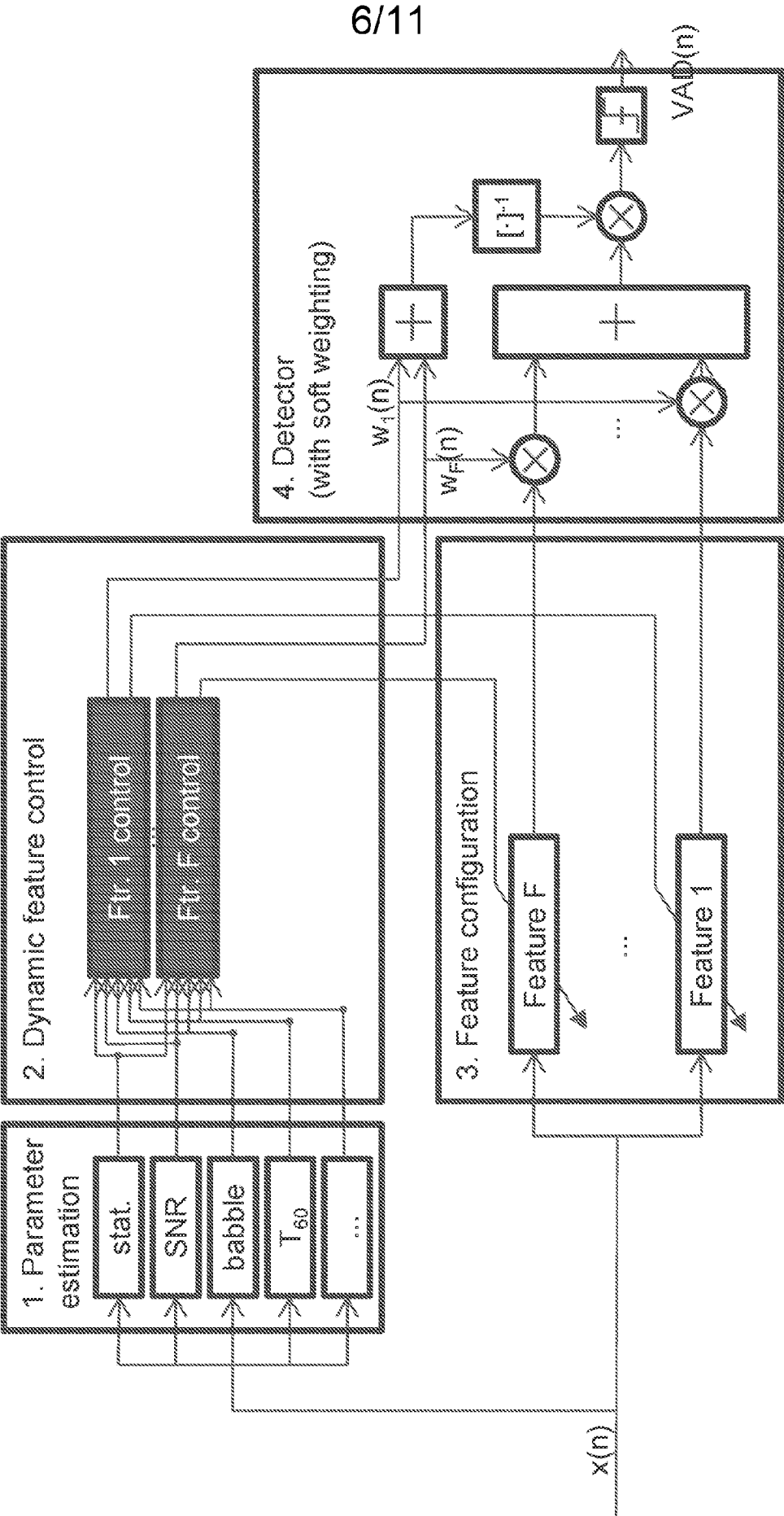


FIG. 5

600



6/11

FIG. 6

7/11

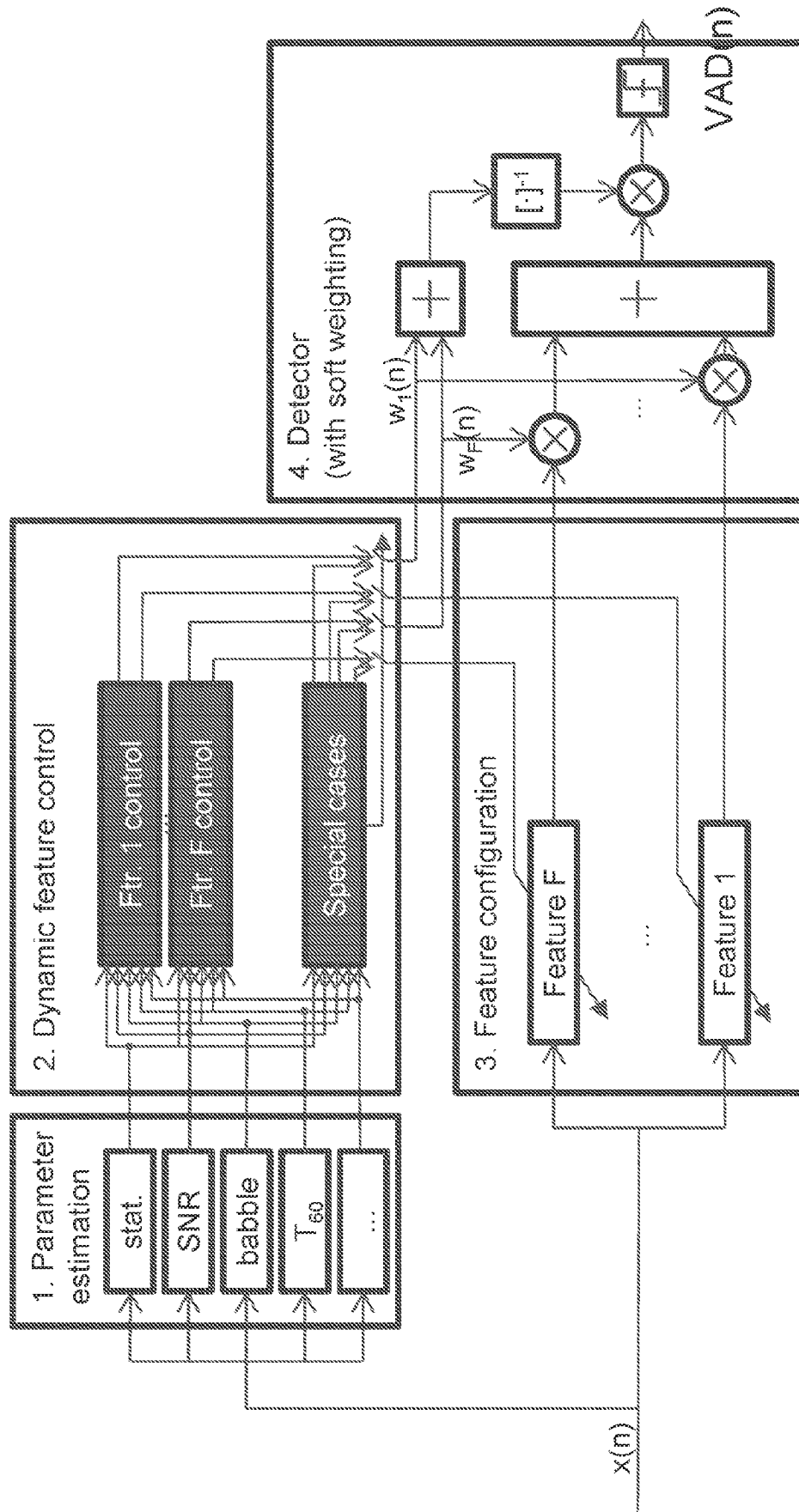
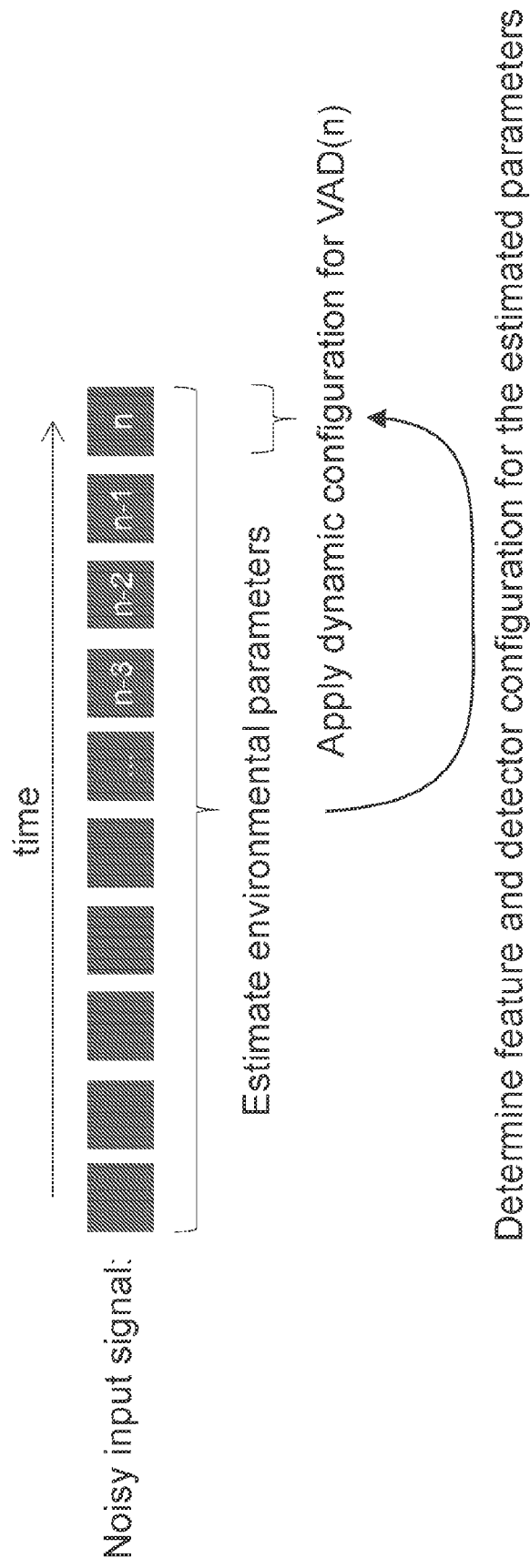


FIG. 7

800



8/11

FIG. 8

900

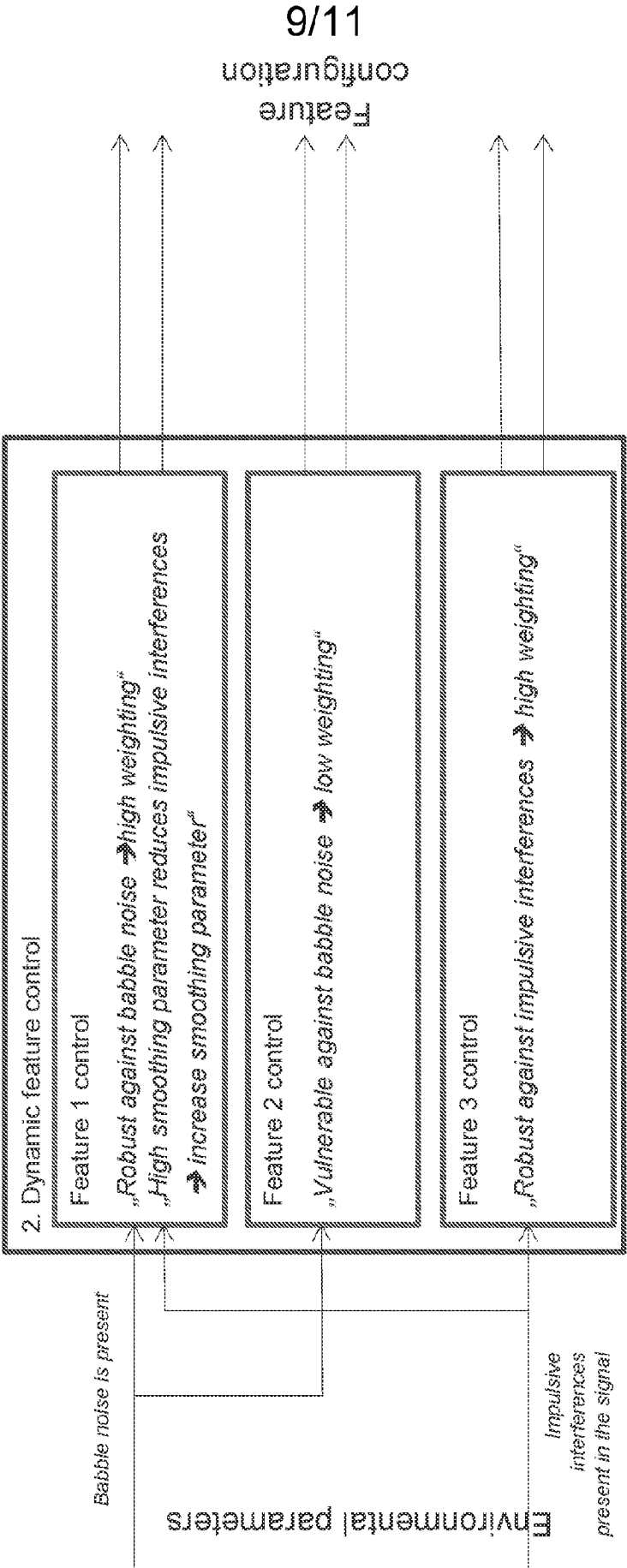
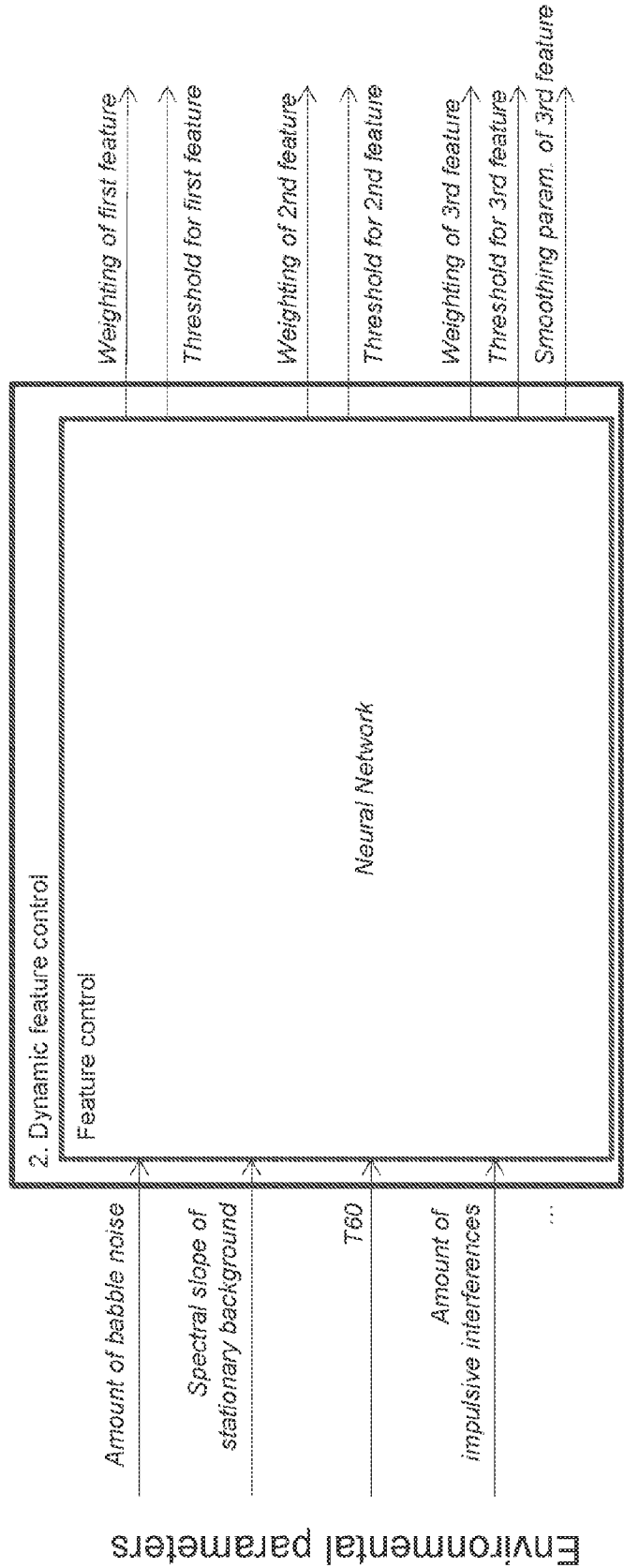


FIG. 9

1000



10/11

FIG. 10

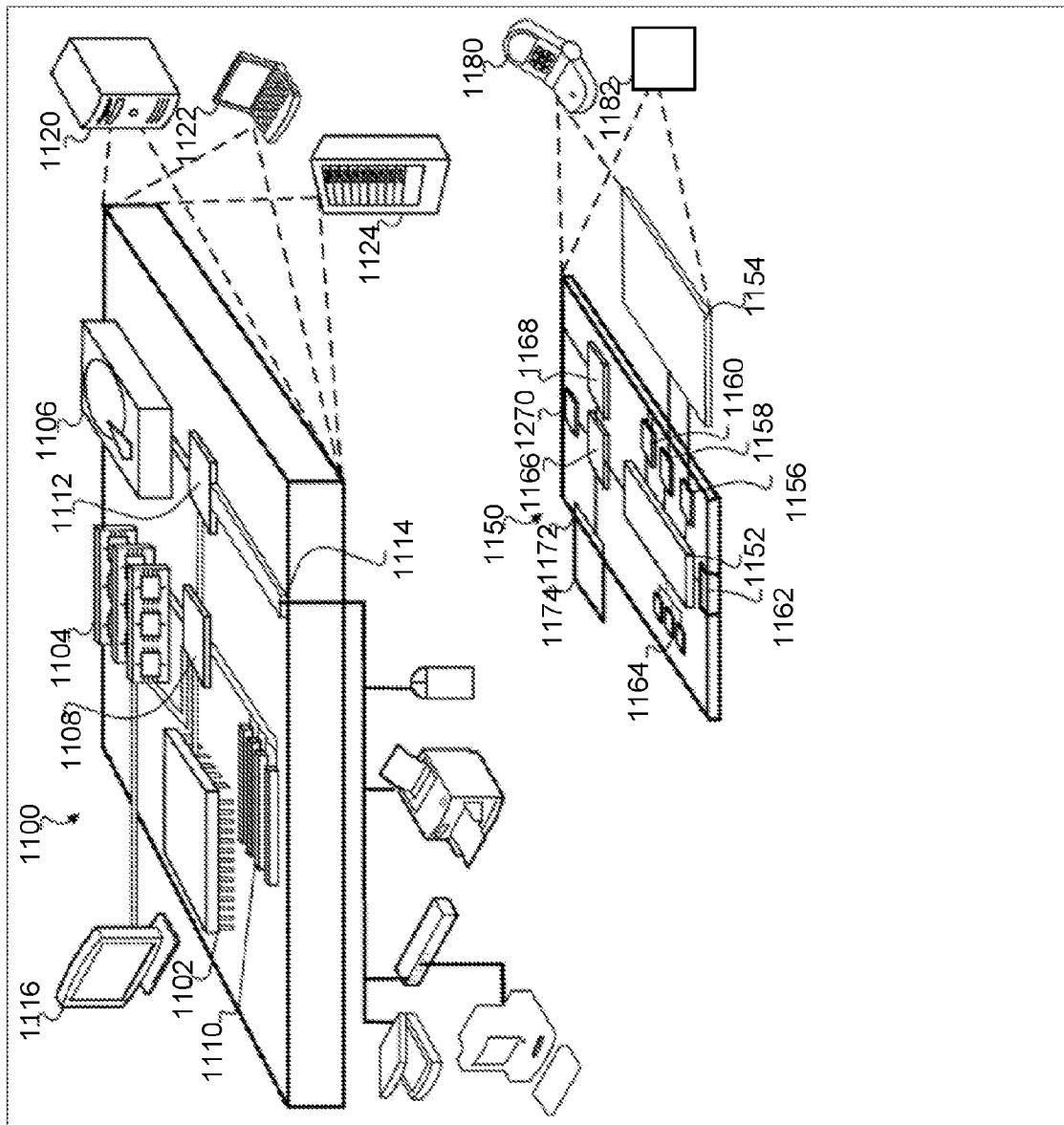


FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2016/012692

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G10L 025/78 (2016.01)

CPC - G10L 2025/786 (2016.01)

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC(8) - G06Q 20/32, 20/00 (2016.01)

CPC - G06F 21/602, 21/35, 21/34; G06Q 20/206, 20/322 (2016.01)

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

USPC - 705/65.000, 67.000; 713/172.000, 169.000 (keyword delimited)

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

PatBase, Orbit, Google Patents, Google Scholar, Google

Search terms used: speech detection, vad context, vad weight, environment, select features

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 8,818,811 B2 (WANG) 26 August 2014 (26.08.2014) entire document	1-3, 7, 10-13, 17, 20
---		---
Y		4-6, 8, 9, 14-16, 18, 19
Y	US 8,503,686 B2 (JING et al) 06 August 2013 (06.08.2013) entire document	4-6, 8, 9, 14-16, 18, 19
A	US 6,044,343 A (CONG et al) 28 March 2000 (28.03.2000) entire document	1-20
A	US 6,993,481 B2 (SKOGLUND et al) 31 January 2006 (31.01.2006) entire document	1-20

☐

Further documents are listed in the continuation of Box C.

☐

See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T"

later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X"

document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y"

document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"

document member of the same patent family

Date of the actual completion of the international search

03 March 2016

Date of mailing of the international search report

12 MAY 2016

Name and mailing address of the ISA/

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774