(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 October 2015 (01.10.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/147662 A1

(51) **International Patent Classification**:
G10L 17/04 (2013.01) *G06K 9/62* (2006.01)

(21) **International Application Number**:
PCT/PL2014/050017

(22) **International Filing Date**:
28 March 2014 (28.03.2014)

(25) **Filing Language**: English

(26) **Publication Language**: English

(71) **Applicant**: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052-8119 (US).

(72) **Inventors**: **BOCKLET, Tobias**; Truderinger Str 110, 81673 Munich, BY (DE). **MAREK, Adam**; 2200 Mission College Boulevard, Santa Clara, CA 95052-8119 (PL).

(74) **Agent**: **BURY, Marek**; Patpol Sp. z o.o., Nowoursynowska 162J, PL-02-776 Warszawa (PL).

(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, LT, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

WO 2015/147662 A1

(54) **Title**: TRAINING CLASSIFIERS USING SELECTED COHORT SAMPLE SUBSETS

(57) **Abstract**: Various systems, apparatuses, and methods for training classifiers using selected cohort sample subsets are disclosed herein. In an example, a set of target supervectors, representing a target class, is received, and a set of cohort supervectors, representing a cohort class, is received. A distance metric is calculated from a respective cohort supervector to a respective target supervector, and a proper subset of cohort supervectors are selected based on the calculated distance metrics. The set of target supervectors and the selected proper subset of cohort supervectors are used to train a classifier. Further examples described herein describe how training classifiers using selected cohort sample subsets may be used to increase performance and decrease resource consumption in voice biometric systems.

TRAINING CLASSIFIERS USING SELECTED COHORT SAMPLE SUBSETS

TECHNICAL FIELD

[0001] Embodiments described herein generally relate to training classifiers using selected cohort sample subsets, and in particular, to training speaker verification classifiers using selected cohort utterance subsets.

BACKGROUND

[0002] Voice biometric systems attempt to verify the claimed identify of a speaker based on a voice sample (e.g., "utterance") from the speaker. Some voice biometric systems utilize machine-learning algorithms, which are trained to distinguish between the target speaker's utterances and other speakers' utterances, known as "cohort/impostor utterances." Increasing the number of cohort utterances may improve the accuracy of the machine-learning algorithm but may also increase the resources and time necessary for the machine-learning algorithm to model the cohort-speaker class and for the classifier to classify an utterance as belonging to either the target-speaker class or the cohort-speaker class, and may have a negative effect on performance.

BRIEF DESCRIPTION OF THE DRAWINGS

[0003] In the drawings, which are not necessarily drawn to scale, like numerals may describe similar components in different views. Like numerals having different letter suffixes may represent different instances of similar components. Some embodiments are illustrated by way of example, and not limitation, in the figures of the accompanying drawings in which:

[0004] FIG. 1 illustrates a system for training a classifier to authenticate a human speaker by using selected cohort speaker sample subsets, in accordance with some embodiments;

[0005] FIG. 2 illustrates a system for classifying a voice authentication attempt using a classifier trained using selected cohort speaker sample subsets, in accordance with some embodiments;

[0006] FIG. 3 illustrates a flowchart for a method for obtaining supervectors from analog audio input, in accordance with some embodiments;

[0007] FIG. 4 illustrates a flowchart for a method for training a classifier, using selected cohort sample subsets, to classify an observation, in accordance with some embodiments;

[0008] FIG. 5 illustrates a block diagram for software and electronic components used to train a classifier to authenticate a human speaker by using selected cohort speaker sample subsets, in accordance with some embodiments; and

[0009] FIG. 6 illustrates a block diagram for an example machine upon which any one or more of the techniques (e.g., operations, processes, methods, and methodologies) discussed herein may be performed, in accordance with some embodiments.

DETAILED DESCRIPTION

[0010] The following description and the drawings illustrate specific embodiments to enable those skilled in the art to practice them. Other embodiments may incorporate structural, logical, electrical, process, and other changes. Portions and features of various embodiments may be included in, or substituted for, those of other embodiments. Embodiments set forth in the claims encompass all available equivalents of those claims.

[0011] Voice biometric systems, which attempt to verify the claimed identify of a speaker based on a voice sample (e.g., "utterance") from the speaker, may be divided into text-dependent and text-independent categories. Text-dependent systems require the user to utter a specific keyword or key-phrase in order to verify the user's identity. Text-independent systems are designed to identify a user by the user's voice, independent of the word(s) or phrase(s) uttered. Text-dependent systems are more suitable for authentication/login scenarios (e.g., telephone banking), whereas text-independent systems are more suited for use in the fields of forensics and secret intelligence, (e.g., wire-tapping).

[0012] A classifier is a process that identifies to which of a set of categories (e.g., sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known. Classifiers, such as Support Vector Machines (SVMs) with or without channel compensation, have often been used in voice biometric systems. Typically, a statistical speaker model, such as a Gaussian Mixture Model (GMM), is created

to model a speaker and a classifier is used to decide whether an utterance was spoken by the speaker. The non-speaker class (e.g., the cohort class) is modeled by a huge set of cohort speakers. Such speaker-model classification systems suffer from at least two drawbacks:

[0013] 1. Modeling the non-speaker class becomes more resource and time consuming as the number of cohort speakers increases.

[0014] 2. Adding too many utterances to the non-speaker class may have a negative effect on the system's performance.

[0015] To overcome these drawbacks, a subset of utterance-specific, non-speaker samples from a set of cohort utterances may be selected and used to model the non-speaker class. A distance metric is calculated to determine the similarity between the cohort utterances and the enrollment/training utterances of a speaker. The "closest" cohort utterances, e.g., utterances with the smallest distance, are then used to model the non-speaker class when training the classifier. This results in a more flexible and cleaner modeling of the non-speaker class because the number of cohort utterances is significantly reduced, thereby improving recognition performance. This approach significantly reduces the computational complexity and memory consumption of the system and makes the system suitable to use on devices with memory and processor constraints, such as application-specific integrated circuits (ASICs).

[0016] FIG. 1 illustrates a system 100 for training a classifier 126 to authenticate a human speaker by using selected cohort speaker sample subsets, in accordance with some embodiments. A target user may wish to enroll into a voice biometric system in order to access a logical and/or physical resource in a secure manner. For example, the target user may wish to enroll into a financial institution's voice biometric system in order to access financial data via telephone. System 100 may be used to enroll the user into such a voice biometric system.

[0017] In some embodiments, system 100 is contained within a single device, such as a smartphone, cellular telephone, mobile phone, laptop computer, tablet computer, desktop computer, server, computer station, computer kiosk, or an ASIC. In some embodiments, the components of system 100 are distributed amongst multiple devices, which may or may not be co-located.

[0018] System 100 includes *n* repetitions of a target training utterance 102 spoken by the target speaker. System 100 also includes various cohort utterances 104 spoken by a plurality of cohort speakers. In some embodiments, the *n* repetitions of a target training utterance 102 and/or the various cohort utterances 104 are received in near real-time by system 100 using an analog audio input component, such as a microphone. In some embodiments, the *n* repetitions of a target training utterance 102 and/or the various cohort utterances 104 are previously recorded audio, and are received or retrieved by system 100.

[0019] Features of speech are extracted 106 from each of the *n* repetitions of a target training utterance 102 spoken by the target speaker. Features of speech are also extracted 108 from the various cohort utterances 104 spoken by the plurality of cohort speakers. In some embodiments, the features of speech extracted may be provided from identified patters or features of audio, such as mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction features (PLPs), Tempo-RAl Patterns (TRAPS), or the like, or other features used in speech verification and/or speech recognition.

[0020] One or more speaker models 112, 114 are adapted to the extracted features 106, 108 to generate statistical target speaker models 116 and statistical cohort speaker models 118, respectively. A universal background model (UBM) is a model trained from numerous hours (e.g., tens or hundreds) of speech data gathered from a large number of speakers. A UBM represents a distribution of the feature vectors that is speaker-independent; thus, a UBM contains data representing general human speech. In some embodiments, during enrollment of a new (target or cohort) speaker into the system, some or all of the parameters of an optional UBM 110 may be adapted to the extracted features 106, 108 of the new speaker to generate the statistical speaker models 116, 118. In some embodiments, the adaptation function is *maximum a posteriori* (MAP), maximum likelihood linear regression (MLLR), or other adaptation functions currently known or unknown in speech verification/recognition arts.

[0021] In some embodiments, one statistical target speaker model 116 is created for each of the *n* repetitions of a target training utterance 102. In some embodiments, the adapted cohort speaker features are converted into statistical cohort speaker models 118. In some embodiments, one statistical cohort speaker model is created for each of the various cohort utterances 104. In some

embodiments, the statistical target speaker models 116 and/or the statistical cohort speaker models 118 are Gaussian Mixture Models (GMMs).

[0022] A supervector, which represents an utterance, is a combination of multiple smaller-dimensional vectors representing features of the utterance, the combination creating one higher-dimensional vector of fixed dimensions. Supervectors are extracted 120, 122 from the statistical target speaker models 116 and the statistical cohort speaker models 118, respectively. In some embodiments, n target speaker supervectors are extracted 120, corresponding to the n repetitions of a target training utterance 102 spoken by the target speaker. A cohort supervector is extracted 122 for each of the various cohort utterances 104 spoken by respective cohort speakers.

[0023] The n extracted target speaker supervectors 120 and the extracted cohort speaker supervectors 122 are used to select 124 a subset of the extracted cohort speaker supervectors 122. In some embodiments, a distance metric is calculated from each cohort speaker supervector to each target speaker supervector, the distance metric representing a similarity between the respective cohort speaker supervector and the respective target speaker supervector. In some embodiments, the distance metric is one of a Mahalanobis, Bhattacharyya, Euclidean, or City Block distance.

[0024] When using City Block distance to calculate the distance metric between supervectors a and b, the following equation may be used:

[0025] $\sum_{i=0}^{D-1} |a_i - b_i|$

[0026] where D is the dimension of supervectors a and b.

[0027] For each target speaker supervector, the k-nearest cohort supervectors are selected. The value of k may vary, depending on the desired accuracy of the classifier 126. The n extracted target speaker supervectors 120 and the selected $k*n$ cohort supervectors 124 are then provided to classifier 126, which uses the supervectors to train to recognize the target speaker's voice. In some embodiments, classifier 126 is a Support Vector Machine (SVM).

[0028] FIG. 2 illustrates a system 200 for classifying a voice authentication attempt 202 using a classifier 126 trained using selected cohort speaker sample subsets, in accordance with some embodiments. In some embodiments, the outcome of the classification of the voice authentication attempt 202 results in allowing or denying some action, such as allowing or denying access to

protected information, or allowing or denying physical access to a protected area or device.

[0029] In some embodiments, system 200 is contained within a single device, such as a smartphone, cellular telephone, mobile phone, laptop computer, tablet computer, desktop computer, server, computer station, computer kiosk, or an ASIC. In some embodiments, the components of system 200 are distributed amongst multiple devices, which may or may not be co-located. In some embodiments, system 200 may be the same device(s) as 100.

[0030] A user makes a voice authentication attempt 202. In some embodiments, the user attempts this voice authentication attempt 202 by uttering the same training utterance used to train the classifier 126. In some embodiments, the user attempts this voice authentication attempt 202 by uttering a different utterance from that which was used to train the classifier 126. In some embodiments, the authentication utterance is received in near real-time by system 200 using an analog audio input component, such as a microphone.

[0031] Features of the user's voice authentication attempt 202 are extracted 204. In some embodiments, the features extracted are MFCCs, PLP, TRAPS, or the like. In some embodiments, the features are extracted using the same process(es) as used in feature extraction 106 and/or 108.

[0032] At this point in the process, it is not yet known whether the user is the same as the target speaker. In some embodiments, a speaker model is adapted 206 to the extracted features 204 to generate a statistical speaker model 208 for the voice authentication attempt 202. In some embodiments, the speaker model is optionally UBM 110. In some embodiments, the extracted features 204 are adapted using MAP adaptation, MLLR adaptation, or other adaptation functions currently known or unknown in speech verification/recognition arts. In some embodiments, the statistical speaker model 208 is a GMM.

[0033] A supervector is then extracted 210 from the statistical speaker model 208. The extracted supervector is then provided to classifier 126, which decides 212 whether the voice authentication attempt 202 was spoken by the claimed speaker. In some embodiments, if the voice authentication attempt 202 was spoken by the claimed speaker, actions such as allowing the claimed speaker access to protected information or physical access to a protected area or device may be performed. In some embodiments, if the voice authentication attempt

202 was not spoken by the claimed speaker, actions such as denying the speaker access to protected information or physical access to a protected area or device may be performed.

[0034] FIG. 3 illustrates a flowchart for a method 300 for obtaining supervectors from analog audio input, in accordance with some embodiments.

[0035] In some embodiments, analog audio input is optionally acquired (operation 305). In some embodiments, the analog audio input may be acquired using an analog audio input component, such as a microphone. In some embodiments, the analog audio input may be acquired from a stored audio recording. In some embodiments, the analog audio input includes repetitions of a training utterance spoken by a target user. In some embodiments, the analog audio input includes cohort utterances spoken by a plurality of cohort speakers.

[0036] In some embodiments, the optionally acquired analog audio input is converted into digital audio (operation 310). In some embodiments, an analog-to-digital converter converts the acquired analog audio input into digital audio.

[0037] Features of speech of each repetition of the training utterance spoken by the target user are extracted from the digital audio (operation 315). In some embodiments, these features may include MFCC, PLP, TRANS, or the like. The digital audio may have been converted from acquired analog audio input (operation 305), or the digital audio may have been received or retrieved from previously converted analog audio input.

[0038] Features of speech of the various utterances spoken by a cohort speaker are extracted from digital audio (operation 320). In some embodiments, these features may include MFCC, PLP, TRANS, or the like. The digital audio may have been converted from acquired analog audio input (operation 305), or the digital audio may have been received or retrieved from previously converted analog audio input.

[0039] A target speaker model is adapted to the extracted features for the target speaker to generate a statistical target speaker model for each repetition of the training utterance by the target speaker (operation 325). In some embodiments, the target speaker model is optionally a UBM (e.g., UBM 110).

[0040] A cohort speaker model is adapted the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for

each utterance spoken by the plurality of cohort speakers (operation 330). In some embodiments, the cohort speaker model is optionally UBM 110.

[0041] A plurality of target supervectors are created by extracting a target supervector from each statistical target speaker model (operation 335), and a plurality of cohort supervectors are created by extracting a cohort supervector from each statistical cohort speaker model (operation 340).

[0042] FIG. 4 illustrates a flowchart for a method 400 for training a classifier 126, using selected cohort sample subsets, to classify an observation, in accordance with some embodiments.

[0043] A plurality of target supervectors, representing a target class, is received or otherwise accessed (operation 405). In some apparatus embodiments, receiving may include reception of signals encoding the target supervectors. In some embodiments, accessing may include requesting a plurality of target supervectors from another component or another device.

[0044] A plurality of cohort supervectors, representing the cohort class, is received or otherwise accessed (operation 410). In some apparatus embodiments, receiving may include reception of signals encoding the cohort supervectors. In some embodiments, accessing may include requesting a plurality of cohort supervectors from another component or another device.

[0045] Distance metrics are calculated from respective cohort supervectors to respective target supervectors. The distance metrics may represent a similarity between the respective cohort supervectors and the respective target supervectors (operation 415).

[0046] Further processing is performed to reduce the number of cohort supervectors. For example, a proper subset of cohort supervectors may be selected, based on the calculated distance metrics, from the plurality of cohort supervectors (operation 420). A proper subset is a subset that is not the same as the original set itself.

[0047] Using the plurality of target supervectors and the proper subset of cohort supervectors, a classifier 126 is trained (operation 425) to classify an observation as belonging to the target class or the cohort class. In some embodiments, a trained classifier 126 is specific to the target speaker, to which the classifier 126 is trained.

[0048] FIG. 5 illustrates a block diagram of software and electronic components 500 used to train a classifier 126 to authenticate a human speaker by using selected cohort speaker sample subsets, within a computer system (such computer system depicted as computing device 502), in accordance with some embodiments. Within the computing device 502, various software and hardware components are implemented in connection with a processor and memory (a processor and memory included in the computing device 502, for example) to train a classifier 126 to authenticate a human speaker by using selected cohort speaker sample subsets or to classify a voice authentication attempt as authentic.

[0049] In some embodiments, computing device 502 includes an analog audio input component 504, such as a microphone for acquiring audio input. This analog audio input component 504 may be integrated into a housing of the computing device 502, or it may be electrically coupled.

[0050] In some embodiments, computing device 502 includes an analog-to-digital converter 506 for converting acquired audio input into digital format.

[0051] In some embodiments, computing device 502 includes a calculation component 508 for calculating a distance metric from a respective cohort supervector to a respective target supervector. In some embodiments, the distance metric represents a similarity between the respective cohort supervector and the respective target supervector.

[0052] In some embodiments, computing device 502 includes a selection component 510 for selecting cohort speaker sample subsets of the cohort speaker supervectors. Selection component 510 selects the cohort sample subsets of the cohort supervectors based on the calculated distance metrics. In some embodiments, in selecting the cohort supervectors, the selection component 510 prefers cohort supervectors with smaller distance metrics to cohort supervectors larger distance metrics. That is, in a set of cohort supervectors with distances 2, 3, 5, 7, and 8, the supervector with distance 2 will be selected before the supervector with distance 3, which will be selected before the supervector with distance 5, etc.

[0053] In some embodiments, computing device 502 includes a classifier 126 that is trained using the target supervectors and the selected cohort speaker sample subsets to recognize the target speaker's voice.

[0054] In some embodiments, computing device 502 is a door lock, a gunlock, a bicycle lock, a vehicle ignition lock, a retail kiosk, a personal computer, a smartphone, a smart television, or combinations thereof.

[0055] FIG. 6 illustrates a block diagram of an example machine 600 upon which any one or more of the techniques (e.g., methodologies) discussed herein may be executed, in accordance with some embodiments. Machine 600 may be embodied by the system 100, system 200, the system performing the operations of method 300, the system performing the operations of method 400, the computing device 502, or some combination thereof.

[0056] In alternative embodiments, the machine 600 may operate as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine 600 may operate in the capacity of a server machine, a client machine, or both in server-client network environments. In an example, the machine 600 may act as a peer machine in peer-to-peer (P2P) (or other distributed) network environment. The machine 600 may be a personal computer (PC), a tablet PC, a set-top box (STB), a personal digital assistant (PDA), a mobile telephone, a web appliance, a network router, switch or bridge, or any machine capable of executing instructions (sequential or otherwise) that specify actions to be taken by that machine. Further, while only a single machine 600 is illustrated, the term "machine" shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein, such as cloud computing, software as a service (SaaS), other computer cluster configurations.

[0057] Examples, as described herein, may include, or may operate on, logic or a number of components, modules, or mechanisms. Modules are tangible entities (e.g., hardware) capable of performing specified operations and may be configured or arranged in a certain manner. In an example, circuits may be arranged (e.g., internally or with respect to external entities such as other circuits) in a specified manner as a module. In an example, the whole or part of one or more computer systems (e.g., a standalone, client or server computer system) or one or more hardware processors may be configured by firmware or software (e.g., instructions, an application portion, or an application) as a module that operates to perform specified operations. In an example, the software may

reside on a machine-readable medium. In an example, the software, when executed by the underlying hardware of the module, causes the hardware to perform the specified operations.

[0058] Accordingly, the term "module" is understood to encompass a tangible entity, be that an entity that is physically constructed, specifically configured (e.g., hardwired), or temporarily (e.g., transitorily) configured (e.g., programmed) to operate in a specified manner or to perform part or all of any operation described herein. Considering examples in which modules are temporarily configured, each of the modules need not be instantiated at any one moment in time. For example, where the modules comprise a general-purpose hardware processor configured using software, the general-purpose hardware processor may be configured as respective different modules at different times. Software may accordingly configure a hardware processor, for example, to constitute a particular module at one instance of time and to constitute a different module at a different instance of time.

[0059] Machine (e.g., computer system) 600 may include a hardware processor 602 (e.g., a central processing unit (CPU), a graphics processing unit (GPU), a hardware processor core, or any combination thereof), a main memory 604 and a static memory 606, some or all of which may communicate with each other via an interlink (e.g., bus) 608. The machine 600 may further include a display unit 610, an alphanumeric input device 612 (e.g., a keyboard), and a user interface (UI) navigation device 614 (e.g., a mouse). In an example, the display unit 610, alphanumeric input device 612, and UI navigation device 614 may be a touch screen display. The machine 600 may additionally include a storage device (e.g., drive unit) 616, a signal generation device 618 (e.g., a speaker), a network interface device 620, and one or more sensors 621, such as a global positioning system (GPS) sensor, compass, accelerometer, or other sensor. The machine 600 may include an output controller 628, such as a serial (e.g., universal serial bus (USB), parallel, or other wired or wireless (e.g., infrared (IR), near field communication (NFC), etc.) connection to communicate or control one or more peripheral devices (e.g., a printer, card reader, etc.).

[0060] The storage device 616 may include a machine-readable medium 622 on which is stored one or more sets of data structures or instructions 624 (e.g., software) embodying or utilized by any one or more of the techniques or

functions described herein. The instructions 624 may also reside, completely or at least partially, within the main memory 604, within static memory 606, or within the hardware processor 602 during execution thereof by the machine 600. In an example, one or any combination of the hardware processor 602, the main memory 604, the static memory 606, or the storage device 616 may constitute machine-readable media.

[0061] Although the machine-readable medium 622 is illustrated as a single medium, the term "machine-readable medium" may include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) configured to store the one or more instructions 624.

[0062] The term "machine-readable medium" may include any medium that is capable of storing, encoding, or carrying instructions 624 for execution by the machine 600 and that cause the machine 600 to perform any one or more of the techniques of the present disclosure, or that is capable of storing, encoding or carrying data structures used by or associated with such instructions 624. Non-limiting machine-readable medium examples may include solid-state memories, and optical and magnetic media. In an example, a massed machine-readable medium comprises a machine-readable medium with a plurality of particles having resting mass. Specific examples of massed machine-readable media may include: non-volatile memory, such as semiconductor memory devices (e.g., Electrically Programmable Read-Only Memory (EPROM), Electrically Erasable Programmable Read-Only Memory (EEPROM)) and flash memory devices; magnetic disks, such as internal hard disks and removable disks; magneto-optical disks; and CD-ROM and DVD-ROM disks.

[0063] The instructions 624 may further be transmitted or received over a communications network 626 using a transmission medium via the network interface device 620 utilizing any one of a number of transfer protocols (e.g., frame relay, internet protocol (IP), transmission control protocol (TCP), user datagram protocol (UDP), hypertext transfer protocol (HTTP), etc.). Example communication networks may include a local area network (LAN), a wide area network (WAN), a packet data network (e.g., the Internet), mobile telephone networks (e.g., cellular networks), Plain Old Telephone (POTS) networks, and wireless data networks (e.g., Institute of Electrical and Electronics Engineers (IEEE) 802.11 family of standards known as Wi-Fi®, IEEE 802.16 family of

standards known as WiMax®), IEEE 802.15.4 family of standards, peer-to-peer (P2P) networks, among others. In an example, the network interface device 620 may include one or more physical jacks (e.g., Ethernet, coaxial, or phone jacks) or one or more antennas to connect to the communications network 626. In an example, the network interface device 620 may include a plurality of antennas to wirelessly communicate using at least one of single-input multiple-output (SIMO), multiple-input multiple-output (MIMO), or multiple-input single-output (MISO) techniques. The term "transmission medium" shall be taken to include any intangible medium that is capable of storing, encoding or carrying instructions 624 for execution by the machine 600, and includes digital or analog communications signals or other intangible medium to facilitate communication of such software.

[0064] The preceding systems, methods, devices, and examples were described in the context of classifying speech. In some embodiments, the preceding systems, methods, devices, and examples may also be used to classify images, videos, non-speech audio, or combinations thereof. For example, a classifier 126 may be trained to classify an image of a target human by providing the classifier 126 images of the target human and images of cohort humans. As another example, a classifier 126 may be trained to classify a video of a target human by providing the classifier 126 videos of the target human and videos of cohort humans.

[0065] Additional examples of the presently described method, system, and device embodiments include the following, non-limiting configurations. Each of the following non-limiting examples may stand on its own, or may be combined in any permutation or combination with any one or more of the other examples provided below or throughout the present disclosure.

[0066] Example 1 includes subject matter (embodied for example by a device, apparatus, machine, or machine-readable medium) of an apparatus to train, using a proper subset of cohort samples, a classifier to classify an observation, the apparatus comprising: a calculation component to calculate, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from a plurality of target supervectors representing a target class, the

respective cohort supervector from a plurality of cohort supervectors representing a cohort class; a selection component to select, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and a training component to train a classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[0067] In Example 2, the subject matter of Example 1 may optionally include a target supervector in the plurality of target supervectors representing an utterance spoken by a target speaker, and a supervector in the plurality of cohort supervectors representing an utterance spoken by a cohort speaker.

[0068] In Example 3, the subject matter of any one or more of Examples 1 to 2 may optionally include a target supervector in the plurality of target supervectors representing an image of a target human, and a cohort supervector in the plurality of cohort supervectors representing an image of a cohort human.

[0069] In Example 4, the subject matter of any one or more of Examples 1 to 3 may optionally include a target supervector in the plurality of target supervectors representing a video of a target human, and a cohort supervector in the plurality of cohort supervectors representing a video of a cohort human.

[0070] In Example 5, the subject matter of any one or more of Examples 1 to 4 may optionally include a target supervector in the plurality of target supervectors representing target audio, and a cohort supervector in the plurality of cohort supervectors representing cohort audio.

[0071] In Example 6, the subject matter of any one or more of Examples 1 to 5 may optionally include an analog audio input component to acquire analog audio input; and an analog-to-digital converter communicatively coupled to the analog audio input component to: receive the analog audio input from the analog audio input component; and convert the analog audio input into digital audio.

[0072] In Example 7, the subject matter of any one or more of Examples 1 to 6 may optionally include the apparatus being further to: extract, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective spoken training repetition; extract, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker; adapt the extracted features

for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker; adapt the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers; create the plurality of target supervectors by extracting a target supervector from respective statistical target speaker models; and create the plurality of cohort supervectors by extracting a cohort supervector from respective statistical cohort speaker models.

[0073] In Example 8, the subject matter of any one or more of Examples 1 to 7 may optionally include the distance metric being one of: City Block, Mahalanobis, Bhattacharya, or Euclidean.

[0074] In Example 9, the subject matter of any one or more of Examples 1 to 8 may optionally include the classifier being a support vector machine.

[0075] Example 10 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-9, to embody subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) of instructions for training a classifier to classify an observation, the training using a proper subset of cohort samples, the instructions which when executed by a machine cause the machine to perform operations including: processing a plurality of target supervectors representing a target class; processing a plurality of cohort supervectors representing a cohort class; calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector; selecting, from the plurality of cohort supervectors and based on the calculated distance metrics, a proper subset of cohort supervectors; and training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[0076] In Example 11, the subject matter of Example 10 may optionally include each target supervector in the plurality of target supervectors representing an utterance spoken by a target speaker, and each cohort supervector in the plurality of cohort supervectors representing an utterance spoken by a cohort speaker.

[0077] In Example 12, the subject matter of any one or more of Examples 10 to 11 may optionally include each target supervector in the plurality of target supervectors representing an image of a target human, and each cohort supervector in the plurality of cohort supervectors representing an image of a cohort human.

[0078] In Example 13, the subject matter of any one or more of Examples 10 to 12 may optionally include each target supervector in the plurality of target supervectors representing a video of a target human, and each cohort supervector in the plurality of cohort supervectors representing a video of a cohort human.

[0079] In Example 14, the subject matter of any one or more of Examples 10 to 13 may optionally include each target supervector in the plurality of target supervectors representing target audio, and each cohort supervector in the plurality of cohort supervectors representing cohort audio.

[0080] In Example 15 the subject matter of any one or more of Examples 10 to 14 may optionally include further instructions, which when executed by the machine, cause the machine to perform operations including: acquiring analog audio input; and converting the analog audio input into digital audio.

[0081] In Example 16 the subject matter of any one or more of Examples 10 to 15 may optionally include further instructions, which when executed by the machine, cause the machine to perform operations including: extracting, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective spoken training repetition; extracting, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker; adapting the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker; adapting the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers; creating the plurality of target supervectors by extracting a target supervector from respective statistical target speaker models; and creating the plurality of cohort supervectors by extracting a cohort supervector from respective statistical cohort speaker models.

[0082] In Example 17 the subject matter of any one or more of Examples 10 to 16 may optionally include the distance metric being one of: City Block, Mahalanobis, Bhattacharya, or Euclidean.

[0083] Example 18 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-17, to embody subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) for training a classifier to classify an observation, the training using a proper subset of cohort samples, the method comprising operations performed by a processor and memory of a computing system, the operations including: processing a plurality of target supervectors representing a target class; processing a plurality of cohort supervectors representing a cohort class; calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector; selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[0084] In Example 19, the subject matter of Example 18 may optionally include each target supervector in the plurality of target supervectors representing an utterance spoken by a target speaker, and each cohort supervector in the plurality of cohort supervectors representing an utterance spoken by a cohort speaker.

[0085] In Example 20, the subject matter of any one or more of Examples 18 to 19 may optionally include each target supervector in the plurality of target supervectors representing an image of a target human, and each cohort supervector in the plurality of cohort supervectors representing an image of a cohort human.

[0086] In Example 21, the subject matter of any one or more of Examples 18 to 20 may optionally include each target supervector in the plurality of target supervectors representing a video of a target human, and each cohort supervector in the plurality of cohort supervectors representing a video of a cohort human.

[0087] In Example 22, the subject matter of any one or more of Examples 18 to 21 may optionally include acquiring analog audio input; and converting the analog audio input into digital audio.

[0088] In Example 23, the subject matter of any one or more of Examples 18 to 22 may optionally include extracting, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective repetition of a training utterance by the target speaker; extracting, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker; adapting the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker; adapting the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers; creating the plurality of target supervectors by extracting a target supervector from a respective statistical target speaker model; and creating the plurality of cohort supervectors by extracting a cohort supervector from a respective statistical cohort speaker model.

[0089] Example 24 includes subject matter for a machine-readable medium including instructions for operation of a computing system, which when executed by a machine, cause the machine to perform operations of any of the methods of Examples 18-23.

[0090] Example 25 includes subject matter for an apparatus comprising means for performing any of the methods of the subject matter of any one of Examples 18 to 23.

[0091] Example 26 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-25, to embody subject matter (e.g., a device, apparatus, machine, or machine-readable medium) of an apparatus for training a classifier to classify an observation, the training using a proper subset of cohort samples, the apparatus comprising: means for processing a plurality of target supervectors representing a target class; means for processing a plurality of cohort supervectors representing a cohort class; means for calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector; means

for selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and means for training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[0092] In Example 27, the subject matter of Example 26 may optionally include each target supervector in the plurality of target supervectors representing an utterance spoken by a target speaker, and each cohort supervector in the plurality of cohort supervectors representing an utterance spoken by a cohort speaker.

[0093] In Example 28, the subject matter of any one or more of Examples 26 to 27 may optionally include each target supervector in the plurality of target supervectors representing an image of a target human, and each cohort supervector in the plurality of cohort supervectors representing an image of a cohort human.

[0094] In Example 29, the subject matter of any one or more of Examples 26 to 28 may optionally include each target supervector in the plurality of target supervectors representing a video of a target human, and each cohort supervector in the plurality of cohort supervectors representing a video of a cohort human.

[0095] In Example 30, the subject matter of any one or more of Examples 26 to 29 may optionally include each target supervector in the plurality of target supervectors representing target audio, and each cohort supervector in the plurality of cohort supervectors representing cohort audio.

[0096] In Example 31, the subject matter of any one or more of Examples 26 to 30 may optionally include means for acquiring analog audio input; and means for converting the analog audio input into digital audio.

[0097] In Example 32, the subject matter of any one or more of Examples 26 to 31 may optionally include means for extracting, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective repetition of a training utterance by the target speaker; means for extracting, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker; means for adapting the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the

training utterance by the target speaker; means for adapting the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers; means for creating the plurality of target supervectors by extracting a target supervector from a respective statistical target speaker model; and means for creating the plurality of cohort supervectors by extracting a cohort supervector from a respective statistical cohort speaker model.

[0098] Example 33 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-32, to embody subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) for enrolling a human user into a voice authentication system, the method comprising operations performed by a processor and memory of a computing system, the operations including: extracting mel-frequency cepstral coefficients (MFCCs) representing features of each repetition of an enrollment utterance spoken by a target speaker; extracting MFCCs representing features of each enrollment utterance spoken by a plurality of cohort speakers; adapting, using *maximum a posteriori* (MAP) adaptation, a Universal Background Model (UBM) to the extracted MFCCs for the target speaker to generate a target speaker Gaussian Mixture Model (GMM) for each repetition of the enrollment utterance by the target speaker; adapting, using MAP adaptation, the UBM to the extracted MFCCs for the plurality of cohort speakers to generate a cohort speaker GMM for each enrollment utterance spoken by the plurality of cohort speakers; creating a plurality of enrollment supervectors by extracting an enrollment supervector from each target speaker GMM; creating a plurality of cohort supervectors by extracting a cohort supervector from each cohort speaker GMM; calculating, from each cohort supervector to each enrollment supervector, a city block distance metric representing a similarity between the cohort supervector and the enrollment supervector, wherein city block distance is the sum of the absolute differences of the projections of a line segment between the n Cartesian coordinates of each supervector; selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and training a Support Vector Machine (SVM) to authenticate the target speaker, the training

initiated by providing the plurality of enrollment supervectors and the selected proper subset of cohort supervectors to the SVM.

[0099] Example 34 includes subject matter (e.g., a device, apparatus, or machine) of an apparatus for performing the operations of Example 33.

[00100] Example 35 includes subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) for enrolling a human user into a voice authentication system, the instructions which when executed by a machine cause the machine to perform the operations of Example 33.

[00101] Example 36 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-35, to embody subject matter subject matter (e.g., a device, apparatus, machine, or machine-readable medium) of an apparatus to train, using a proper subset of cohort samples, a classifier to classify an observation, the apparatus comprising: means for extracting mel-frequency cepstral coefficients (MFCCs) representing features of each repetition of an enrollment utterance spoken by a target speaker; means for extracting MFCCs representing features of each enrollment utterance spoken by a plurality of cohort speakers; means for adapting, using *maximum a posteriori* (MAP) adaptation, a Universal Background Model (UBM) to the extracted MFCCs for the target speaker to generate a target speaker Gaussian Mixture Model (GMM) for each repetition of the enrollment utterance by the target speaker; means for adapting, using MAP adaptation, the UBM to the extracted MFCCs for the plurality of cohort speakers to generate a cohort speaker GMM for each enrollment utterance spoken by the plurality of cohort speakers; means for creating a plurality of enrollment supervectors by extracting an enrollment supervector from each target speaker GMM; means for creating a plurality of cohort supervectors by extracting a cohort supervector from each cohort speaker GMM; means for calculating, from each cohort supervector to each enrollment supervector, a city block distance metric representing a similarity between the cohort supervector and the enrollment supervector, wherein city block distance is the sum of the absolute differences of the projections of a line segment between the n Cartesian coordinates of each supervector; means for selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics;

and means for training a Support Vector Machine (SVM) to authenticate the target speaker, the training initiated by providing the plurality of enrollment supervectors and the selected proper subset of cohort supervectors to the SVM.

[00102] Example 37 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-36, to embody subject matter subject matter (e.g., a device, apparatus, machine, or machine-readable medium) of an apparatus to train, using a proper subset of cohort samples, a classifier to classify an observation, the apparatus comprising: an analog audio input component to acquire analog audio input; an analog-to-digital converter communicatively coupled to the analog audio input component to: receive the analog audio input from the analog audio input component; and convert the analog audio input into digital audio; a calculation component to calculate, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from a plurality of target supervectors representing a target class, the respective cohort supervector from a plurality of cohort supervectors representing a cohort class; a selection component to select, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and a training component to train a classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[00103] In Example 38, the subject matter of Example 37 may optionally include the apparatus being further to: extract mel-frequency cepstral coefficients (MFCCs) representing features of each repetition of an enrollment utterance spoken by a target speaker; extract MFCCs representing features of each utterance spoken by a plurality of cohort speakers; adapt, using *maximum a posteriori* (MAP) adaptation, a Universal Background Model (UBM) to the extracted MFCCs for the target speaker to generate a target speaker Gaussian Mixture Model (GMM) for each repetition of the enrollment utterance by the target speaker; adapt, using MAP adaptation, the UBM to the extracted MFCCs for the plurality of cohort speakers to generate a cohort speaker GMM for each utterance spoken by the plurality of cohort speakers; create the plurality of

enrollment supervectors by extracting an enrollment supervector from each target speaker GMM; and create the plurality of cohort supervectors by extracting a cohort supervector from each cohort speaker GMM.

[00104] In Example 39, the subject matter of any one or more of Examples 37 to 38 may optionally include the apparatus being a door lock.

[00105] In Example 40, the subject matter of any one or more of Examples 37 to 39 may optionally include the apparatus being a gunlock.

[00106] In Example 41, the subject matter of any one or more of Examples 37 to 40 may optionally include the apparatus being a bicycle lock.

[00107] In Example 42, the subject matter of any one or more of Examples 37 to 41 may optionally include the apparatus being a vehicle ignition lock.

[00108] In Example 43, the subject matter of any one or more of Examples 37 to 42 may optionally include the apparatus being a retail kiosk.

[00109] In Example 44, the subject matter of any one or more of Examples 37 to 43 may optionally include the apparatus being a personal computer.

[00110] In Example 45, the subject matter of any one or more of Examples 37 to 44 may optionally include the apparatus being a smartphone.

[00111] In Example 46, the subject matter of any one or more of Examples 37 to 45 may optionally include the apparatus being a smart television.

[00112] Example 47 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-46, to embody subject matter subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) for training a classifier to classify an observation, the training using a proper subset of cohort samples, the method comprising operations performed by a processor and memory of a computing system, the operations including: receiving a plurality of target supervectors representing a target class; receiving a plurality of cohort supervectors representing a cohort class; calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from the plurality of target supervectors, the respective cohort supervector from the plurality of cohort supervectors; selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and

training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[00113] Example 48 includes subject matter (e.g., a method, machine-readable medium, or operations arranged or configured from an apparatus or machine) for enrolling a human user into a voice authentication system, the instructions which when executed by a machine cause the machine to perform the operations of Example 47.

[00114] Example 49 includes subject matter (e.g., a device, apparatus, or machine) of an apparatus for performing the operations of Example 47.

[00115] Example 50 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-49, to embody subject matter subject matter (e.g., a device, apparatus, machine, or machine-readable medium) of an apparatus to train, using a proper subset of cohort samples, a classifier to classify an observation, the training using a proper subset of cohort samples, the apparatus comprising: means for receiving a plurality of target supervectors representing a target class; means for receiving a plurality of cohort supervectors representing a cohort class; means for calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from the plurality of target supervectors, the respective cohort supervector from the plurality of cohort supervectors; means for selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and means for training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

[00116] Example 51 includes, or may optionally be combined with all or portions of the subject matter of one or any combination of Examples 1-50, to embody subject matter subject matter (e.g., a device, apparatus, machine, or machine-readable medium) of an apparatus to train, using a proper subset of cohort samples, a statistical classifier to classify an observation, the apparatus

comprising: a first reception component to receive a plurality of target supervectors representing a target class; a second reception component to receive a plurality of cohort supervectors representing a cohort class; a calculation component to calculate, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from the plurality of target supervectors, the respective cohort supervector from the plurality of cohort supervectors; a selection component to select, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and a training component to train a statistical classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the statistical classifier.

[00117] In Example 52, the subject matter of Example 51 may optionally include the second reception component being the first reception component.

[00118] The above detailed description includes references to the accompanying drawings, which form a part of the detailed description. The drawings show, by way of illustration, specific embodiments that may be practiced. These embodiments are also referred to herein as "examples." Such examples may include elements in addition to those shown or described. However, also contemplated are examples that include the elements shown or described. Moreover, also contemplate are examples using any combination or permutation of those elements shown or described (or one or more aspects thereof), either with respect to a particular example (or one or more aspects thereof), or with respect to other examples (or one or more aspects thereof) shown or described herein.

[00119] In this document, the terms "a" or "an" are used, as is common in patent documents, to include one or more than one, independent of any other instances or usages of "at least one" or "one or more." In this document, the term "or" is used to refer to a nonexclusive or, such that "A or B" includes "A but not B," "B but not A," and "A and B," unless otherwise indicated. In the appended claims, the terms "including" and "in which" are used as the plain-English equivalents of the respective terms "comprising" and "wherein." Also,

in the following claims, the terms "including" and "comprising" are open-ended, that is, a system, device, article, or process that includes elements in addition to those listed after such a term in a claim are still deemed to fall within the scope of that claim. Moreover, in the following claims, the terms "first," "second," and "third," etc. are used merely as labels, and are not intended to suggest a numerical order for their objects.

[00120] The above description is intended to be illustrative, and not restrictive. For example, the above-described examples (or one or more aspects thereof) may be used in combination with others. Other embodiments may be used, such as by one of ordinary skill in the art upon reviewing the above description. Also, in the above Detailed Description, various features may be grouped together to streamline the disclosure. However, the claims may not set forth every feature disclosed herein and embodiments may feature a subset of said features. Further, embodiments may include fewer features than those disclosed in a particular example. Thus, the following claims are hereby incorporated into the Detailed Description, with a claim standing on its own as a separate embodiment. The scope of the embodiments disclosed herein is to be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

CLAIMS

What is claimed is:

1. An apparatus to train, using a proper subset of cohort samples, a classifier to classify an observation, the apparatus comprising:

a calculation component to calculate, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector, the respective target supervector from a plurality of target supervectors representing a target class, the respective cohort supervector from a plurality of cohort supervectors representing a cohort class;

a selection component to select, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and

a training component to train a classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

2. The apparatus of claim 1, wherein a target supervector in the plurality of target supervectors represents an utterance spoken by a target speaker, and wherein a supervector in the plurality of cohort supervectors represents an utterance spoken by a cohort speaker.

3. The apparatus of claim 1, wherein a target supervector in the plurality of target supervectors represents an image of a target human, and wherein a cohort supervector in the plurality of cohort supervectors represents an image of a cohort human.

4. The apparatus of claim 1, wherein a target supervector in the plurality of target supervectors represents a video of a target human, and wherein a cohort supervector in the plurality of cohort supervectors represents a video of a cohort human.

5. The apparatus of claim 1, wherein a target supervector in the plurality of target supervectors represents target audio, and wherein a cohort supervector in the plurality of cohort supervectors represents cohort audio.

6. The apparatus of claim 1, further comprising:

an analog audio input component to acquire analog audio input; and

an analog-to-digital converter communicatively coupled to the analog audio input component to:

receive the analog audio input from the analog audio input component; and

convert the analog audio input into digital audio.

7. The apparatus of claim 6, wherein the apparatus is further to:

extract, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective spoken training repetition;

extract, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker;

adapt the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker;

adapt the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers;

create the plurality of target supervectors by extracting a target supervector from respective statistical target speaker models; and

create the plurality of cohort supervectors by extracting a cohort supervector from respective statistical cohort speaker models.

8. The apparatus of claim 1, wherein the distance metric is one of: City Block, Mahalanobis, Bhattacharya, or Euclidean.

9. The apparatus of claim 1, wherein the classifier is a support vector machine.

10. A machine-readable medium including instructions for training a classifier to classify an observation, the training using a proper subset of cohort samples, the instructions which when executed by a machine cause the machine to perform operations including:

processing a plurality of target supervectors representing a target class;

processing a plurality of cohort supervectors representing a cohort class;

calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector;

selecting, from the plurality of cohort supervectors and based on the calculated distance metrics, a proper subset of cohort supervectors; and

training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

11. The machine-readable medium of claim 10, wherein each target supervector in the plurality of target supervectors represents an utterance spoken by a target speaker, and wherein each cohort supervector in the plurality of cohort supervectors represents an utterance spoken by a cohort speaker.

12. The machine-readable medium of claim 10, wherein each target supervector in the plurality of target supervectors represents an image of a target human, and wherein each cohort supervector in the plurality of cohort supervectors represents an image of a cohort human.

13. The machine-readable medium of claim 10, wherein each target supervector in the plurality of target supervectors represents a video of a target human, and wherein each cohort supervector in the plurality of cohort supervectors represents a video of a cohort human.

14. The machine-readable medium of claim 10, wherein each target supervector in the plurality of target supervectors represents target audio, and wherein each cohort supervector in the plurality of cohort supervectors represents cohort audio.

15. The machine-readable medium of claim 10, further comprising instructions, which when executed by the machine, cause the machine to perform operations including:

acquiring analog audio input; and

converting the analog audio input into digital audio.

16. The machine-readable medium of claim 15, further comprising instructions, which when executed by the machine, cause the machine to perform operations including:

extracting, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective spoken training repetition;

extracting, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker;

adapting the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker;

adapting the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers;

creating the plurality of target supervectors by extracting a target supervector from respective statistical target speaker models; and

creating the plurality of cohort supervectors by extracting a cohort supervector from respective statistical cohort speaker models.

17. The machine-readable medium of claim 10, wherein the distance metric is one of: City Block, Mahalanobis, Bhattacharya, or Euclidean.

18. A method for training a classifier to classify an observation, the training using a proper subset of cohort samples, the method comprising operations performed by a processor and memory of a computing system, the operations including:

processing a plurality of target supervectors representing a target class;

processing a plurality of cohort supervectors representing a cohort class;

calculating, from a respective cohort supervector to a respective target supervector, a distance metric representing a similarity between the respective cohort supervector and the respective target supervector;

selecting, from the plurality of cohort supervectors, a proper subset of cohort supervectors based on the calculated distance metrics; and

training the classifier to classify the observation as belonging to the target class or the cohort class, the training initiated by providing the plurality of target supervectors and the selected proper subset of cohort supervectors to the classifier.

19. The method of claim 18, wherein each target supervector in the plurality of target supervectors represents an utterance spoken by a target speaker, and wherein each cohort supervector in the plurality of cohort supervectors represents an utterance spoken by a cohort speaker.

20. The method of claim 18, wherein each target supervector in the plurality of target supervectors represents an image of a target human, and wherein each cohort supervector in the plurality of cohort supervectors represents an image of a cohort human.

21. The method of claim 18, wherein each target supervector in the plurality of target supervectors represents a video of a target human, and wherein each cohort supervector in the plurality of cohort supervectors represents a video of a cohort human.

22. The method of claim 18, further comprising:

acquiring analog audio input; and

converting the analog audio input into digital audio.

23. The method of claim 22, further comprising:

extracting, from digital audio representing spoken repetitions of a training utterance by a target speaker, features of a respective repetition of a training utterance by the target speaker;

extracting, from digital audio representing various utterances spoken by a plurality of cohort speakers, features of a respective utterance spoken by a cohort speaker;

adapting the extracted features for the target speaker to generate a statistical target speaker model for a respective repetition of the training utterance by the target speaker;

adapting the extracted features for the plurality of cohort speakers to generate a statistical cohort speaker model for a respective utterance spoken by the plurality of cohort speakers;

creating the plurality of target supervectors by extracting a target supervector from a respective statistical target speaker model; and

creating the plurality of cohort supervectors by extracting a cohort supervector from a respective statistical cohort speaker model.

24. A machine-readable medium including instructions for operation of a computing system, which when executed by a machine, cause the machine to perform operations of any of the methods of claims 18-23.

25. An apparatus comprising means for performing any of the methods of claims 18-23.

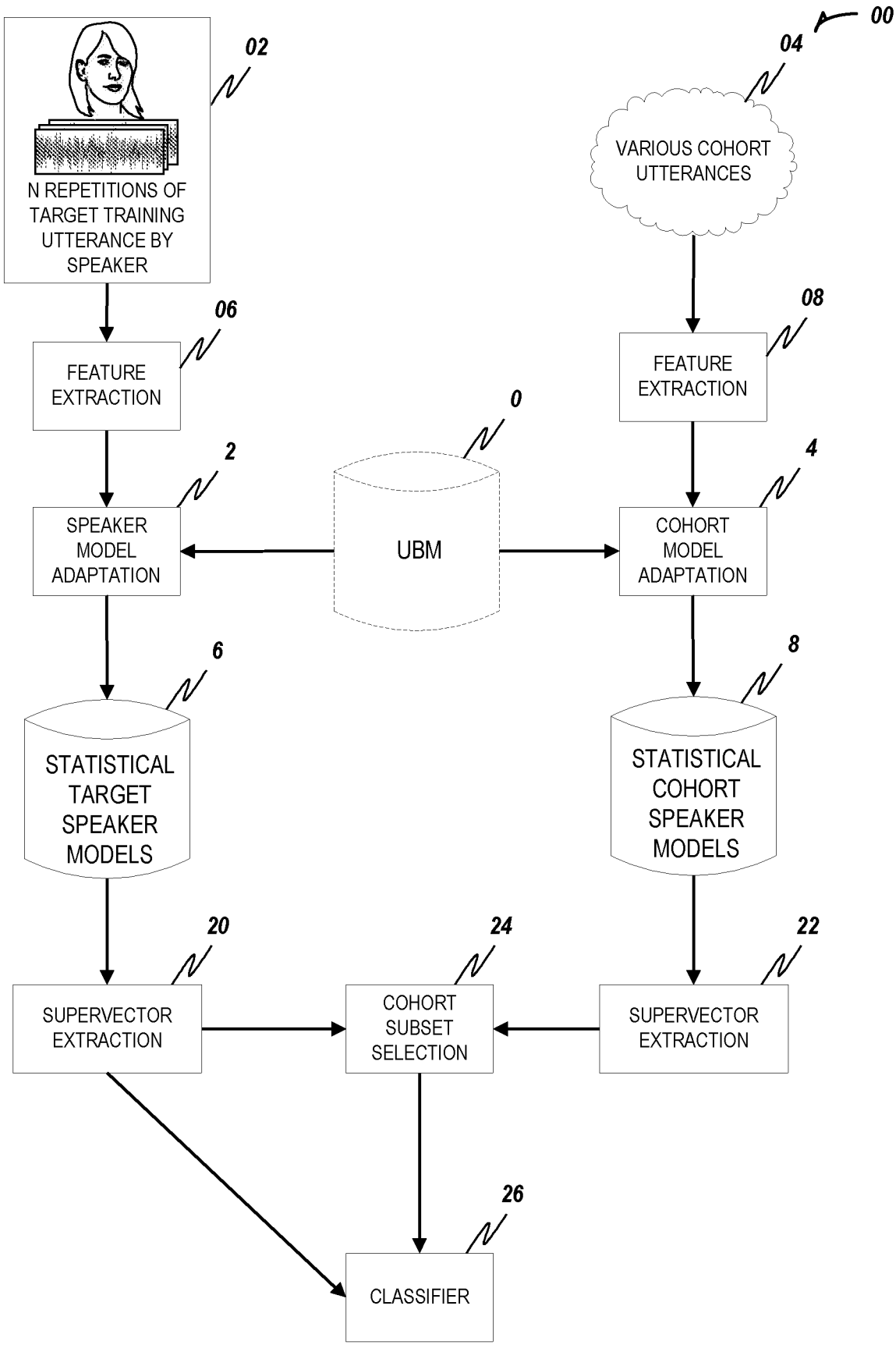
00
02
N REPETITIONS OF TARGET TRAINING UTTERANCE BY SPEAKER
04
VARIOUS COHORT UTTERANCES
06
FEATURE EXTRACTION
08
FEATURE EXTRACTION
0
UBM
2
SPEAKER MODEL ADAPTATION
4
COHORT MODEL ADAPTATION
6
STATISTICAL TARGET SPEAKER MODELS
8
STATISTICAL COHORT SPEAKER MODELS
20
SUPERVECTOR EXTRACTION
24
COHORT SUBSET SELECTION
22
SUPERVECTOR EXTRACTION
26
CLASSIFIER

FIG. 1

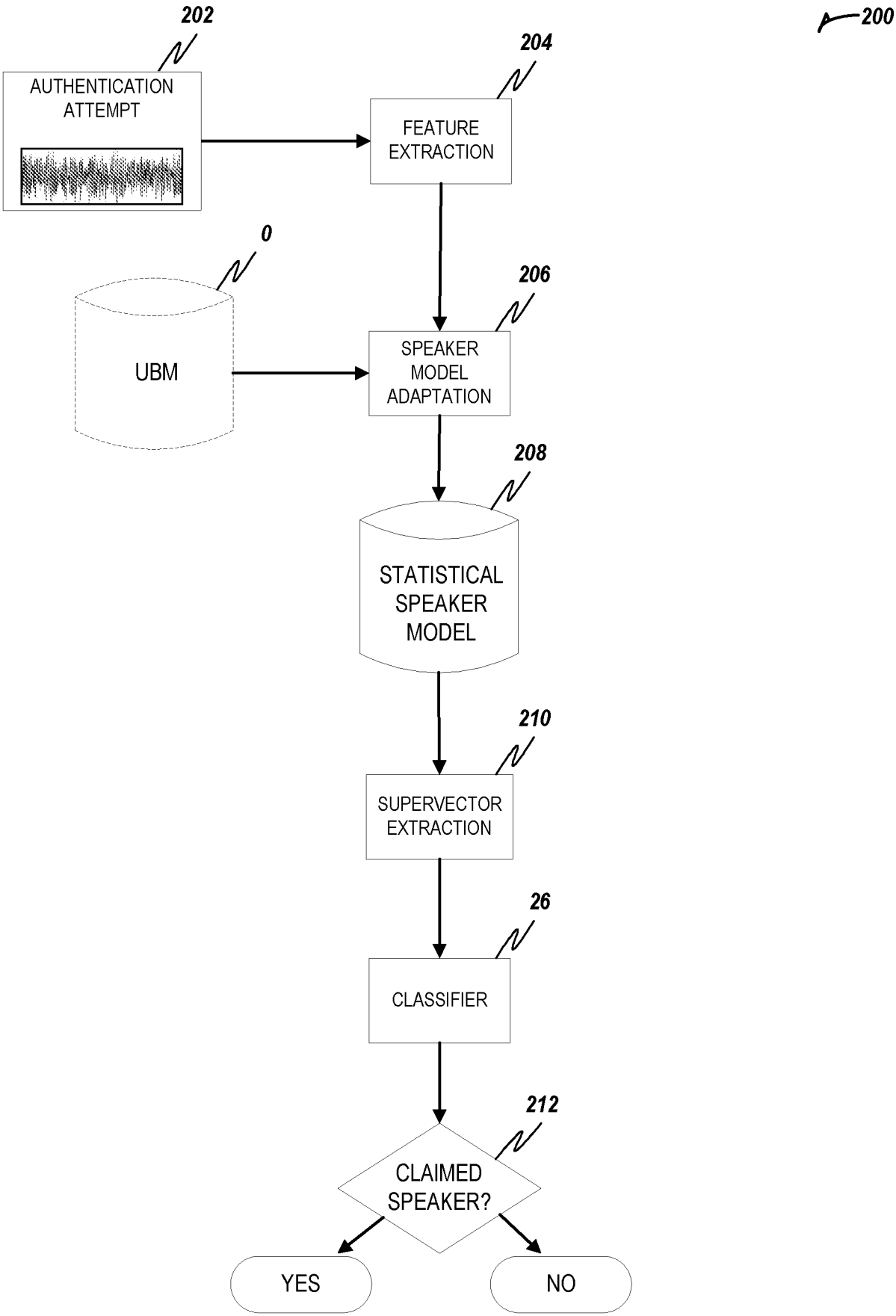
200
202
AUTHENTICATION ATTEMPT
204
FEATURE EXTRACTION
0
UBM
206
SPEAKER MODEL ADAPTATION
208
STATISTICAL SPEAKER MODEL
210
SUPERVECTOR EXTRACTION
26
CLASSIFIER
212
CLAIMED SPEAKER?
YES
NO

FIG. 2

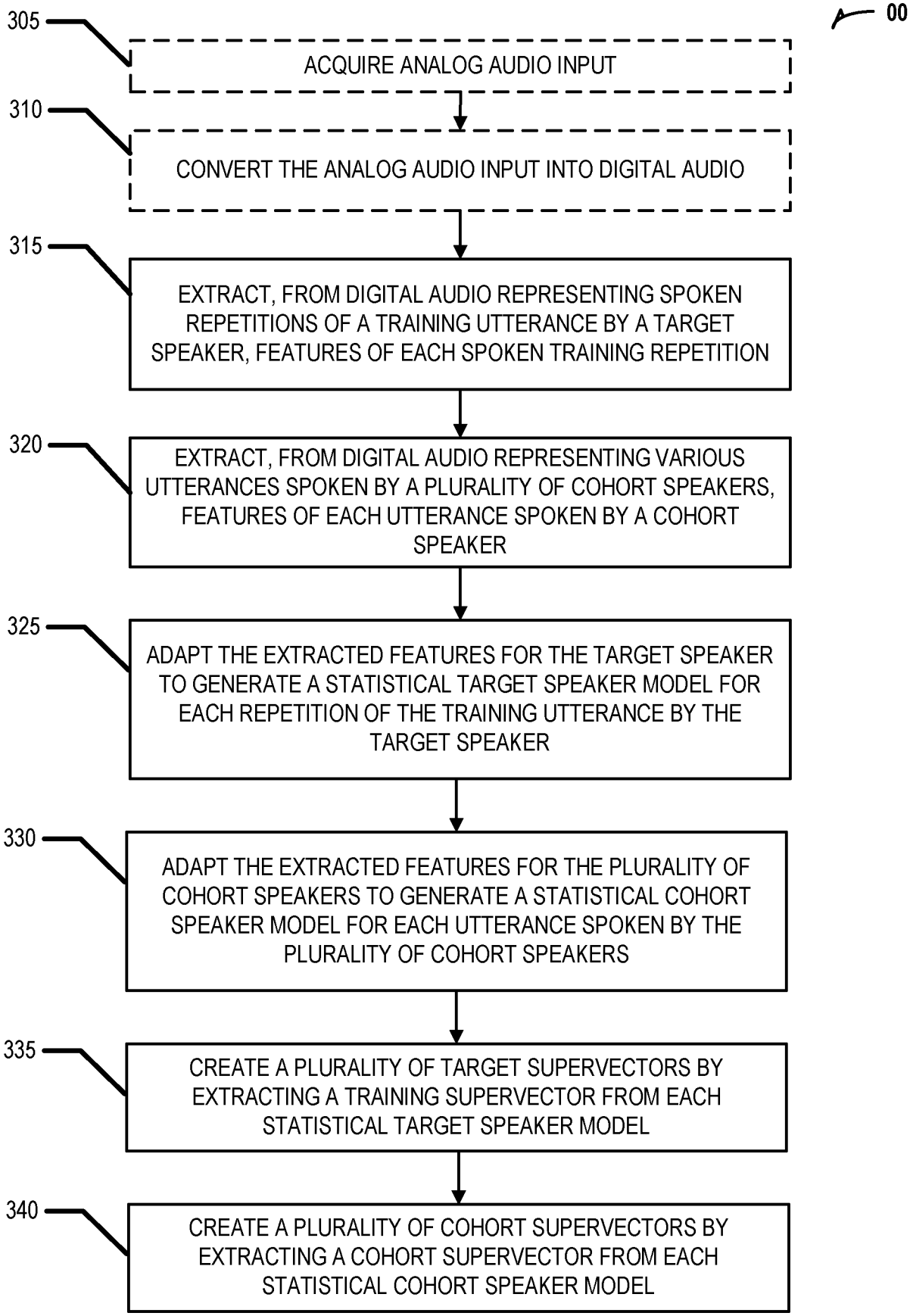
00
305
ACQUIRE ANALOG AUDIO INPUT
310
CONVERT THE ANALOG AUDIO INPUT INTO DIGITAL AUDIO
315
EXTRACT, FROM DIGITAL AUDIO REPRESENTING SPOKEN REPETITIONS OF A TRAINING UTTERANCE BY A TARGET SPEAKER, FEATURES OF EACH SPOKEN TRAINING REPETITION
320
EXTRACT, FROM DIGITAL AUDIO REPRESENTING VARIOUS UTTERANCES SPOKEN BY A PLURALITY OF COHORT SPEAKERS, FEATURES OF EACH UTTERANCE SPOKEN BY A COHORT SPEAKER
325
ADAPT THE EXTRACTED FEATURES FOR THE TARGET SPEAKER TO GENERATE A STATISTICAL TARGET SPEAKER MODEL FOR EACH REPETITION OF THE TRAINING UTTERANCE BY THE TARGET SPEAKER
330
ADAPT THE EXTRACTED FEATURES FOR THE PLURALITY OF COHORT SPEAKERS TO GENERATE A STATISTICAL COHORT SPEAKER MODEL FOR EACH UTTERANCE SPOKEN BY THE PLURALITY OF COHORT SPEAKERS
335
CREATE A PLURALITY OF TARGET SUPERVECTORS BY EXTRACTING A TRAINING SUPERVECTOR FROM EACH STATISTICAL TARGET SPEAKER MODEL
340
CREATE A PLURALITY OF COHORT SUPERVECTORS BY EXTRACTING A COHORT SUPERVECTOR FROM EACH STATISTICAL COHORT SPEAKER MODEL

FIG. 3

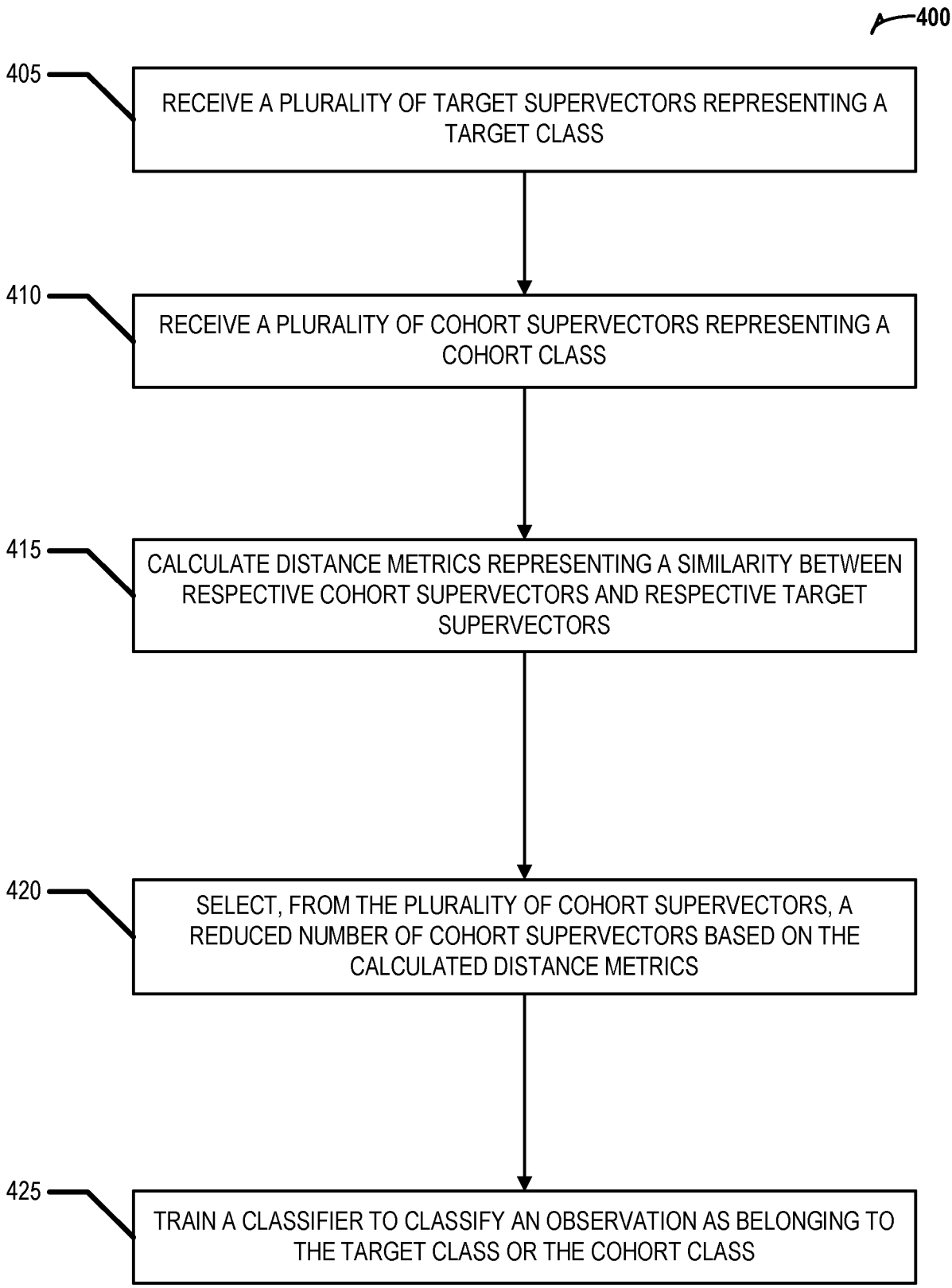
400
405
RECEIVE A PLURALITY OF TARGET SUPERVECTORS REPRESENTING A TARGET CLASS
410
RECEIVE A PLURALITY OF COHORT SUPERVECTORS REPRESENTING A COHORT CLASS
415
CALCULATE DISTANCE METRICS REPRESENTING A SIMILARITY BETWEEN RESPECTIVE COHORT SUPERVECTORS AND RESPECTIVE TARGET SUPERVECTORS
420
SELECT, FROM THE PLURALITY OF COHORT SUPERVECTORS, A REDUCED NUMBER OF COHORT SUPERVECTORS BASED ON THE CALCULATED DISTANCE METRICS
425
TRAIN A CLASSIFIER TO CLASSIFY AN OBSERVATION AS BELONGING TO THE TARGET CLASS OR THE COHORT CLASS

FIG. 4

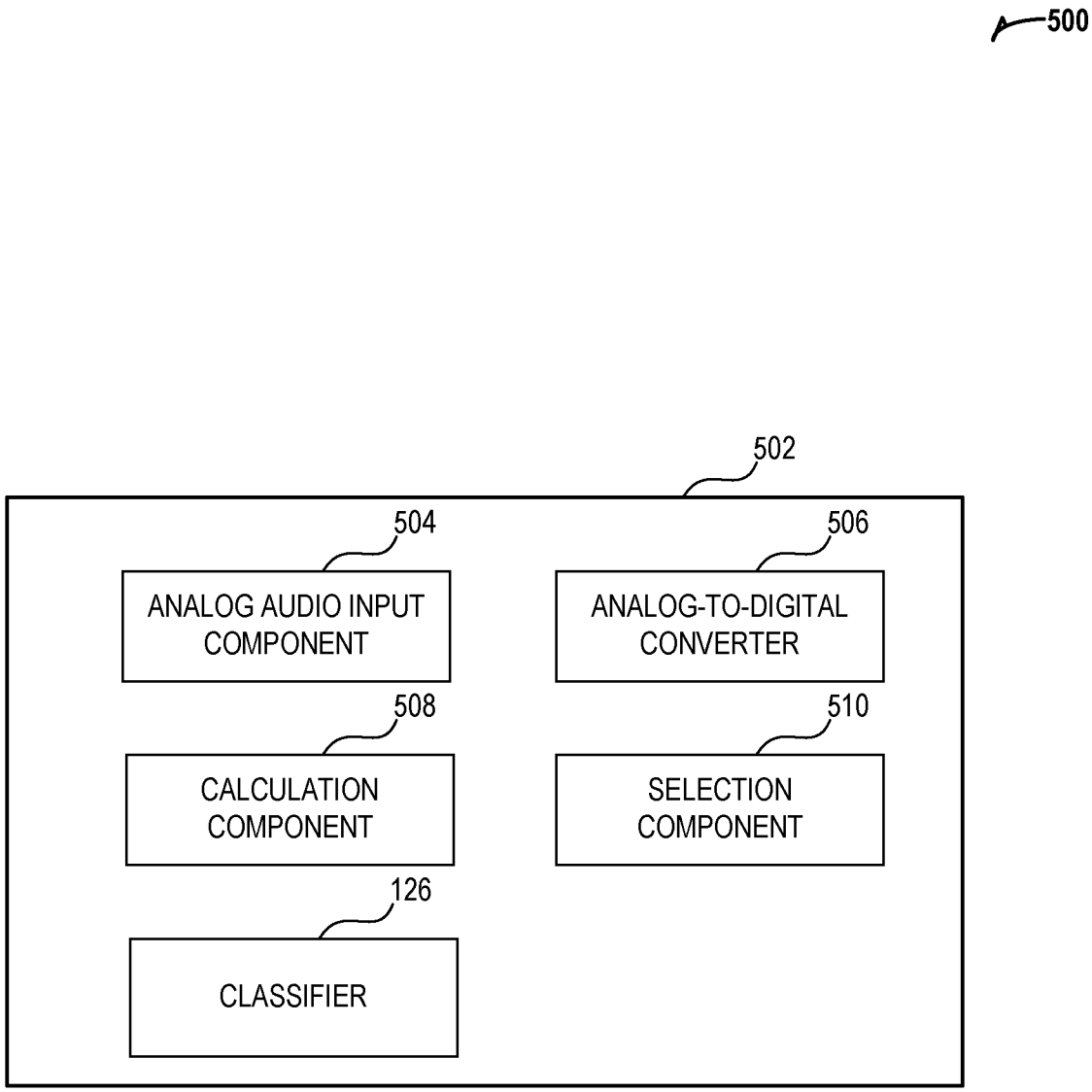
500
502
504
ANALOG AUDIO INPUT COMPONENT
506
ANALOG-TO-DIGITAL CONVERTER
508
CALCULATION COMPONENT
510
SELECTION COMPONENT
126
CLASSIFIER

FIG. 5

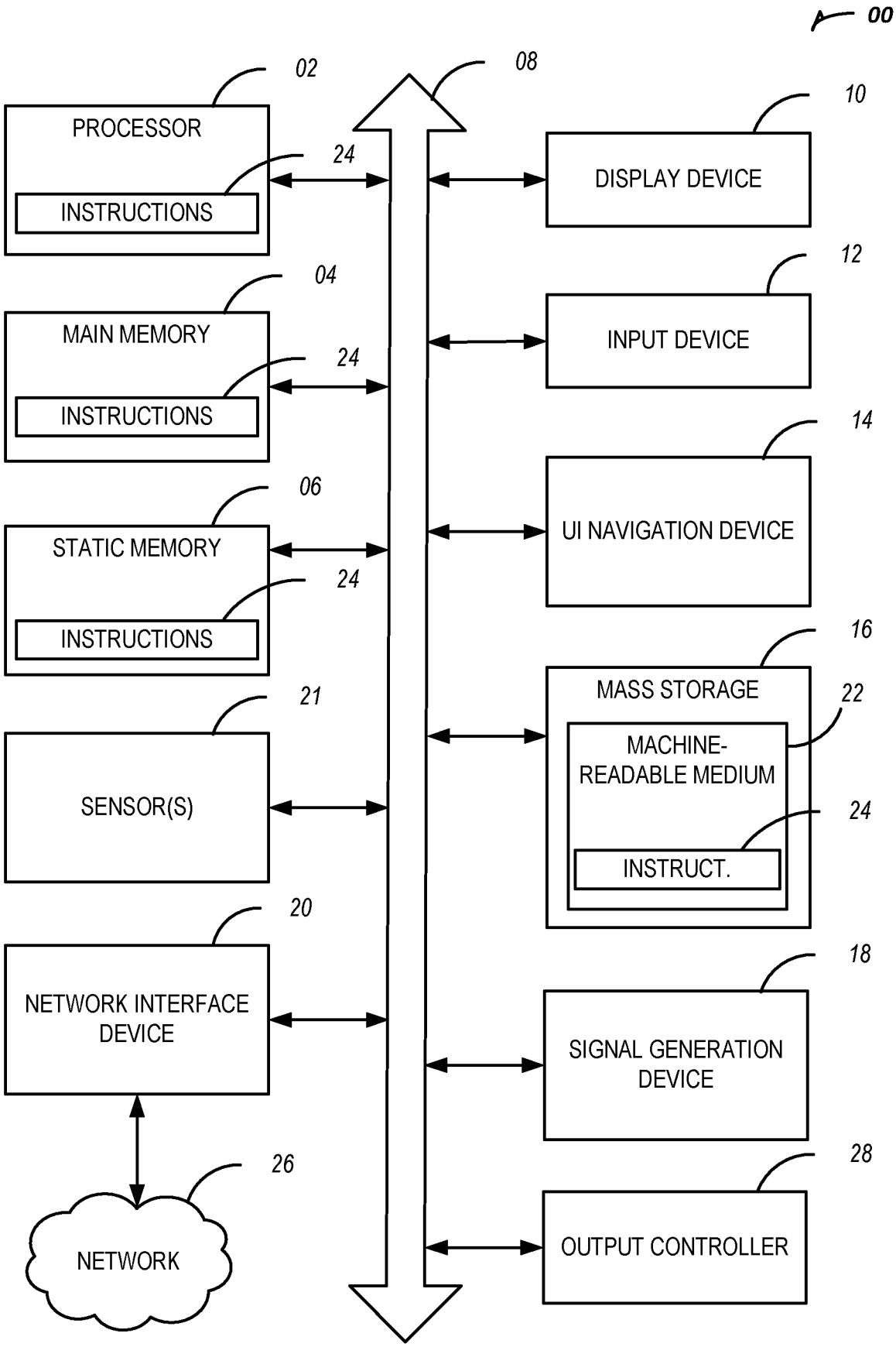
00
02
PROCESSOR
INSTRUCTIONS
24
04
MAIN MEMORY
INSTRUCTIONS
24
06
STATIC MEMORY
INSTRUCTIONS
24
21
SENSOR(S)
20
NETWORK INTERFACE DEVICE
26
NETWORK
08
10
DISPLAY DEVICE
12
INPUT DEVICE
14
UI NAVIGATION DEVICE
16
MASS STORAGE
22
MACHINE-READABLE MEDIUM
24
INSTRUCT.
18
SIGNAL GENERATION DEVICE
28
OUTPUT CONTROLLER

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/PL2014/050017

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L17/04 G06K9/62
ADD.

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)
G10L G06K

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

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

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JOHN H. L. HANSEN ET AL: "Effective background data selection for SVM-based speaker recognition with unseen test environments: more is not always better", INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY, 10 January 2014 (2014-01-10), XP055127882, ISSN: 1381-2416, DOI: 10.1007/s10772-013-9219-z abstract section 3.2; page 4 page 5; figure 4 section 4, first paragraph; page 6 section 5.1 and section 5.3; page 9 ----- -/--	1-25

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
4 August 2014	11/08/2014

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ziegler, Stefan

1

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	WO 2005/043450 A1 (UNIV QUEENSLAND [AU]; GATES KEVIN E [AU]) 12 May 2005 (2005-05-12) abstract page 4, line 14 - line 23 figure 5	1,9,10, 18 2-8, 11-17, 19-25
X	Mitchell Mclaren: "Improving automatic speaker verification using SVM techniques", PhD Thesis - Queensland University of Technology, 30 April 2010 (2010-04-30), pages 1-230, XP055126224, Brisbane Retrieved from the Internet: URL:http://eprints.qut.edu.au/32063/ [retrieved on 2014-07-01] section 6.4.1; page 129 - page 130 equation 6.1; page 130 section 6.4.2; page 130 - page 131 section 4.2.2; 2nd paragraph; page 80 page 73; figure 3.2 section 4.2; page 79 - page 83 section 3.7; page 72	1-25
A	EP 0 887 761 A2 (LUCENT TECHNOLOGIES INC [US]) 30 December 1998 (1998-12-30) figure 3 abstract	1-25
A	W.M. CAMPBELL ET AL: "Support vector machines using GMM supervectors for speaker verification", IEEE SIGNAL PROCESSING LETTERS, vol. 13, no. 5, 1 May 2006 (2006-05-01), pages 308-311, XP055126500, ISSN: 1070-9908, DOI: 10.1109/LSP.2006.870086 the whole document	1-25

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/PL2014/050017

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2005043450 A1	12-05-2005	US 2007203861 A1	30-08-2007
		WO 2005043450 A1	12-05-2005
EP 0887761 A2	30-12-1998	CA 2238164 A1	26-12-1998
		EP 0887761 A2	30-12-1998
		JP H1173406 A	16-03-1999
		US 6134344 A	17-10-2000