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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 December 2013 (12.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/184821 A1

(51) International Patent Classification:
*G10L 15/20* (2006.01) *G10L 21/0216* (2013.01)
*G10L 17/00* (2013.01) *G10L 21/0208* (2013.01)
*G10L 21/028* (2013.01)

(21) International Application Number:
PCT/US2013/044338

(22) International Filing Date:
5 June 2013 (05.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
13/490,161 6 June 2012 (06.06.2012) US

(71) Applicant: **QUALCOMM INCORPORATED** [US/US]; Attn: International IP Administration, 5775 Morehouse Drive, San Diego, California 92121-1714 (US).

(72) Inventors: **BECKLEY, Jeffrey B.**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **AGGARWAL, Pooja**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). **BALASUBRAMANYAM, Shivakumar**; 5775 Morehouse Drive, San Diego, California 92121-1714 (US).

(74) Agent: **JACOBS, Jeffrey D.**; Attn: International IP Administration, 5775 Morehouse Drive, San Diego, Califonia 92121-1714 (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, 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, 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).

**Declarations under Rule 4.17**:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

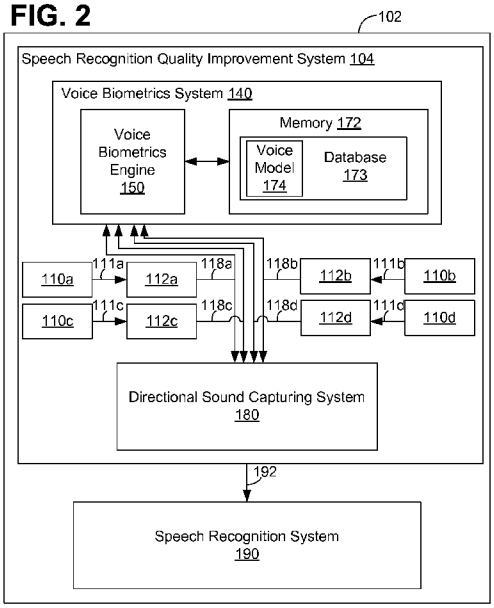
**Published**:

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

*[Continued on next page]*

(54) Title: METHOD AND SYSTEMS HAVING IMPROVED SPEECH RECOGNITION

FIG. 2




(57) Abstract: A method for improving speech recognition by a speech recognition system includes obtaining a voice sample from a speaker; storing the voice sample of the speaker as a voice model in a voice model database; identifying an area from which sound matching the voice model for the speaker is coming; providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

WO 2013/184821 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

# METHOD AND SYSTEMS HAVING IMPROVED SPEECH RECOGNITION

## BACKGROUND

### 1. Field

**[0001]** The disclosure relates generally to the field of speech recognition systems and methods, and, in particular, to speech recognition systems and methods having improved speech recognition.

### 2. Background

**[0002]** Speech recognition (SR) (also commonly referred to as voice recognition) represents one of the most important techniques to endow a machine with simulated intelligence to recognize user or user-voiced commands and to facilitate human interface with the machine. SR also represents a key technique for human speech understanding. Systems that employ techniques to recover a linguistic message from an acoustic speech signal are called voice recognizers. The term "speech recognizer" is used herein to mean generally any spoken-user-interface-enabled device or system.

**[0003]** The use of SR is becoming increasingly important for safety reasons. For example, SR may be used to replace the manual task of pushing buttons on a wireless telephone keypad. This is especially important when a user is initiating a telephone call while driving a car. When using a phone without SR, the driver must remove one hand from the steering wheel and look at the phone keypad while pushing the buttons to dial the call. These acts increase the likelihood of a car accident. A speech-enabled phone (i.e., a phone designed for speech recognition) would allow the driver to place telephone calls while continuously watching the road. In addition, a hands-free car-kit system would permit the driver to maintain both hands on the steering wheel during call initiation.

**[0004]** Speech recognition (ASR) systems, such as always-on speech recognition (ASR) systems, have difficulty handling ambient noise, such as background conversation or other undesired noise. The presence of background conversation, for instance, may cause the system to recognize a command that was not intended by the user of the system, thus leading to a number of false positives and misrecognitions. For example, in a car environment, while the driver might be the issuer of voice commands to control the ASR system, the presence of passengers and ensuing conversation

between them can greatly reduce the performance of the ASR system as their conversations may lead to commands that are false positives or misrecognitions.

## SUMMARY

**[0005]** A method for improving speech recognition by a speech recognition system includes, but is not limited to any one or combination of: (i) obtaining a voice sample from a speaker; (ii) storing the voice sample of the speaker as a voice model in a voice model database; (iii) identifying an area from which sound matching the voice model for the speaker is coming; and (iv) providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

**[0006]** In various embodiments, the method further includes controlling microphones to receive sound from the identified area.

**[0007]** In some embodiments, the method further includes controlling the microphones to not receive sound from areas other than the identified area.

**[0008]** In various embodiments, the method further includes filtering sound received from areas other than the identified area.

**[0009]** In various embodiments, the method further includes separating ambient noise from speech in the one or more audio signals; and providing the resulting speech signal to the speech recognition system for processing.

**[0010]** In various embodiments, the identifying includes receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas; and determining whether the received audio signals corresponding to the sound coming from the first area match the voice model for the speaker. The providing includes selectively providing audio signals corresponding to sound from the first area to the speech recognition system for processing thereof based on the determination.

**[0011]** In some embodiments, if the one or more received audio signals corresponding to the sound coming from the first area match the voice model for the speaker, one or more audio signals corresponding to sound from the first area are provided to the speech recognition system for processing thereof. If the one or more received audio signals corresponding to the sound coming from the first area do not match the voice model for

the speaker, one or more audio signals corresponding to sound from the first area are not provided to the speech recognition system for processing thereof.

**[0012]** In some embodiments, one or more audio signals corresponding to sound from the first area received after the determination are selectively provided to the speech recognition system for processing thereof based on the determination.

**[0013]** In some embodiments, the one or more microphones comprise at least two microphones.

**[0014]** In some embodiments, the method further includes: receiving, from the one or more microphones, one or more audio signals corresponding to sound coming from a second area of the plurality of areas; determining whether the one or more received audio signals corresponding to the sound coming from the second area match the voice model for the speaker; and selectively providing audio signals corresponding to sound from the second area to the speech recognition system for processing thereof based on the determination.

**[0015]** In further embodiments, if the one or more audio signals corresponding to the sound coming from the second area match the voice model for the speaker, one or more audio signals corresponding to sound from the second area are provided to the speech recognition system for processing thereof. If the one or more audio signals corresponding to the sound coming from the second area do not match the voice model for the speaker, one or more audio signals corresponding to sound from the second area are not provided to the speech recognition system for processing thereof.

**[0016]** In some embodiments, the determining includes: characterizing the sound coming from the first area; and comparing the characterized sound to the voice model for the speaker.

**[0017]** In further embodiments, if the characterized sound compares with the voice model for the speaker, the sound coming from the first area is determined to be speech from the speaker and one or more audio signals corresponding to sound coming from the first area is provided to the speech recognition system. If the characterized sound does not compare with the voice model for the speaker, the sound coming from the first area is determined not to be speech from the speaker and one or more audio signals corresponding to sound coming from the first area is not provided to the speech recognition system.

**[0018]** In further embodiments, the sound coming from the first area is characterized through analysis of an acoustic spectrum of the sound.

**[0019]** In further embodiments, the one or more microphones comprises a first microphone associated with a first area of the plurality of areas and a second microphone associated a second area of the plurality of areas. If the one or more received audio signals corresponding to the sound coming from the first area matches the voice model for the speaker, one or more audio signals corresponding to sound from the first area are received from the first microphone for providing to the speech recognition system. If the one or more received audio signals corresponding to the sound coming from the first area do not matches the voice model for the speaker, one or more audio signals corresponding to sound from the first area are not received from the first microphone for providing to the speech recognition system.

**[0020]** In yet further embodiments, if the one or more received audio signals corresponding to the sound coming from the first area matches the voice model for the speaker and the one or more received audio signals corresponding to the sound coming from the second area does not match the voice model for the speaker, one or more audio signals corresponding to sound from the first area are received from the first microphone for providing to the speech recognition system. If the one or more received audio signals corresponding to the sound coming from the first area does not match the voice model for the speaker and the one or more received audio signals corresponding to the sound coming from the second area matches the voice model for the speaker, one or more audio signals corresponding to sound from the second area are received from the second microphone.

**[0021]** In various embodiments, the audio signals provided to the speech recognition system correspond to speech commands for the speech recognition system. The speech recognition system issues instructions based on the speech.

**[0022]** An apparatus for improving speech recognition by a speech recognition system, includes, but is not limited to any one or combination of: means for obtaining a voice sample from a speaker; means for storing the voice sample of the speaker as a voice model in a voice model database; means for identifying an area from which sound matching the voice model for the speaker is coming; and means for providing one or

more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

**[0023]** An apparatus for improving speech recognition by a speech recognition system includes a processor configured for (but is not limited to any one or combination of): obtaining a voice sample from a speaker; storing the voice sample of the speaker as a voice model in a voice model database; receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas; identifying an area from which sound matching the voice model for the speaker is coming; and providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

**[0024]** A computer program product for improving speech recognition by a speech recognition system includes a computer-readable storage medium comprising code for (but is not limited to any one or combination of): obtaining a voice sample from a speaker; storing the voice sample of the speaker as a voice model in a voice model database; receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas; identifying an area from which sound matching the voice model for the speaker is coming; and providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0025]** FIG. 1 is an illustration of an electronics device according to various embodiments of the disclosure.

**[0026]** FIG. 2 is a block diagram of various components in an electronics device according to various embodiments of the disclosure.

**[0027]** FIG. 3 is a flow chart of a speaker authentication method according to various embodiments of the disclosure.

**[0028]** FIG. 4A illustrates a flow chart of a method for improving speech quality recognition according to various embodiments of the disclosure.

**[0029]** FIG. 4B illustrates a flow chart of a method for improving speech quality recognition according to various embodiments of the disclosure.

**[0030]** FIG. 5 is a block diagram of a directional sound capturing system according to various embodiments of the disclosure.

**[0031]** FIG. 6A is a block diagram of a directional sound capturing system according to various embodiments of the disclosure.

**[0032]** FIG. 6B is a block diagram of a directional sound capturing system according to various embodiments of the disclosure.

**[0033]** FIG. 7A illustrates a flow chart of a method for improving speech quality recognition according to various embodiments of the disclosure.

**[0034]** FIG. 7B illustrates a flow chart of a method for improving speech quality recognition according to various embodiments of the disclosure.

**[0035]** FIG. 8 is a block diagram illustrating a communication device according to various embodiments of the disclosure.

## DETAILED DESCRIPTION

**[0036]** Various embodiments relate to a system for improving quality of speech recognition performed by a speech recognition system. Particular embodiments relate to a speech recognition quality improvement (SRQI) system implementing voice biometrics to authenticate speakers and directional sound capturing for discriminating (e.g., via directional filtering and/or directional tracking by microphones) speech from the authenticated speaker against undesired sound (e.g., ambient noise, background conversations, etc.) such that the speech is provided to and/or processed by the speech recognition system while reducing undesired sound being provided to and/or processed by the speech recognition system.

**[0037]** For instance, in some embodiments, the SRQI system may include a voice biometrics system for performing speaker identification by providing a voice sample to enroll a speaker (user) as a valid speaker for issuing commands to the speech recognition system and then identifying speech matching the voice sample as coming from the valid speaker. In further embodiments, the SRQI system also includes a directional sound capturing system (e.g., via directional filtering and/or directional tracking by microphones) for capturing sound, based on the voice sample of the valid speaker, from an area in which the valid speaker is located and providing the sound in the valid speaker's area to the speech recognition system and rejecting sound coming

from other areas so that such sound is not provided to or otherwise processed by the speech recognition system.

**[0038]** For example, in the context of a vehicle (or home, building, outside, or any other environment), according to various embodiments, a voice biometrics system allows for enrolling a driver of a car by providing a voice sample and then identifying the driver of the car as the valid speaker for issuing commands to the speech recognition system. By identifying the driver via voice biometrics, a directional sound capturing system may selectively provide sound (e.g., speech commands) from a direction or area of the driver to the speech recognition system for processing corresponding commands and prevent or mitigate undesired sound, such as speech or other sound from an occupant or music, from being provided to or otherwise processed by the speech recognition system. Accordingly, by implementing such a SRQI system to reduce at least some of the undesired sound being provided to the speech recognition system, the speech recognition system will be less likely to perform a task based on sound that was not intended by the speaker, thus leading to fewer false positives and misrecognitions by the speech recognition system.

**[0039]** FIGS. 1 and 2 illustrate an electronics device 102 having a speech recognition system 190 and a speech recognition quality improvement (SRQI) system 104 for improving quality of audio or speech input provided to the speech recognition system 190 for processing thereof. The electronics device 102 may be, but is not limited to, a mobile phone, "land line" phone, wired headset, wireless headset (e.g. Bluetooth®), hearing aid, audio/video recording device, head unit (e.g., in a vehicle), or other electronic device that utilizes transducers/microphones for receiving audio.

**[0040]** The speech recognition system 190 may be any system that accepts speech (also referred to as voice) input in order to control certain functions, or may otherwise benefit from separation of desired noises from background noises, such as (but not limited to) communication devices. The speech recognition system 190 may include human-machine interfaces in electronic or computational devices that incorporate capabilities such as voice recognition and detection, speech enhancement and separation, voice-activated control, and the like. For instance, the speech recognition 190 system may be for performing functions based on speech (e.g., verbal commands) 106 received from a user (speaker) of the electronics device 102 for instance via

microphones 110a, 110b, 110c, 110d. In particular embodiments, the speech recognition system 190 is an always-on speech recognition (ASR) system. It should be noted that the terms "user" and "speaker" may be used interchangeable unless otherwise noted. It should also be noted that the terms "speech" and "voice" may be used interchangeably unless otherwise noted. It should also be noted that the terms "sound" and "audio" may be used interchangeably unless otherwise noted.

**[0041]** In some embodiments, the microphones 110a-110d (which may be referred to collectively as microphones 110) are housed in the electronics device 102. In other embodiments, the microphones 110 are external (e.g., separate from) the electronics device 102. Each of the microphones 110 may have any suitable response, such as (but not limited to) omnidirectional, bidirectional, unidirectional (e.g., cardioid), and/or the like. The various types of microphones that may be used include, but are not limited to, piezoelectric microphones, dynamic microphones, electret microphones, and/or the like. In other embodiments, any suitable electro-acoustic transducer may be used in place of or in addition to the microphones 110. In some embodiments, four microphones 110 are implemented. In other embodiments, any suitable number of microphones 110 (e.g., less than four or more than four) may be implemented. The microphones 110 may be referred to as an array of microphones. In various embodiments, the SRQI system 104 may include and/or control operation of the microphones 110.

**[0042]** In various embodiments, the SRQI system 104 includes a voice biometrics system 140, which may include a voice biometrics engine 150, for authenticating a speaker (or user) for issuing commands or the like based on speech from the speaker to the speech recognition system 190. The authentication of a speaker may be based on a voice sample of the speaker.

**[0043]** In some embodiments, the voice biometrics system 140 (or components thereof such as the voice biometrics engine 150, the database 173, and/or the like) may be provided on a host device, such as a remote server or other electronic device. The electronics device 102 may be connectable to the host device via a network. The network may be a local area network (LAN), a wide area network (WAN), a telephone network such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, or a combination of networks. In other embodiments, the electronics device 102 may be connectable directly to the host device (e.g., USB, IR, Bluetooth, etc.).

**[0044]** The voice biometrics system 140 (e.g., via the voice biometrics engine 150) may be configured to provide an application service of authentication that includes, for example, enrollment and/or identification for enrolling a speaker and/or identifying the speaker based on speech from the speaker. Embodiments are provided wherein the speaker is a human being engaged in producing sounds in the form of utterances, which are recognized as speech, such as, for example, oral communication. According to various embodiments, the SRQI system 104 may implement voice biometrics to authenticate speakers and, once authenticated, may be a basis for discriminating desired sound (e.g., speech from the authenticated speaker) against undesired sound such that the desired sound is provided to and/or processed by the speech recognition system 190 while reducing undesired sound being fed to and/or processed by the speech recognition system 190.

**[0045]** In various embodiments, authentication includes an enrollment process and an identification process. The enrollment process may be performed to captures nuances (e.g., tones, patterns, accents, inflections, etc.) of a voice (speech) of a user enrolling in the system. A voice sample representing the nuances may be stored as a voice model 174 in a database 173 of voice models provided on a memory 172. In some embodiments, the voice model 174 (also referred to as a known acoustic model) may be a frequency distribution generated from training data (e.g., voice sample) obtained during the enrollment process. The voice model 174 obtained during the enrollment process may be used in authenticating (e.g., identifying) the user.

**[0046]** The identification process is the process of finding and attaching a speaker identity (e.g., as provided during the enrollment process) to a voice of an unknown speaker. In some embodiments, the voice biometrics system 140 compares captured sound (e.g., that includes speech 106 from an unknown speaker) with the voice models (voice prints) 174 stored on the database 173 of voice models. For instance, the voice biometrics system 140 may characterize the captured sound (e.g., through analysis of an acoustic spectrum of the sound) and compare the characterized sound to a known acoustic model (e.g., the voice model 174). In some embodiments, a time domain signal of the captured sound may be analyzed over a predetermined time window and a fast Fourier transform (FFT) may be performed to obtain a frequency distribution characteristic of the captured sound. The detected frequency distribution may be compared to the known acoustic model. If that comparison is favorable, the unknown

speaker is identified as a valid user having the identity of the matching voice model. Thus, by comparing the captured sound with the voice models 174, the voice biometrics system 140 may identify the captured sound as speech 106 from a valid user (speaker) or not (e.g., ambient noise 108). In some embodiments, if the comparison is not favorable (e.g., the unknown speaker's voice does not match any of the voice models 174), the enrollment process may be performed to enroll the unknown speaker.

**[0047]** Examples of voice model (or voice template) management are described in (but are not limited to) U.S. Patent Application Ser. No. 09/760,076, filed Jan. 12,2001, which is assigned to the assignee of the present disclosure and incorporated herein by reference in its entirety and U.S. Patent Application Ser. No. 09/703,191, filed Oct. 30, 2000, which is assigned to the assignee of the present disclosure and incorporated herein by reference in its entirety.

**[0048]** The voice biometrics engine 150 may be Large Vocabulary Continuous Speech Recognition (LVCSR), phonetic-based, text-dependent, text independent, or any other suitable type of voice biometrics engine. Enrollment and identification/verification for the voice biometrics engine 150 may be performed based on the type of the voice biometrics engine 150.

**[0049]** For embodiments in which the voice biometrics engine 150 is configured for LVCSR, the LVCSR may be based on a Hidden Markov Model (HMM) for training and recognition of spoken words. The LVCSR-based voice biometrics engine 150 does not split the spoken words into phonemes for training and recognition. Instead, the LVCSR-based voice biometrics engine 150 looks for entire words, as is, for training and recognition.

**[0050]** For embodiments in which the voice biometrics engine 150 is a phonetic-based voice biometrics engine, the words are split into phoneme units or sometimes even into sub-phoneme units. Next, the voice biometrics engine 150 is trained with those phonemes to create a voice model (e.g., 174) for a particular speaker. For embodiments in which the voice biometrics engine 150 is text dependent, text dependent speaker enrollment and identification is performed with a predefined utterance for both training (enrollment) and identification of the speakers.

**[0051]** FIG. 3 illustrates a flow chart of an illustrative method B500 for authenticating (e.g., enrolling and/or identifying) a speaker based on speech (voice sample) of the

speaker by the voice biometrics system 140 (e.g., the voice biometrics engine 150) of the SRQI system 104. With reference to FIGS. 1-3, at block B510, one or more of the microphones 110a-110d capture speech (voice) 106 from the speaker. At block B520, one or more signals 118a-118d corresponding to the captured speech 106 is transmitted to the SRQI system 104, for example to the voice biometrics system 140. In particular embodiments, one or more analog-to-digital converters 112a-112d may be used to convert one or more analog signals 111a-111d corresponding to the captured speech 106 into the one or more signals 118a-118d. In some embodiments, a signal is transmitted for less than all of the microphones 110a, 110b (e.g., only microphone 110a, two microphones in an electronics device having three microphones, etc.) to the voice biometrics system 140. In other embodiments, a signal is transmitted for every microphone (e.g., each of the microphones 110a, 110b, 110c, 110d) to the voice biometrics system 140.

**[0052]** At block B530, the voice biometrics system 140 is configured to perform an identification process to identify the speaker. As part of the identification process, at block B535, the speaker's speech 106 (as represented by the one or more signals 118a-118d, 118b) is compared with voice models 174 in the database 173 of voice models on the memory 172.

**[0053]** If the speaker's voice matches one of the voice models 174 (B535: Yes), then at block B550, the speaker is identified as the speaker having the matching voice model. If the speaker's voice does not match any of the voice models 174 in the database 173 (B535: No), then the voice biometrics engine 150 may perform a predetermined action. For instance, in some embodiments, at block B540, the voice biometrics system 140 may perform the enrollment process, for instance (but not limited to) as described in the disclosure. This may occur, for example, during an initial use of the SRQI system 104. As such, for example, at block B545, the voice biometrics system 140 may generate at least one voice model 174 for the speaker (e.g., based on a voice sample of the speaker) and store the voice model 174 in the database 172 as part of the enrollment process. After the enrollment process, the method B500 may be configured to proceed to any suitable block. For instance, after the enrollment process, the method B500 may return to block B510 such that speech subsequently captured from the speaker may be compared with the voice model 174 stored during the enrollment process (block B545) to identify the speaker. In addition or in the alternative, the method B500 may proceed

to block B550 (or any other block) after the enrollment process as the speaker will have been identified as the newly enrolled speaker.

**[0054]** Alternatively, if the speaker's voice does not match any of the voice models 174 in the database 173 and the enrollment process is not to be initiated (e.g., because enrollment has already been performed for a speaker and the captured speech does not match that of the enrolled speaker) then at block B542, the voice biometric system 140 may perform any other suitable predetermined action, such as (but not limited to), providing an indication of an error of a non-match, requesting an additional voice sample, disregarding speech from the user, and/or the like.

**[0055]** In some embodiments, the identification process (e.g., block B530) may be initiated before the enrollment process (e.g., block B540). For instance, if the identification process fails, for example, because the captured speech does not match any voice samples in the database 173, then the enrollment process is initiated. In particular embodiments, the captured speech (as used in the failed identification process) may be used as the voice model 174. In other particular embodiments, new and/or additional speech from the user is captured to generate the voice model 174. In other embodiments, the enrollment process (e.g., block B540) is performed before the identification process (e.g., block B530).

**[0056]** The SRQI system 104 (e.g., the voice biometrics system 140) may be configured to perform the authentication process at any suitable time and/or any suitable number of times. For instance, the voice biometrics engine 140 may authenticate the speaker's speech at predefined intervals (e.g., every five minutes, every five hundred words, etc.) or upon occurrence of predetermined events (e.g., upon usage of the electronics device 102 after a predetermined amount of time, powering on the electronics device 102, enabling the speech recognition system 190 or other related component, a presence of an undesired or unknown sound, etc.). In some embodiments, authentication may be performed at all times, such that, for example, each word or phrase (groups of words, such as a sentence or the like) captured by the microphones 110 is compared with the voice model 174 to ensure that such speech belongs to the authenticated (identified) speaker.

**[0057]** Electronic devices may be used in an environment that may include ambient noise 108 (also referred to as background noise or as undesired sound). As such, in

some instances, ambient noise 108 in addition to the speech 106 may be received by one or more of the microphones 110a-110d. Ambient noise 108 may be, but is not limited to, ancillary or background conversations (i.e., speech from other individuals), music or other media playing, and/or the like. In particular embodiments, ambient noise 108 may refer to undesired sound or any noise other than speech 106 from the speaker. Quality of sound being provided as input to a speech recognition system for processing may be affected by the presence of ambient noise 108.

**[0058]** Thus in various embodiments, the SRQI system 104 includes a directional sound capturing system 180 configured to discriminate, based on the voice biometrics of the identified speaker (e.g., as identified in the method B500), speech from the identified speaker against undesired sound (e.g., ambient noise, background conversations, etc.) such that the desired sound is provided to and/or processed by the speech recognition system 190 while reducing undesired sound being provided to and/or processed by the speech recognition system 190.

**[0059]** In some embodiments, the SRQI system 104 (e.g., the directional sound capturing system 180) may be configured to identify an area from which speech is coming from the speaker based on voice biometrics (e.g., voice model 174 as provided in block B540-545) of the speaker such that sound in the identified area is emphasized relative to sounds from other areas.

**[0060]** For instance, the directional sound capturing system 180 may capture or process sound coming from an area or direction of the speaker for providing to the speech recognition system 190 and not capture or process sound in the other areas or directions or otherwise mitigate such sound from being provided to and/or processed by the speech recognition system 190.

**[0061]** For example, in some embodiments, when the electronics device 102 is orientated such that desired sound (e.g., speech determined to match the voice sample 174 of the speaker) arrives from a direction in area A1, the directional sound capturing system 180 may be configured to use a (first) filter that is directional to the area A1 and tends to attenuate audio coming from other directions (e.g., area A2, area A3). Likewise, when the electronics device 102 is oriented such that desired speech arrives from a direction in area A2, the directional sound capturing system 180 may be configured to use a second filter that is directional to the area A2 and tends to attenuate

audio coming from other directions (e.g., area A1, area A3). When the electronics device 102 is oriented such that desired speech arrives from a direction in area A3, the directional sound capturing system 180 may be configured to use a third filter, which is directional to the area A3, and tends to attenuate audio coming from other directions (e.g., area A1, area A2).

**[0062]** For various embodiments, it is noted that the area boundaries shown in FIG. 1 are for visual illustrative purposes only. The illustrated boundaries do not purport to show the actual boundaries between areas associated with the various orientation states of the electronics device 102. In some embodiments, two or more of the filters may perform equally well for a sound source that is beyond some distance from the electronics device (such an orientation is also called a "far-field scenario"). This distance may depend largely on the distance between the microphones 110 of the electronics device 102. In some embodiments, two areas may overlap such that the two corresponding filters may be expected to perform equally well for a desired source located in the overlap region. In some embodiments, the areas may be predetermined in shape or size (e.g., area or volume) relative to the electronics device 102. For example, four areas may be provided to represent four quadrants around the electronics device 102 (e.g., sound coming from a direction toward bottom left of the device, bottom right of the device, top left of the device, and top right of the device). In other embodiments, the directional sound capturing system 180 may be configured to enter a single-channel mode such that only one microphone is active (e.g., microphone 110a, 110b, 110c, or 110d) or such that the microphones currently active are mixed down to a single channel, and possibly to suspend spatial processing operations. In some embodiments, the microphones 110 may be part of and/or controlled by the directional sound capturing system 180.

**[0063]** In some embodiments, the directional sound capturing system 180 is associated and discrete from the voice biometrics system 140. In other embodiments, the directional sound capturing system 180 includes or is integrated with the voice biometrics system 140 or portions thereof.

**[0064]** FIG. 4A illustrates a method B600 for improving quality of speech recognition performed by a speech recognition system (e.g., 190) according to various embodiments of the disclosure. FIG. 4B illustrates means-plus-function blocks B600' for improving

quality of speech recognition performed by a speech recognition system (e.g., 190) according to various embodiments of the disclosure. One or more of the method B600 and the means-plus-function blocks B600' may be implemented with or as part of the SRQI system 104 (e.g., FIGS. 1-3) and/or (but not limited to) any of the other embodiments of the disclosure. Although one or more of the method B600 and the means-plus-function blocks B600' may include features similar or used with the embodiments of FIGS. 1-3, it should be understood that one or more of the method B600 and the means-plus-function blocks B600' may also include (but is not limited to) some or all of the same features and operate in a manner similar to that shown and described in the embodiments of FIGS. 5-8. In addition, some or all of the features shown in FIGS. 1-3 and 5-8 may be combined in various ways and included in one or more of the embodiments relating to FIGS. 4A and 4B. Likewise, it should be understood that any of the features of the embodiments relating to FIGS. 4A and 4B may be combined or otherwise incorporated into any other embodiments relating to FIGS. 4A and 4B as well as (but not limited to) any other embodiment herein discussed.

**[0065]** With reference to FIGS. 1-4A, in various embodiments, the SRQI system 104 via implementation of the method B600 is configured to identify an area or a direction from which a desired sound, such as speech from the speaker, is coming based on the voice biometrics (e.g., voice model 174) of the speaker (e.g., as performed in the method B500) such that sound in the area or coming from the direction is emphasized relative to sounds in other areas or coming from other directions (in which the desired sound is not in or coming from). For instance, sound in the area or coming from the direction may be captured and/or provided to the speech recognition system 190 for processing whereas sounds from the other areas or directions are not captured and/or processed by the speech recognition system 190.

**[0066]** At blocks B610 and B620, a voice sample is obtained from the user and stored as a voice model 174 in a database 173 of voice models, for example as part of or similar to the authentication method B500, such as the enrollment process (e.g., blocks B540-B545), and/or the like.

**[0067]** Sound from one or more of the areas A1, A2, A3 may be detected or otherwise captured using the microphones 110 (e.g., microphone 110a and microphone 110b). The captured sound may include desired sound (e.g., speech 106 from the identified

speaker) and undesired sound (e.g., ambient noise 108). At block B630, the SRQI system 104 may identify an area from which sound matching the voice model for the speaker is coming. For instance, the SRQI system 104 may capture sound with one or more of the microphones 110 from a first area (e.g., A1) of the one of more areas A1-A3 in any suitable manner, such as using through filtering (e.g., as discussed in the disclosure and/or the like), selective use of the microphones (e.g., emphasizing sound captured by one or more of the microphones 110a-110d nearest and/or associated with the first area A1), and/or the like. Accordingly, at block B730, audio signals corresponding to the sound coming from the first area (e.g., A1) captured by the one or more microphones 110 is received by the SRQI 104.

**[0068]** Then the SRQI system 104 may determine whether the received audio signals corresponding to the sound coming from the first area A1 match the voice model 174 for the speaker. For instance, the SRQI system may compare the captured sound (as represented in the signals) with the voice model 174 to identify the sound (or source thereof), for example as part of or similar to the authentication method B500, such as the identification process (e.g., blocks B530-B550), and/or the like. By comparing the captured sound with the voice model 174, the captured sound may be determined to be speech 106 from the speaker or ambient noise 108.

**[0069]** At block B640, based upon the determination of the captured sound, the SRQI 104 may selectively take appropriate action depending upon whether the captured sound is identified as speech 106 from the speaker or ambient noise 108. For instance, if the captured speech coming the first area A1 is determined to be speech 106 from the valid speaker, the SRQI 104 may emphasize or amplify sound captured from the first area A1 (relative to sounds coming from the other areas A2, A3) and/or take other appropriate action such that captured sound coming from the first area A1 is provided to the speech recognition system 190. In some embodiments, amplifying sound or taking other appropriate action may include reducing noise disturbances associated with a source of sound. For example, the directional sound capturing system 180 may use a corresponding filter F10-1 through F10-n (e.g., refer to FIG. 5). Likewise, if the captured sound coming from the first area A1 is determined to be ambient noise 108, the SRQI system 104 may not provide sounds coming from the first area A1 to the speech recognition system 190. For instance, the SRQI system 104 may filter out or otherwise deemphasize sounds coming from the first area A1 or take other appropriate action.

**[0070]** After block B640, the method B600 may be repeated for a second area (e.g., A2) and any other areas, such as a third area (e.g., A3), to determine whether any of the sounds from those areas is speech from the speaker and, based on the determinations, to selectively provide audio signals corresponding to sounds from each of the areas to the speech recognition system 190.

**[0071]** Accordingly, for example, for each of the second area A2 and the third area A3, the captured sound coming from each area may be characterized and compared to the database 173 of voice models 174. If the captured sound coming from each area does not match the voice models, the SRQI system 104 may determine that the captured sound coming from the second area A2 and the third area A3 is ambient noise 108. For the first area A1, the captured sound coming from the area A1 may be characterized and compared to the database 173 of voice models 174. If the captured sound matches the voice model 174, the captured sound may be identified as the speech 106 from the speaker (matching the voice model). As a result, because the speech 106 is coming from the first area A1, the sound in the first area A1 continues to be captured and/or provided to the speech recognition system 190 whereas sounds from the second area A2 and the third area A3 are attenuated or otherwise not provided to the speech recognition system 190.

**[0072]** In some embodiments, the voice sample 174 is obtained with one or more of the microphones 110. In other embodiments, the voice sample is obtained using a different microphone, for example, of a home computer or the like.

**[0073]** The method B600 described in FIG. 4A above may be performed by various hardware and/or software component(s) and/or module(s) corresponding to the means-plus-function blocks B600' illustrated in FIG. 4B. In other words, blocks B610 through B640 illustrated in FIG. 4A correspond to means-plus-function blocks B610' through B640' illustrated in FIG. 4B.

**[0074]** In some embodiments, the SRQI system 104 may be configured to separate (filter) undesired sound (e.g., ambient noise) using any suitable signal processing techniques or other techniques from desired sound (e.g., speech matching the identified speech) based on the voice biometrics (e.g., voice model 174) of the speaker to produce a resulting signal that is provided to the speech recognition system 190 for processing. For instance, in particular embodiments, the SRQI system 104 may be configured to

separate undesired sound from an area in which sound from the authenticated speaker is coming.

**[0075]** FIG. 5 shows a block diagram of an illustrative directional sound capturing system 180' according to various embodiments of the disclosure. The directional sound capturing system 180' may be implemented with the SRQI system 104 (e.g., FIGS. 1-4B) and/or (but not limited to) any of the other embodiments of the disclosure. For instance, the directional sound capturing system 180' may include similar features as, employed as, and/or provided with an embodiment of the directional sound capturing system 180 (e.g., FIG. 3) and/or (but not limited to) any of the other embodiments of the disclosure. Although the directional sound capturing system 180' may include features similar or used with the embodiments of FIGS. 1-4B, it should be understood that the directional sound capturing system 180' may also include (but is not limited to) some or all of the same features and operate in a manner similar to that shown and described in the embodiments of FIGS. 6A-8. In addition, some or all of the features shown in FIGS. 1-4B and 6A-8 may be combined in various ways and included in the embodiments relating to FIG. 5. Likewise, it should be understood that any of the features of the embodiments relating to FIG. 5 may be combined or otherwise incorporated into any other embodiments relating to FIG. 5 as well as (but not limited to) any other embodiment herein discussed.

**[0076]** In various embodiments, the sound capturing system 180' is configured to separate undesired sound (e.g., ambient noise 108) from desired sound (e.g., speech matching the identified speech) based on the voice biometrics (e.g., voice model 174) of the speaker to produce a resulting signal 192 that is fed to the speech recognition system 190 for processing.

**[0077]** With reference to FIGS. 1-5, in some embodiments, the directional sound capturing system 180' includes a filter bank 182 that is configured to receive an M-channel input signal S10, where M is an integer greater than one and each of the M channels is based on an output signal (e.g., 118a-118d) of a corresponding one of M microphones (e.g., the microphones 110a-110d). In various embodiments, the microphone signals may be (but is not limited to any one or combination of) sampled, pre-processed (e.g., filtered for echo cancellation, noise reduction, spectrum shaping, etc.), pre-separated (e.g., by another spatial separation filter or adaptive filter as

described in the disclosure). For acoustic applications, such as speech, sampling rates may range for example from 8 kHz to 16 kHz.

**[0078]** The filter bank 182 may include n spatial separation filters F10-1 to F10-n (where n is an integer greater than one), each of which is configured to filter the M-channel input signal S10 to produce a corresponding spatially processed M-channel signal. Each of the spatial separation filters F10-1 to F10-n is configured to separate one or more directional desired sound components (speaker's speech 106) of the M-channel input signal from one or more other components of the signal (e.g., ambient noise 108).

**[0079]** For instance, the filter F10-1 produces an M-channel signal that includes filtered channels S2011 to S201n, the filter F10-2 produces an M-channel signal that includes filtered channels S2012 to S202n, and so on. Each of the filters F10-1 to F10-n may be characterized by one or more matrices of coefficient values, which may be calculated using a blind source separation (BSS), beamforming, or combined BSS/beamforming method (e.g., an ICA, or IVA method or a variation thereof) and/or may be trained (e.g., based on the voice model 174). In some cases, a matrix of coefficient values may be only a vector (i.e., a one-dimensional matrix) of coefficient values. In various embodiments, the one or more of the filters F10-1 may be configured based on the speaker's voice model 174 to separate speech 106 matching the speaker's voice 174 from ambient noise 108.

**[0080]** The directional sound capturing system 180' may also includes a switching mechanism 184 that is configured to receive the M-channel filtered signal from each filter F10-1 to F10-n to determine which of these filters currently best separates at least one desired component of the input signal S10 from one or more other components, and to produce an M-channel output signal S40 accordingly.

**[0081]** The M-channel output signal S40 (e.g., 192 in FIG. 2), which is provided by the filter determined by the switching mechanism 184 to best separate the at least one desired component (e.g., speech) from one or more other components (e.g., ambient noise) may be transmitted to the speech recognition system 190. Thus, for example, if speech from the identified speaker is coming in a direction from the area A3 (toward the microphone 110b), the switching mechanism 184 may determine that the filter F10-2 best separates the speech from ambient noise. Accordingly, the output channel signal S40-2 may be transmitted to the speech recognition system 190. As such, by selecting

the filter F10-2, sound, including the speech, coming from the direction from the area A3 is fed to the speech recognition system 190 whereas sound coming from directions of the other areas A1, A2 is at least partially attenuated.

**[0082]** In some embodiments, the SRQI system 104 (e.g., the directional sound capturing system 180, 180') may implement beamforming. Beamforming techniques use the time difference between channels that results from the spatial diversity of the microphones to enhance a component of the signal that arrives from a particular direction. More particularly, it is likely that one of the microphones will be oriented more directly at the desired source (e.g., the user's mouth), whereas the other microphone may generate a signal from this source that is relatively attenuated. These beamforming techniques are methods for spatial filtering that steer a beam towards a sound source, putting a null at the other directions. Beamforming techniques make no assumption on the sound source but assume that the geometry between source and sensors, or the sound signal itself, is known for the purpose of dereverberating the signal or localizing the sound source. For instance, in some embodiments, one or more of the filters of the filter bank 182 may be configured according to a data-dependent or data-independent beamformer design (e.g., a superdirective beamformer, leastsquares beamformer, or statistically optimal beamformer design). In the case of a data-independent beamformer design, the beam pattern may be shaped to cover a desired spatial area (e.g., by tuning the noise correlation matrix).

**[0083]** FIGS. 6A and 6B show block diagrams of illustrative directional sound capturing system 200a, 200b according to various embodiments of the disclosure. One or more of the directional sound capturing systems 200a, 200b may be implemented with the SRQI system 104 (e.g., FIGS. 1-5) and/or (but not limited to) any of the other embodiments of the disclosure. For instance, one or more of the directional sound capturing systems 200a, 200b may include similar features as, employed as, and/or provided with an embodiment of the directional sound capturing systems 180, 180' (e.g., FIGS. 2 and 5) and/or (but not limited to) any of the other embodiments of the disclosure. Although one or more of the directional sound capturing systems 200a, 200b may include features similar or used with the embodiments of FIGS. 1-5, it should be understood that one or more of the directional sound capturing systems 200a, 200b may also include (but is not limited to) some or all of the same features and operate in a manner similar to that shown and described in the embodiments of FIGS. 7A-8. In

addition, some or all of the features shown in FIGS. 1-5 and 7A-8 may be combined in various ways and included in one or more of the embodiments relating to FIGS. 6A and 6B. Likewise, it should be understood that any of the features of the embodiments relating to FIGS. 6A and 6B may be combined or otherwise incorporated into any other embodiments relating to FIGS. 6A and 6B as well as (but not limited to) any other embodiment herein discussed.

**[0084]** With reference to FIGS. 1-6A, in various embodiments, the directional sound capturing system 200a (in addition to or in place of the sound capturing system 180) is configured to separate undesired sound (e.g., ambient noise 108) from desired sound (e.g., speech matching the identified speech) based on the voice biometrics (e.g., voice model 174) of the speaker to produce a resulting signal 192 that is fed to the speech recognition system 190 for processing.

**[0085]** In some embodiments, the directional sound capturing system 200a may include a beamformer 214 and/or a noise reference refiner 220a. The directional sound capturing system 200a may be configured to receive digital audio signals 212a, 212b. The digital audio signals 212a, 212b may or may not have matching or similar energy levels. The digital audio signals 212a, 212b may be signals from audio sources (e.g., the signals 118a, 118b from the microphones 110a, 110b in the electronics device 102). The digital audio signals 212a, 212b may have matching or similar signal characteristics. For example, both signals 212a, 212b may include a desired audio signal (e.g., speech 106). The digital audio signals 212a, 212b may also include ambient noise 108.

**[0086]** The digital audio signals 212a, 212b may be received by the beamformer 214. One of the digital audio signals 212a may also be routed to the noise reference refiner 220a. The beamformer 214 may generate a desired audio reference signal 216 (e.g., a voice/speech reference signal). The beamformer 214 may generate a noise reference signal 218. The noise reference signal 218 may contain residual desired audio. The noise reference refiner 220a may reduce or effectively eliminate the residual desired audio from the noise reference signal 218 in order to generate a refined noise reference signal 222a. The noise reference refiner 220a may utilize one of the digital audio signals 212a, 212b to generate a refined noise reference signal 222a. The desired audio reference signal 216 and the refined noise reference signal 222a may be utilized to

provide the resulting signal 192, which has improved desired audio output. For example, the refined noise reference signal 222a may be filtered and subtracted from the desired audio reference signal 216 in order to reduce noise in the desired audio. The desired audio reference signal 216 and the refined noise reference signal 222a and/or the resulting signal 192 (e.g., a signal resulting from subtracting the refined noise reference signal 222a from the desired audio reference signal 216) may be transmitted to the speech recognition system 190. In some embodiments, the refined noise reference signal 222a and the desired audio reference signal 216 may also be further processed to reduce noise in the desired audio.

**[0087]** With reference to FIGS. 1-6B, in various embodiments, the directional sound capturing system 200b (in addition to or in place of the sound capturing system 180 and/or 200a) is configured to separate undesired sound (e.g., ambient noise 108) from desired sound (e.g., speech matching the identified speech) based on the voice biometrics (e.g., voice model 174) of the speaker to produce a resulting signal 192 that is fed to the speech recognition system 190 for processing.

**[0088]** The directional sound capturing system 200b may include digital audio signals 212a, 212b, a beamformer 214, a desired audio reference signal 216, a noise reference signal 218, a noise reference refiner 220b, and a refined noise reference signal 222b. As the noise reference signal 218 may include residual desired audio, the noise reference refiner 220b may reduce or effectively eliminate residual desired audio from the noise reference signal 218. The noise reference refiner 220b may utilize both digital audio signals 212a, 212b in addition to the noise reference signal 218 in order to generate a refined noise reference signal 222b. The refined noise reference signal 222b and the desired audio reference signal 216 may be utilized to provide the resulting signal 192, which has improved desired audio output. The desired audio reference signal 216 and the refined noise reference signal 222b and/or the resulting signal 192 (e.g., a signal resulting from subtracting the refined noise reference signal 222b from the desired audio reference signal 216) may be transmitted to the speech recognition system 190. In some embodiments, the refined noise reference signal 222b and the desired audio reference signal 216 may also be further processed to reduce noise in the desired audio. Further illustrative beamformer and/or noise reference refiner examples are disclosed in (but are not limited to) U.S. Application Nos. 12/334,246 (filed on December 12, 2008);

12/323,200 (filed on November 25, 2008), all of which are assigned to Qualcomm, Inc. and are incorporated by reference in their entirety.

**[0089]** It should be noted that the directional sound capturing systems, components, and/or configurations thereof discussed with respect to FIGS. 5-6B as well as any other embodiments discussed in the disclosure are merely illustrative. Further non-limiting examples of systems and methods for separating an undesired sound from a desired sound or otherwise emphasizing a desired sound (e.g., sound, such as speech, coming from a desired direction) and attenuating undesired sound (e.g., ambient sound) are disclosed in (but are not limited to) U.S. Application No. 12/334,246 (filed on December 12, 2008); U.S. Application No. 12/323,200 (filed on November 25, 2008); U.S. Application No. 12/473,492 (filed on May 28, 2009); U.S. Application No. 12/277,283 (filed on November 24, 2008); U.S. Application No. 12/796,566 (filed on June 8, 2010); U.S. Patent No. 7,464,029 (filed on July 22, 2005), all of which are assigned to Qualcomm, Inc. and are incorporated by reference in their entirety. Such systems and methods, for instance, may be used in placed of or in addition to the directional sound capturing systems 180, 200a, 20b (e.g., FIGS. 1-6B).

**[0090]** FIGS. 7A illustrates a method B700 for improving quality of speech recognition performed by a speech recognition system (e.g., 190) according to various embodiments of the disclosure. FIG. 7B illustrates means-plus-function blocks B700' for improving quality of speech recognition performed by a speech recognition system (e.g., 190) according to various embodiments of the disclosure. One or more of the method B700 and the means-plus-function blocks B700' may be implemented with or as part of the SRQI system 104 (e.g., FIGS. 1-6B) and/or (but not limited to) any of the other embodiments of the disclosure. For instance, at least a portion of the method B700 and/or the means-plus-function blocks B700' may be implemented by the directional sound capturing system 180, 180', 200a, 200b (e.g., FIGS. 2 and 5-6B) and/or (but not limited to) any of the other embodiments of the disclosure. Although one or more of the method B700 and the means-plus-function blocks B700' may include features similar or used with the embodiments of FIGS. 1-6B, it should be understood that one or more of the method B700 and the means-plus-function blocks B700' may also include (but is not limited to) some or all of the same features and operate in a manner similar to that shown and described in the embodiments of FIG. 8. In addition, some or all of the features shown in FIGS. 1-6B and 8 may be combined in

various ways and included in one or more of the embodiments relating to FIGS. 7A and 7B. Likewise, it should be understood that any of the features of the embodiments relating to FIGS. 7A and 7B may be combined or otherwise incorporated into any other embodiments relating to FIGS. 7A and 7B as well as (but not limited to) any other embodiment herein discussed.

**[0091]** With reference to FIGS. 1-7A, at blocks B710 and B720, a voice sample is obtained from the user and stored as a voice model 174 in a database 173 of voice models, for example as part of in a manner similar to the authentication method B500, such as the enrollment process (e.g., blocks B540-B545), and/or the like.

**[0092]** The SRQI system 104 (e.g., the microphones 110a-110d) may capture sound. The captured sound may include desired sound (e.g., speech 106 from a speaker matching the voice model provided in blocks B710 and B720) and undesired sound (e.g., ambient noise 108). Accordingly, at block B730, one or more audio signals corresponding to the sound captured by at least two of the microphones 110 is received by the SRQI system 104.

**[0093]** At block B740, the SRQI system 104 (e.g., the directional sound capturing system 180) may separate (e.g., filters out, attenuates, etc.) the ambient noise 108 from the speech 106 in the one or more audio signals based on the voice model 174 stored in the database 173 using any suitable signal processing techniques or other techniques to obtain a resulting speech signal 192. That is, based on the voice model 174 of the speaker, undesired sound (e.g., ambient noise) may be separated using any suitable signal processing techniques or other techniques from desired sound, such as speech matching the voice model 174. For instance, the voice model 174 may provide a reference for filtering undesired audio (e.g., ambient noise 108) from speech 106 (desired audio) matching the voice model 174 to produce the resulting speech signal 192 that is fed to the speech recognition system 190 for processing. For example, sound from other areas than that of the area in which the identified speaker is located may be subtracted from sounds of the speaker's area. At block B750, the SRQI system 104 (e.g., directional sound capturing system 180) provides the resulting speech signal 192 to the speech recognition system 190.

**[0094]** The method B700 described in FIG. 7A above may be performed by various hardware and/or software component(s) and/or module(s) corresponding to the means-

plus-function blocks B700' illustrated in FIG. 7B. In other words, blocks B710 through B740 illustrated in FIG. 7A correspond to means-plus-function blocks B710' through B740' illustrated in FIG. 7B.

**[0095]** FIG. 8 illustrates certain components that may be included within an electronics device (e.g., 102 in FIGS. 1-4B) according to various embodiments of the disclosure. The electronics device 102 may include a SRQI system 104 and/or be implemented with any of the features of the embodiments relating to FIGS. 1-7B.

**[0096]** With reference to FIGS. 1-8, the electronics device 102 includes a processor 170. The processor 170 may be a general purpose single or multi-chip microprocessor (e.g., an ARM), a special purpose microprocessor (e.g., a digital signal processor (DSP)), a microcontroller, a programmable gate array, etc. The processor 170 may be referred to as a central processing unit (CPU). The processor may be a single processor 170 or a combination of processors (e.g., an ARM and DSP) depending on the embodiment.

**[0097]** The electronics device 102 may include a memory 172. The memory 172 may be any electronic component capable of storing electronic information. The memory 172 may be embodied as random access memory (RAM), read only memory (ROM), magnetic disk storage media, optical storage media, flash memory devices in RAM, on-board memory included with the processor, EPROM memory, EEPROM memory, registers, and so forth, including combinations thereof.

**[0098]** Data 174 and instructions 176 may be stored in the memory 172. The data 174 and/or the instructions 176 may relate to the voice biometrics portion, the directional sound capturing system 180, and/or the speech recognition system 190 of the electronics device. In some embodiments, the SRQI system 104, portions thereof, and/or methods of implementation may be provided on the memory 172. For example, the data 174 may include the voice models 174 and/or the instructions 176 may include instructions for executing the voice biometrics engine 150 and/or the directional sound capturing system 180 or the like. The instructions 176 may be executable by the processor 170 to implement the methods disclosed in the disclosure. Executing the instructions 176 may involve the use of some of the data 174 that is stored in the memory 172. The electronics device 102 may also multiple microphones 110a, 110b, through 110n. The

microphones 110 may receive audio signals that include speech 106 and/or ambient noise 108.

**[0099]** In some embodiments, the electronics device 102 may also include a transmitter 162 and a receiver 164 to allow wireless transmission and reception of signals between the communication device 102 and a remote location. The transmitter 162 and the receiver 164 may be collectively referred to as a transceiver 160. An antenna 166 may be electrically coupled to the transceiver 160. The communication device 102 may also include (not shown) multiple transmitters, multiple receivers, multiple transceivers and/or multiple antenna. The various components of the communication device 102 may be coupled by one or more buses 168, which may include a power bus, a control signal bus, a status signal bus, a data bus, and/or the like.

**[0100]** In some embodiments, the SRQI system 104 may omit the directional sound capturing system 180 such that the speech 106 (e.g., speech used to identify or enroll the speaker and/or subsequent speech from the speaker) from a speaker identified by the voice biometrics system 140 may be input into the speech recognition system 190 for performing functions based on the input without.

**[0101]** Various embodiments as described in the disclosure may be incorporated into an electronic device that accepts speech input in order to control certain functions, or may otherwise benefit from separation of desired noises from background noises, such as communication devices. Many applications may benefit from enhancing or separating clear desired sound from background sounds originating from multiple directions. Such applications may include human-machine interfaces in electronic or computational devices that incorporate capabilities such as voice recognition and detection, speech enhancement and separation, voice-activated control, and the like.

**[0102]** It is understood that the specific order or hierarchy of steps in the processes disclosed is an example of illustrative approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the processes may be rearranged while remaining within the scope of the present disclosure. The accompanying method claims present elements of the various steps in a sample order, and are not meant to be limited to the specific order or hierarchy presented.

**[0103]** Those of skill in the art would understand that information and signals may be represented using any of a variety of different technologies and techniques. For

example, data, instructions, commands, information, signals, bits, symbols, and chips that may be referenced throughout the above description may be represented by voltages, currents, electromagnetic waves, magnetic fields or particles, optical fields or particles, or any combination thereof.

**[0104]** Those of skill would further appreciate that the various illustrative logical blocks, modules, circuits, and algorithm steps described in connection with the implementations disclosed herein may be implemented as electronic hardware, computer software embodied on a tangible medium, or combinations of both. To clearly illustrate this interchangeability of hardware and software, various illustrative components, blocks, modules, circuits, and steps have been described above generally in terms of their functionality. Whether such functionality is implemented as hardware or software embodied on a tangible medium depends upon the particular application and design constraints imposed on the overall system. Skilled artisans may implement the described functionality in varying ways for each particular application, but such implementation decisions should not be interpreted as causing a departure from the scope of the present disclosure.

**[0105]** The various illustrative logical blocks, modules, and circuits described in connection with the implementations disclosed herein may be implemented or performed with a general purpose processor, a digital signal processor (DSP), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A general-purpose processor may be a microprocessor, but in the alternative, the processor may be any conventional processor, controller, microcontroller, or state machine. A processor may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration.

**[0106]** The steps of a method or algorithm described in connection with the implementations disclosed herein may be embodied directly in hardware, in a software module executed by a processor, or in a combination of the two. A software module may reside in RAM memory, flash memory, ROM memory, EPROM memory,

EEPROM memory, registers, hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art. An illustrative storage medium is coupled to the processor such the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an ASIC. The ASIC may reside in a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a user terminal.

**[0107]** In one or more illustrative implementations, the functions described may be implemented in hardware, software or firmware embodied on a tangible medium, or any combination thereof. If implemented in software, the functions may be stored on or transmitted over as one or more instructions or code on a computer-readable medium. Computer-readable media includes both computer storage media and communication media including any medium that facilitates transfer of a computer program from one place to another. A storage media may be any available media that can be accessed by a computer. By way of example, and not limitation, such computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium that can be used to carry or store desired program code in the form of instructions or data structures and that can be accessed by a computer. In addition, any connection is properly termed a computer-readable medium. For example, if the software is transmitted from a website, server, or other remote source using a coaxial cable, fiber optic cable, twisted pair, digital subscriber line (DSL), or wireless technologies such as infrared, radio, and microwave, then the coaxial cable, fiber optic cable, twisted pair, DSL, or wireless technologies such as infrared, radio, and microwave are included in the definition of medium. Disk and disc, as used herein, includes compact disc (CD), laser disc, optical disc, digital versatile disc (DVD), floppy disk, and Blu-Ray disc where disks usually reproduce data magnetically, while discs reproduce data optically with lasers. Combinations of the above should also be included within the scope of computer-readable media.

**[0108]** The previous description of the disclosed implementations is provided to enable any person skilled in the art to make or use the present disclosure. Various modifications to these implementations will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other implementations

without departing from the spirit or scope of the disclosure. Thus, the present disclosure is not intended to be limited to the implementations shown herein but is to be accorded the widest scope consistent with the principles and novel features disclosed herein.

**WHAT IS CLAIMED IS:**

1. A method for improving speech recognition by a speech recognition system, comprising:

obtaining a voice sample from a speaker;

storing the voice sample of the speaker as a voice model in a voice model database;

identifying an area from which sound matching the voice model for the speaker is coming; and

providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

2. The method of claim 1, further comprising:

controlling microphones to receive sound from the identified area.

3. The method of claim 2, further comprising:

controlling the microphones to not receive sound from areas other than the identified area.

4. The method of claim 1, further comprising:

filtering sound received from areas other than the identified area.

5. The method of claim 1, further comprising:

separating ambient noise from speech in the one or more audio signals based on the voice model stored in the voice model database to obtain a resulting speech signal; and

providing the resulting speech signal to the speech recognition system for processing.

6. The method of claim 1,

wherein the identifying comprises:

receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas; and

determining whether the received audio signals corresponding to the sound coming from the first area match the voice model for the speaker;

wherein the providing comprises:

selectively providing audio signals corresponding to sound from the first area to the speech recognition system for processing thereof based on the determination.

7. The method of claim 6,

wherein if the one or more received audio signals corresponding to the sound coming from the first area match the voice model for the speaker, one or more audio signals corresponding to sound from the first area are provided to the speech recognition system for processing thereof; and

wherein if the one or more received audio signals corresponding to the sound coming from the first area do not match the voice model for the speaker, one or more audio signals corresponding to sound from the first area are not provided to the speech recognition system for processing thereof.

8. The method of claim 6, wherein one or more audio signals corresponding to sound from the first area received after the determination are selectively provided to the speech recognition system for processing thereof based on the determination.

9. The method of claim 6, wherein the one or more microphones comprises at least two microphones.

10. The method of claim 6, further comprising:

receiving, from the one or more microphones, one or more audio signals corresponding to sound coming from a second area of the plurality of areas;

determining whether the one or more received audio signals corresponding to the sound coming from the second area match the voice model for the speaker; and

selectively providing audio signals corresponding to sound from the second area to the speech recognition system for processing thereof based on the determination.

11. The method of claim 10,

wherein if the one or more audio signals corresponding to the sound coming from the second area match the voice model for the speaker, one or more audio signals corresponding to sound from the second area are provided to the speech recognition system for processing thereof; and

wherein if the one or more audio signals corresponding to the sound coming from the second area do not match the voice model for the speaker, one or more audio signals corresponding to sound from the second area are not provided to the speech recognition system for processing thereof.

12. The method of claim 6, wherein the determining comprises:

characterizing the sound coming from the first area; and

comparing the characterized sound to the voice model for the speaker.

13. The method of claim 12,

wherein if the characterized sound compares with the voice model for the speaker, the sound coming from the first area is determined to be speech from the speaker and one or more audio signals corresponding to sound coming from the first area is provided to the speech recognition system; and

wherein if the characterized sound does not compare with the voice model for the speaker, the sound coming from the first area is determined not to be speech from the speaker and one or more audio signals corresponding to sound coming from the first area is not provided to the speech recognition system.

14. The method of claim 12, wherein the sound coming from the first area is characterized through analysis of an acoustic spectrum of the sound.

15. The method of claim 12,

wherein the one or more microphones comprise a first microphone associated with a first area of the plurality of areas and a second microphone associated a second area of the plurality of areas;

wherein if the one or more received audio signals corresponding to the sound coming from the first area matches the voice model for the speaker, one or more audio signals corresponding to sound from the first area are received from the first microphone for providing to the speech recognition system; and

wherein if the one or more received audio signals corresponding to the sound coming from the first area do not matches the voice model for the speaker, one or more audio signals corresponding to sound from the first area are not received from the first microphone for providing to the speech recognition system.

16. The method of claim 15,

wherein if the one or more received audio signals corresponding to the sound coming from the first area matches the voice model for the speaker and the one or more received audio signals corresponding to the sound coming from the second area does not match the voice model for the speaker, one or more audio signals corresponding to sound from the first area are received from the first microphone for providing to the speech recognition system; and

wherein if the one or more received audio signals corresponding to the sound coming from the first area does not match the voice model for the speaker and the one or more received audio signals corresponding to the sound coming from the second area matches the voice model for the speaker, one or more audio signals corresponding to sound from the second area are received from the second microphone.

17. The method of claim 1,

wherein the audio signals provided to the speech recognition system correspond to speech commands for the speech recognition system; and

wherein the speech recognition system issues instructions based on the speech.

18. An apparatus for improving speech recognition by a speech recognition system, the apparatus comprising:

means for obtaining a voice sample from a speaker;

means for storing the voice sample of the speaker as a voice model in a voice model database;

means for identifying an area from which sound matching the voice model for the speaker is coming; and

means for providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

19. An apparatus for improving speech recognition by a speech recognition system, the apparatus comprising:

a processor configured for:

obtaining a voice sample from a speaker;

storing the voice sample of the speaker as a voice model in a voice model database;

receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas;

identifying an area from which sound matching the voice model for the speaker is coming; and

providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

20. A computer program product for improving speech recognition by a speech recognition system, the computer program product comprising:

a computer-readable storage medium comprising code for:

obtaining a voice sample from a speaker;

storing the voice sample of the speaker as a voice model in a voice model database;

receiving, from one or more microphones, one or more audio signals corresponding to sound coming from a first area of a plurality of areas;

identifying an area from which sound matching the voice model for the speaker is coming; and

providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing.

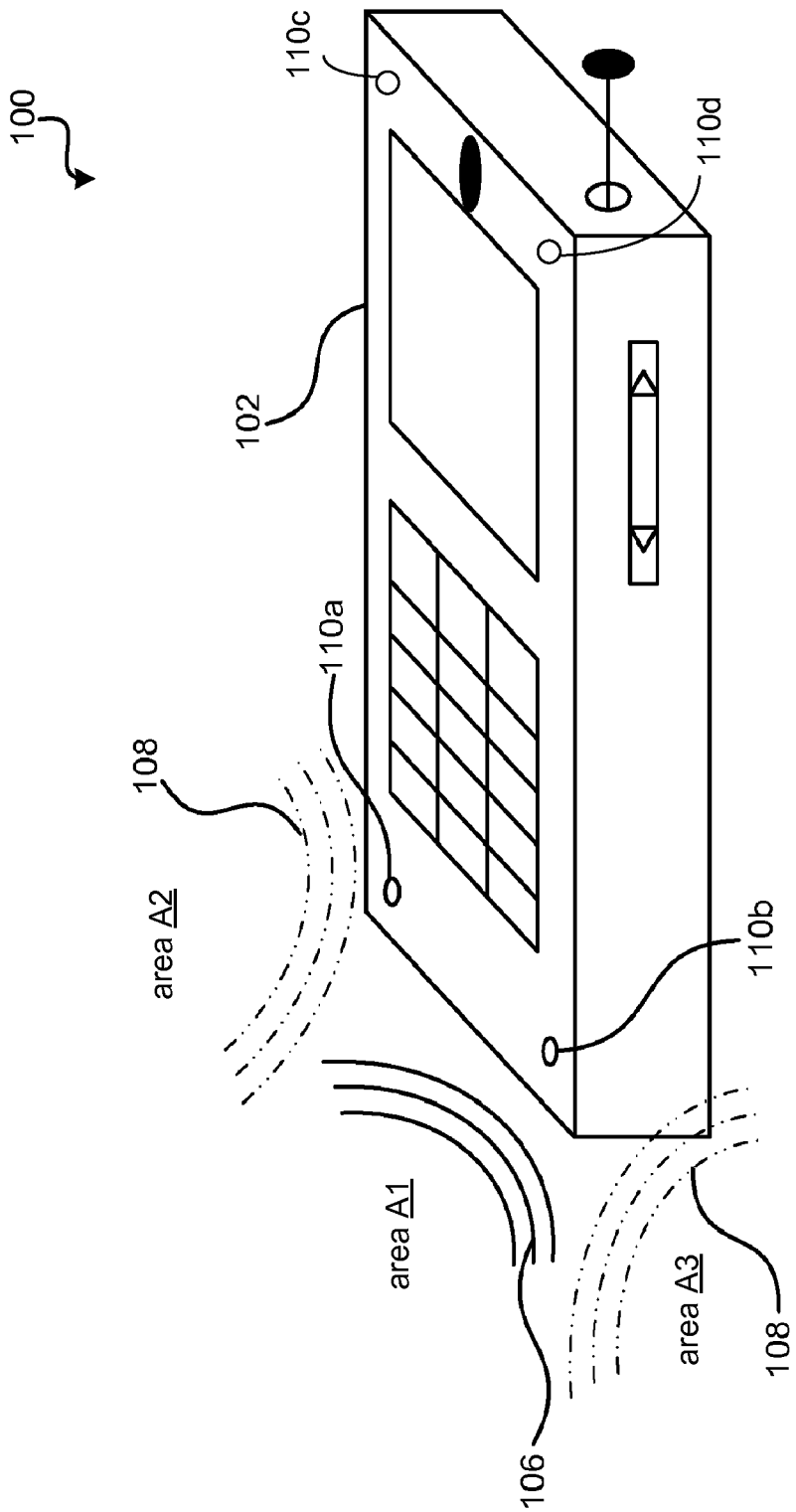



FIG. 1

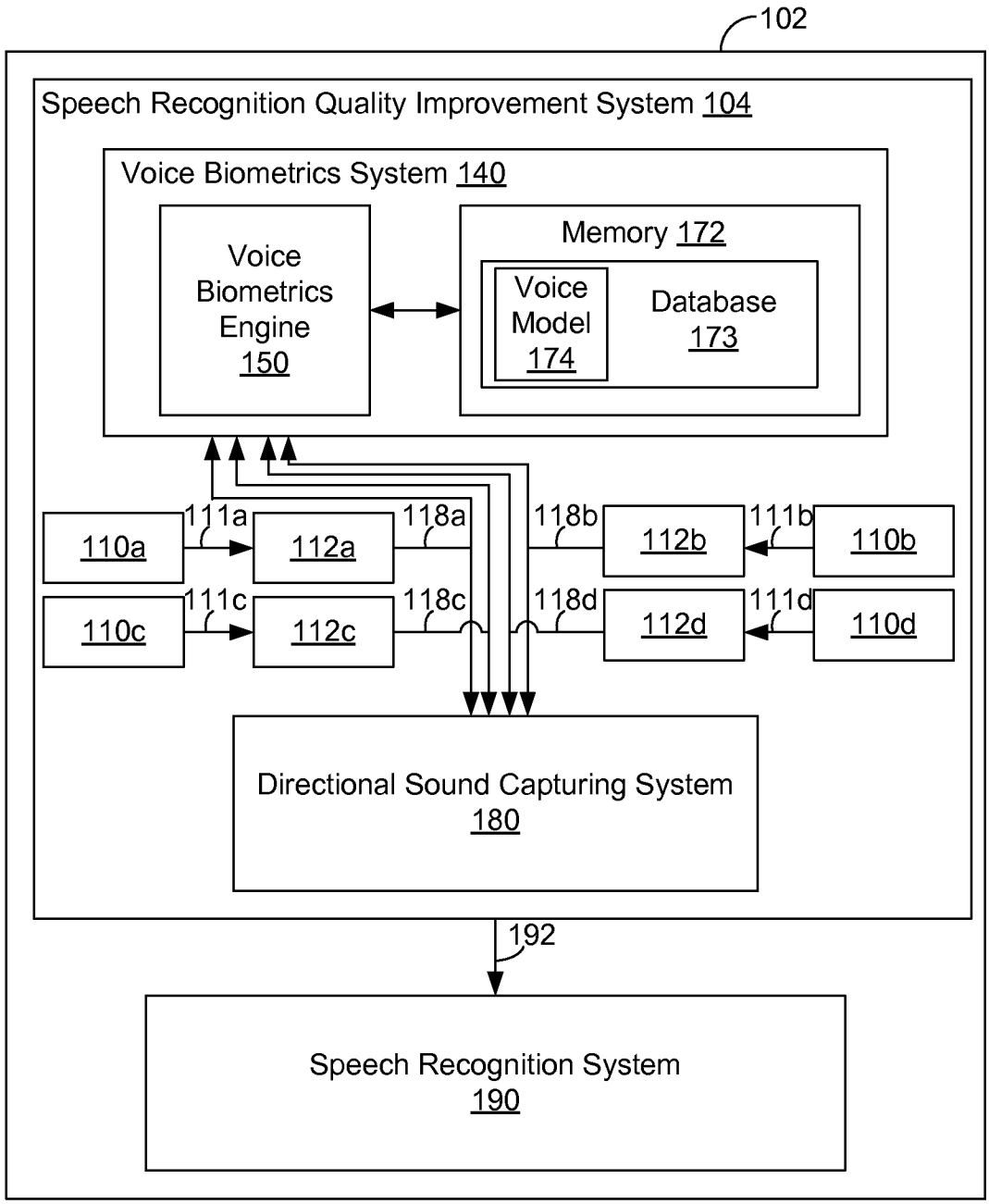

102
Speech Recognition Quality Improvement System 104
Voice Biometrics System 140
Voice Biometrics Engine 150
Memory 172
Voice Model 174
Database 173
110a
111a
112a
118a
118b
112b
111b
110b
110c
111c
112c
118c
118d
112d
111d
110d
Directional Sound Capturing System 180
192
Speech Recognition System 190

FIG. 2

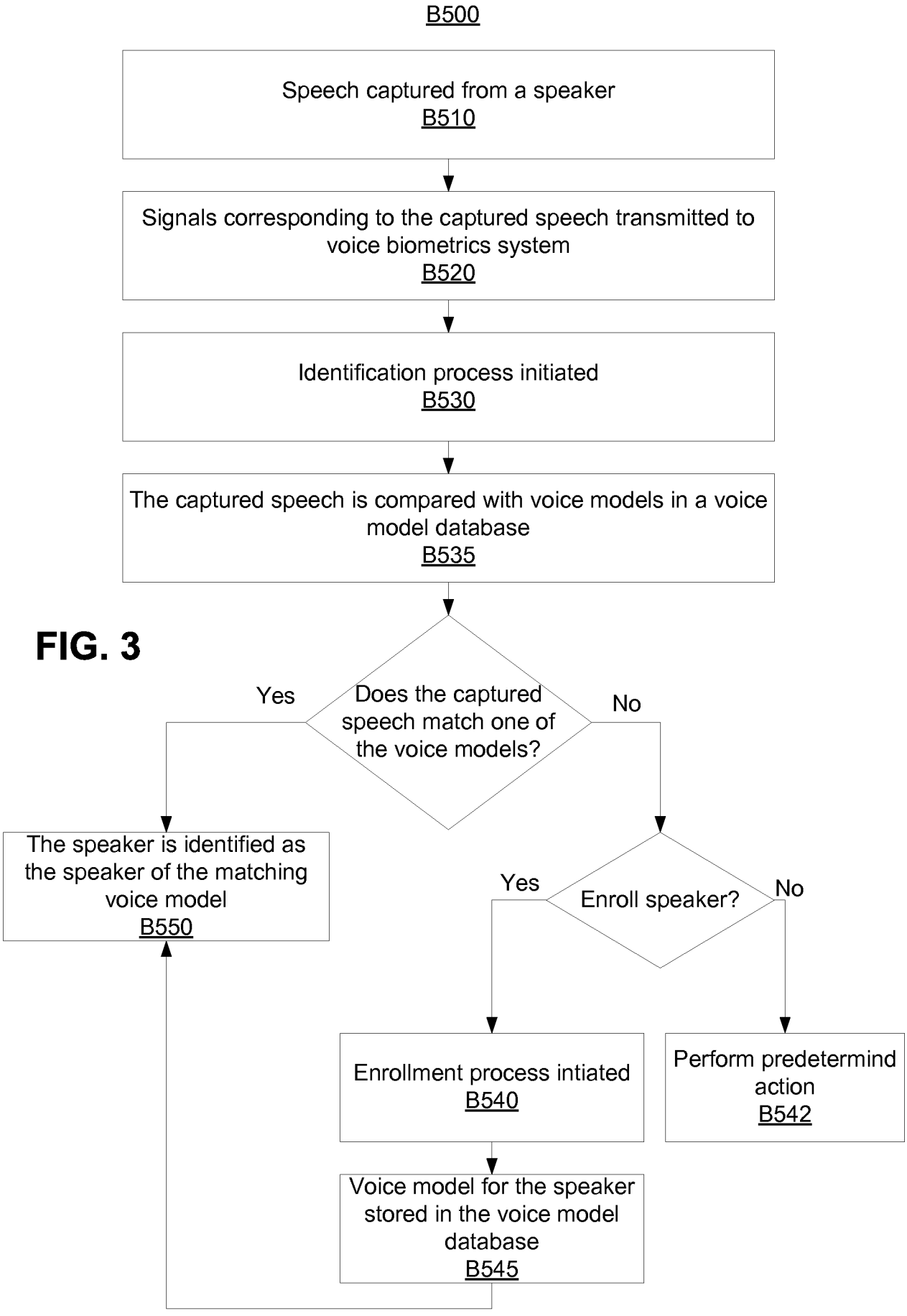

B500
Speech captured from a speaker
B510
Signals corresponding to the captured speech transmitted to voice biometrics system
B520
Identification process initiated
B530
The captured speech is compared with voice models in a voice model database
B535
Does the captured speech match one of the voice models?
Yes
No
The speaker is identified as the speaker of the matching voice model
B550
Enroll speaker?
Yes
No
Enrollment process intiated
B540
Perform predetermind action
B542
Voice model for the speaker stored in the voice model database
B545

FIG. 3

FIG. 4A

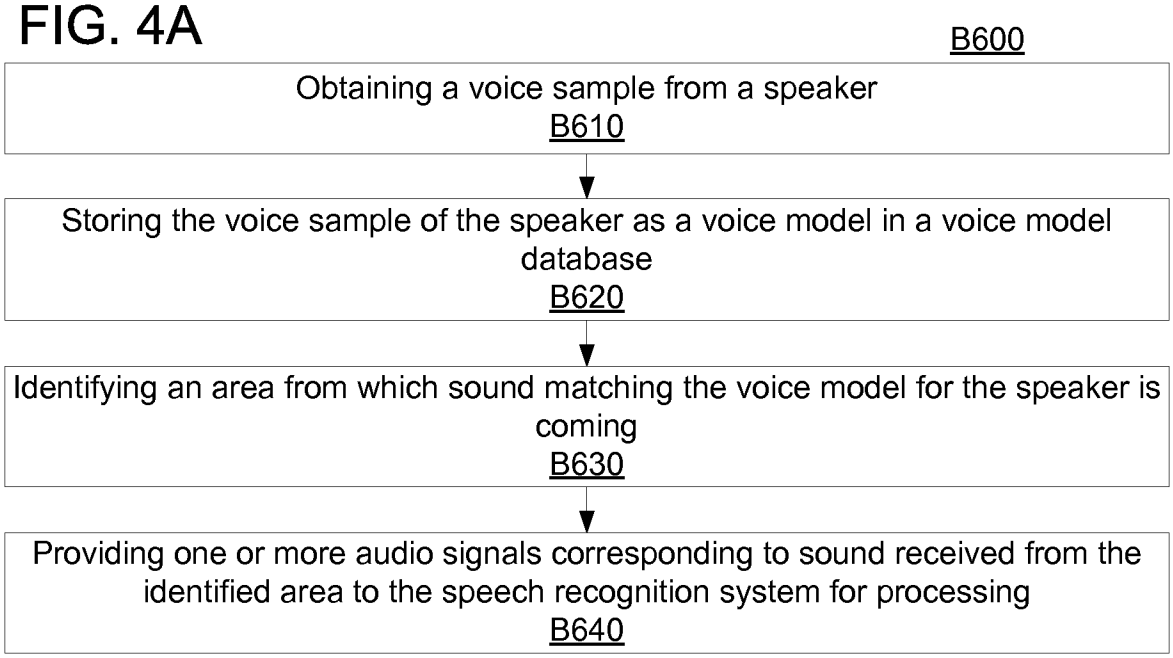
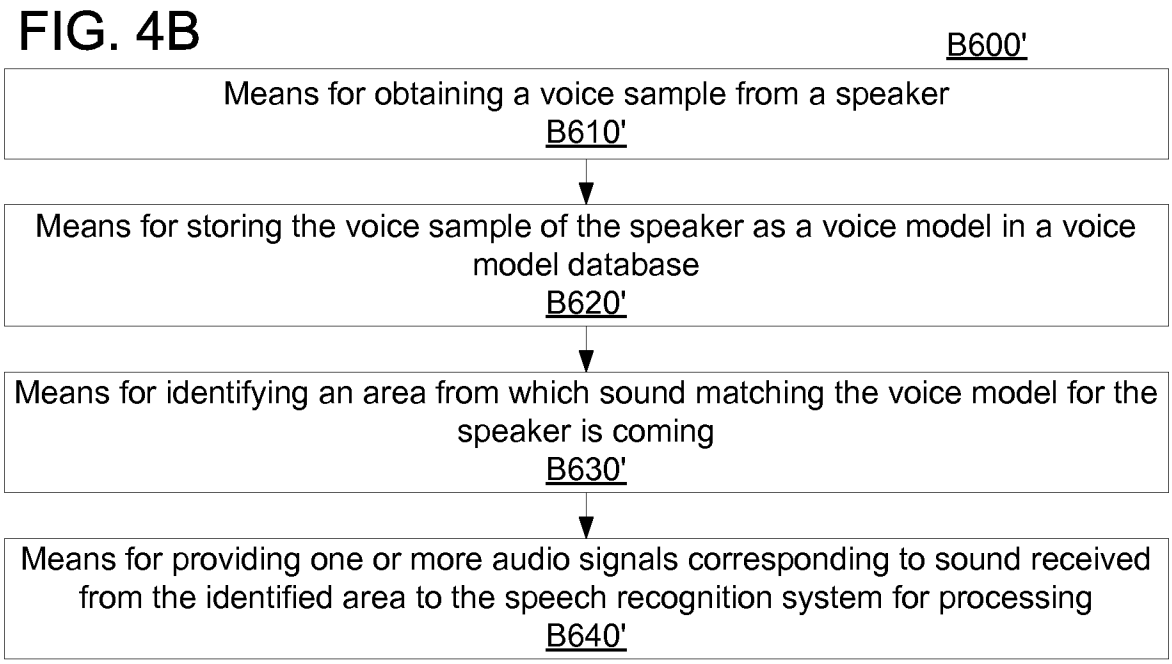

B600
Obtaining a voice sample from a speaker
B610
Storing the voice sample of the speaker as a voice model in a voice model database
B620
Identifying an area from which sound matching the voice model for the speaker is coming
B630
Providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing
B640

FIG. 4B


B600'
Means for obtaining a voice sample from a speaker
B610'
Means for storing the voice sample of the speaker as a voice model in a voice model database
B620'
Means for identifying an area from which sound matching the voice model for the speaker is coming
B630'
Means for providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing
B640'

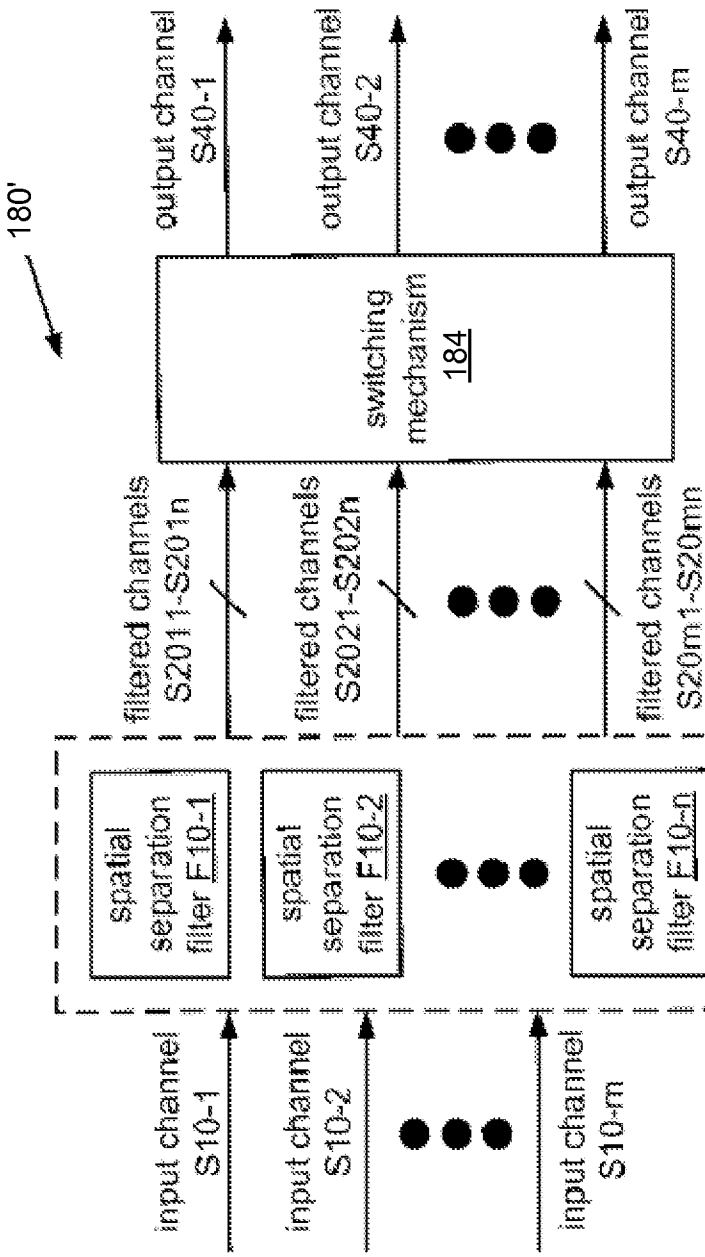

180'
input channel S10-1
input channel S10-2
input channel S10-m
spatial separation filter F10-1
spatial separation filter F10-2
spatial separation filter F10-n
filter bank 182
filtered channels S2011-S201n
filtered channels S2021-S202n
filtered channels S20m1-S20mn
switching mechanism 184
output channel S40-1
output channel S40-2
output channel S40-m

FIG. 5

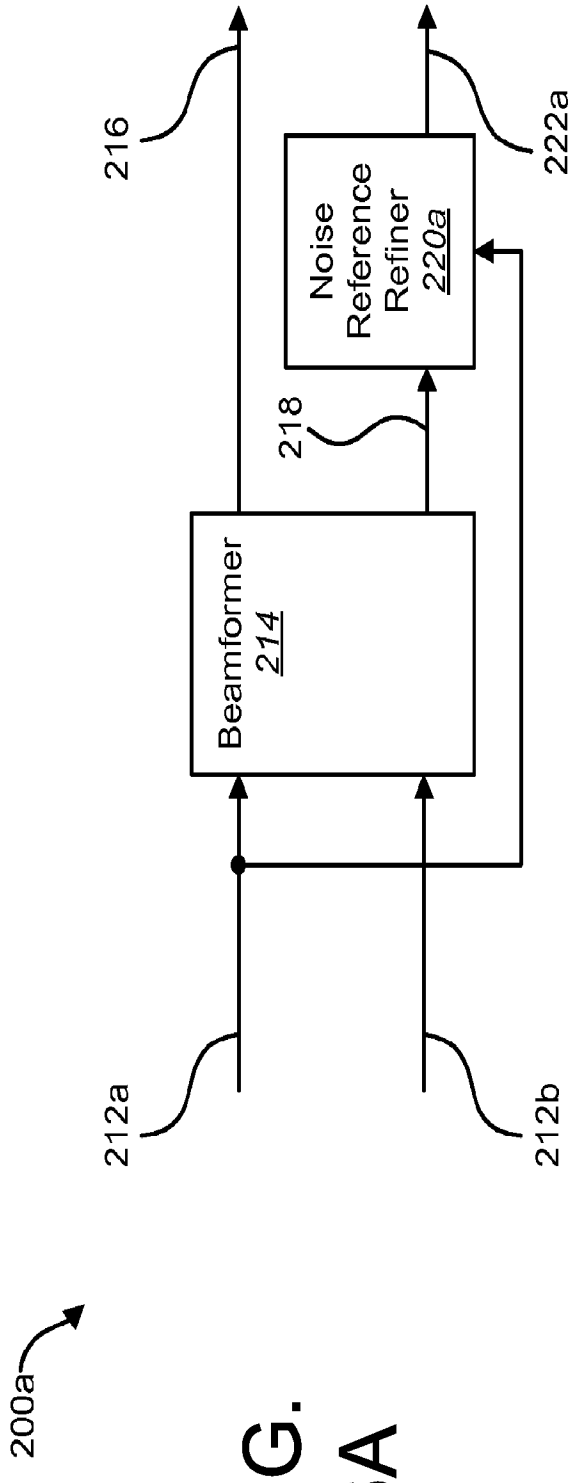

200a
216
222a
Noise Reference Refiner 220a
218
Beamformer 214
212a
212b

FIG. 6A

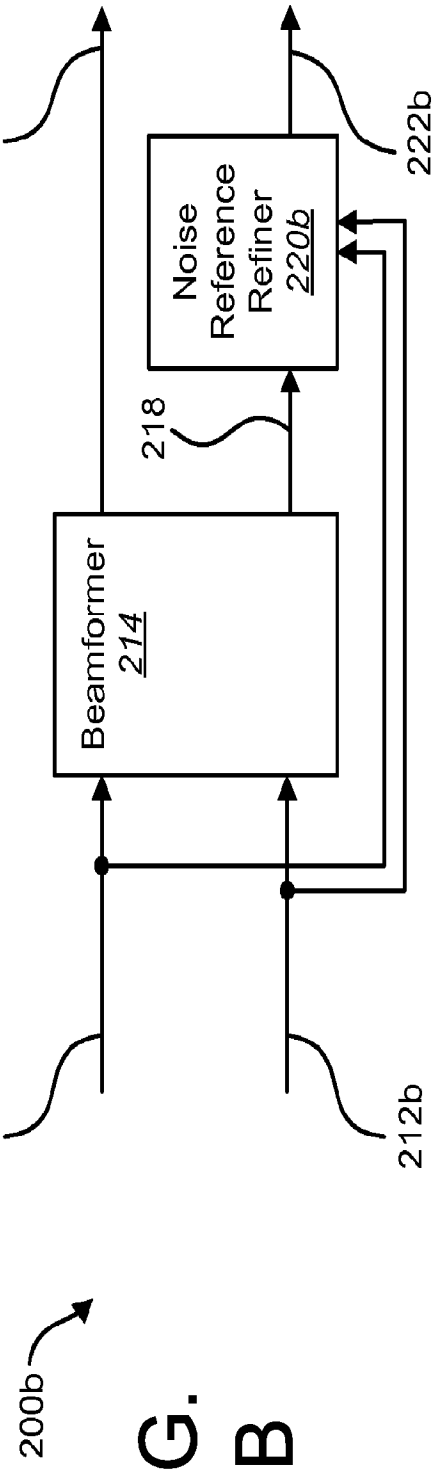

200b
216
222b
Noise Reference Refiner 220b
218
Beamformer 214
212a
212b

FIG. 6B

FIG. 7A

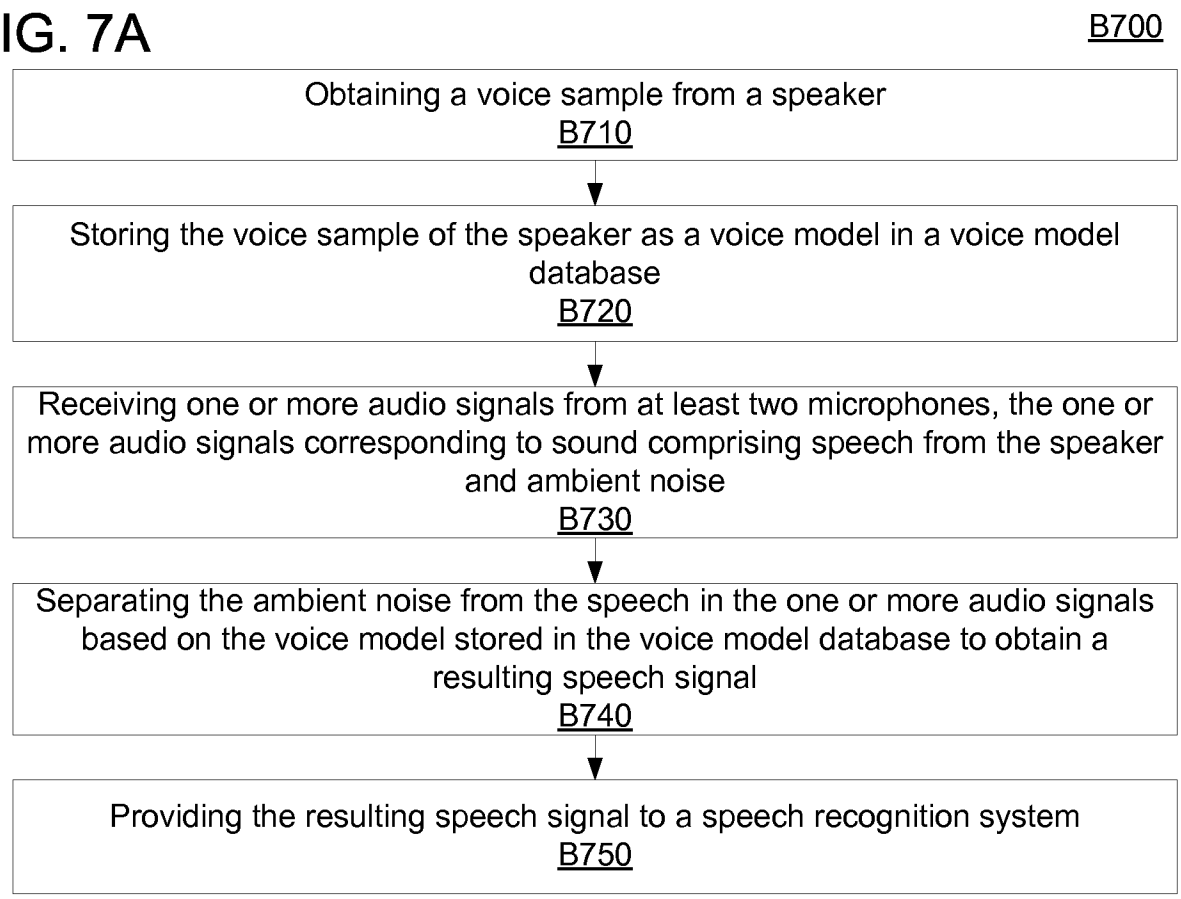
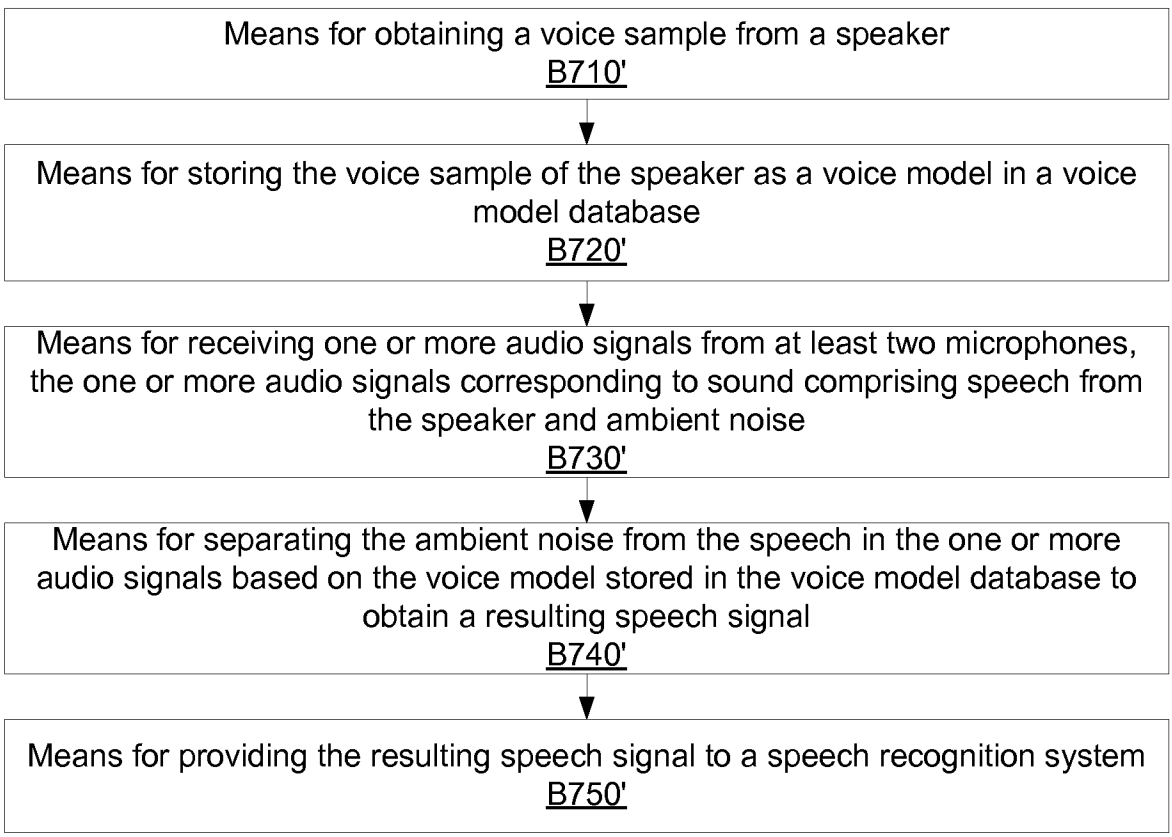

B700
Obtaining a voice sample from a speaker
B710
Storing the voice sample of the speaker as a voice model in a voice model database
B720
Receiving one or more audio signals from at least two microphones, the one or more audio signals corresponding to sound comprising speech from the speaker and ambient noise
B730
Separating the ambient noise from the speech in the one or more audio signals based on the voice model stored in the voice model database to obtain a resulting speech signal
B740
Providing the resulting speech signal to a speech recognition system
B750

FIG. 7B


B700'
Means for obtaining a voice sample from a speaker
B710'
Means for storing the voice sample of the speaker as a voice model in a voice model database
B720'
Means for receiving one or more audio signals from at least two microphones, the one or more audio signals corresponding to sound comprising speech from the speaker and ambient noise
B730'
Means for separating the ambient noise from the speech in the one or more audio signals based on the voice model stored in the voice model database to obtain a resulting speech signal
B740'
Means for providing the resulting speech signal to a speech recognition system
B750'

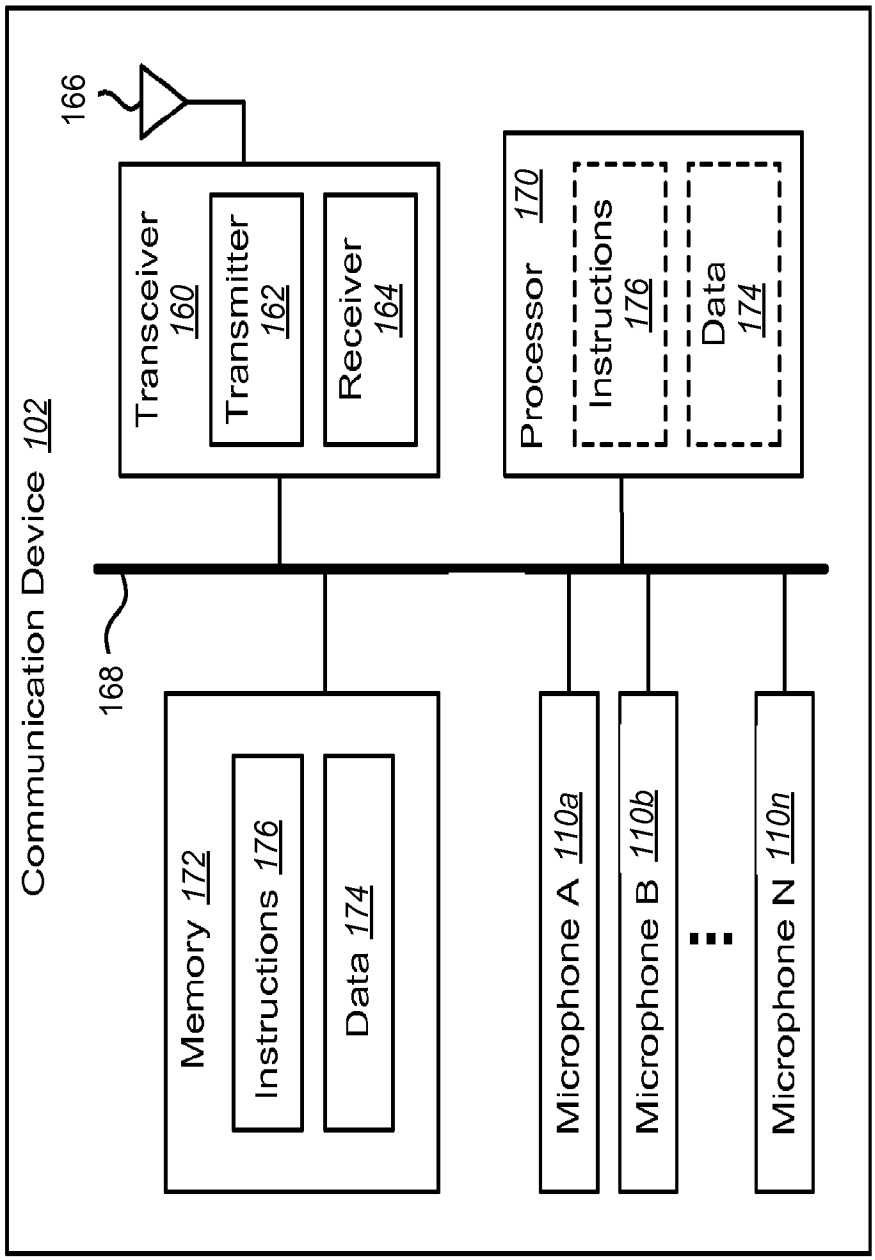

Communication Device 102
166
Transceiver 160
Transmitter 162
Receiver 164
Processor 170
Instructions 176
Data 174
168
Memory 172
Instructions 176
Data 174
Microphone A 110a
Microphone B 110b
Microphone N 110n

FIG. 8

# INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/044338

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/20
ADD. G10L17/00 G10L21/028 G10L21/0216 G10L21/0208

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

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 | LLEIDA E ET AL: "Robust continuous speech recognition system based on a microphone array", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1998. PROCEEDINGS OF THE 1998 IEEE INTERNATIONAL CONFERENCE ON SEATTLE, WA, USA 12-15 MAY 1998, NEW YORK, NY, USA,IEEE, US, vol. 1, 12 May 1998 (1998-05-12), pages 241-244, XP010279154, DOI: 10.1109/ICASSP.1998.674412 ISBN: 978-0-7803-4428-0 the whole document<br>-----<br>-/-- | 1-20 |

[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 |
|---|---|
| 31 October 2013 | 07/11/2013 |

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

1

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

# INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/044338

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | WO 03/107327 A1 (KONINKL PHILIPS ELECTRONICS NV [NL]; VIGNOLI FABIO [NL]) 24 December 2003 (2003-12-24) | 1-5, 17-20 |
| A | abstract<br>pages 1, 3-4, 6-7<br>figures 1, 3 | 6-16 |
| | ----- | |
| X | EP 2 028 061 A2 (DELPHI TECH INC [US]) 25 February 2009 (2009-02-25)<br>paragraphs 5-6, 11-15, 17, 19, 22, 24-25, 28-32, 34-35;<br>figures 1, 3-5 | 1-5, 17-20 |
| | ----- | |
| X | US 2006/212291 A1 (MATSUO NAOSHI [JP]) 21 September 2006 (2006-09-21) | 1,17-20 |
| A | paragraphs 7-9, 12, 49, 51-53, 59, 67, 74;<br>figures 1, 5 | 6-16 |
| | ----- | |
| A | DE 10 2009 051508 A1 (CONTINENTAL AUTOMOTIVE GMBH [DE])<br>5 May 2011 (2011-05-05)<br>the whole document | 1-20 |
| | ----- | |

1

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2013/044338

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| WO 03107327 A1 | 24-12-2003 | AU 2003240193 A1 | 31-12-2003 |
| | | WO 03107327 A1 | 24-12-2003 |
| EP 2028061 A2 | 25-02-2009 | EP 2028061 A2 | 25-02-2009 |
| | | US 2009055178 A1 | 26-02-2009 |
| US 2006212291 A1 | 21-09-2006 | JP 4346571 B2 | 21-10-2009 |
| | | JP 2006259164 A | 28-09-2006 |
| | | US 2006212291 A1 | 21-09-2006 |
| DE 102009051508 A1 | 05-05-2011 | CN 102054481 A | 11-05-2011 |
| | | DE 102009051508 A1 | 05-05-2011 |
| | | EP 2333768 A2 | 15-06-2011 |
| | | US 2011145000 A1 | 16-06-2011 |