



(86) Date de dépôt PCT/PCT Filing Date: 2012/06/18
 (87) Date publication PCT/PCT Publication Date: 2012/12/20
 (85) Entrée phase nationale/National Entry: 2013/12/12
 (86) N° demande PCT/PCT Application No.: US 2012/042916
 (87) N° publication PCT/PCT Publication No.: 2012/174515
 (30) Priorités/Priorities: 2011/06/16 (US61/497,699);
 2012/06/15 (US13/524,695)

(51) Cl.Int./Int.Cl. *G10L 21/00* (2013.01)
 (71) Demandeur/Applicant:
 AGERO CONNECTED SERVICES, INC., US
 (72) Inventeurs/Inventors:
 SCHALK, THOMAS BARTON, US;
 SAENZ, LEONEL, US;
 BURCH, BARRY, US
 (74) Agent: FASKEN MARTINEAU DUMOULIN LLP

(54) Titre : RECONNAISSANCE DE PAROLES DE DIALOGUE HYBRIDE POUR INTERACTION AUTOMATISEE DANS UN VEHICULE ET INTERFACES UTILISATEUR DANS LE VEHICULE NECESSITANT UN TRAITEMENT DE COMMANDE COGNITIVE MINIMAL POUR CELLE-CI
 (54) Title: HYBRID DIALOG SPEECH RECOGNITION FOR IN-VEHICLE AUTOMATED INTERACTION AND IN-VEHICLE USER INTERFACES REQUIRING MINIMAL COGNITIVE DRIVER PROCESSING FOR SAME

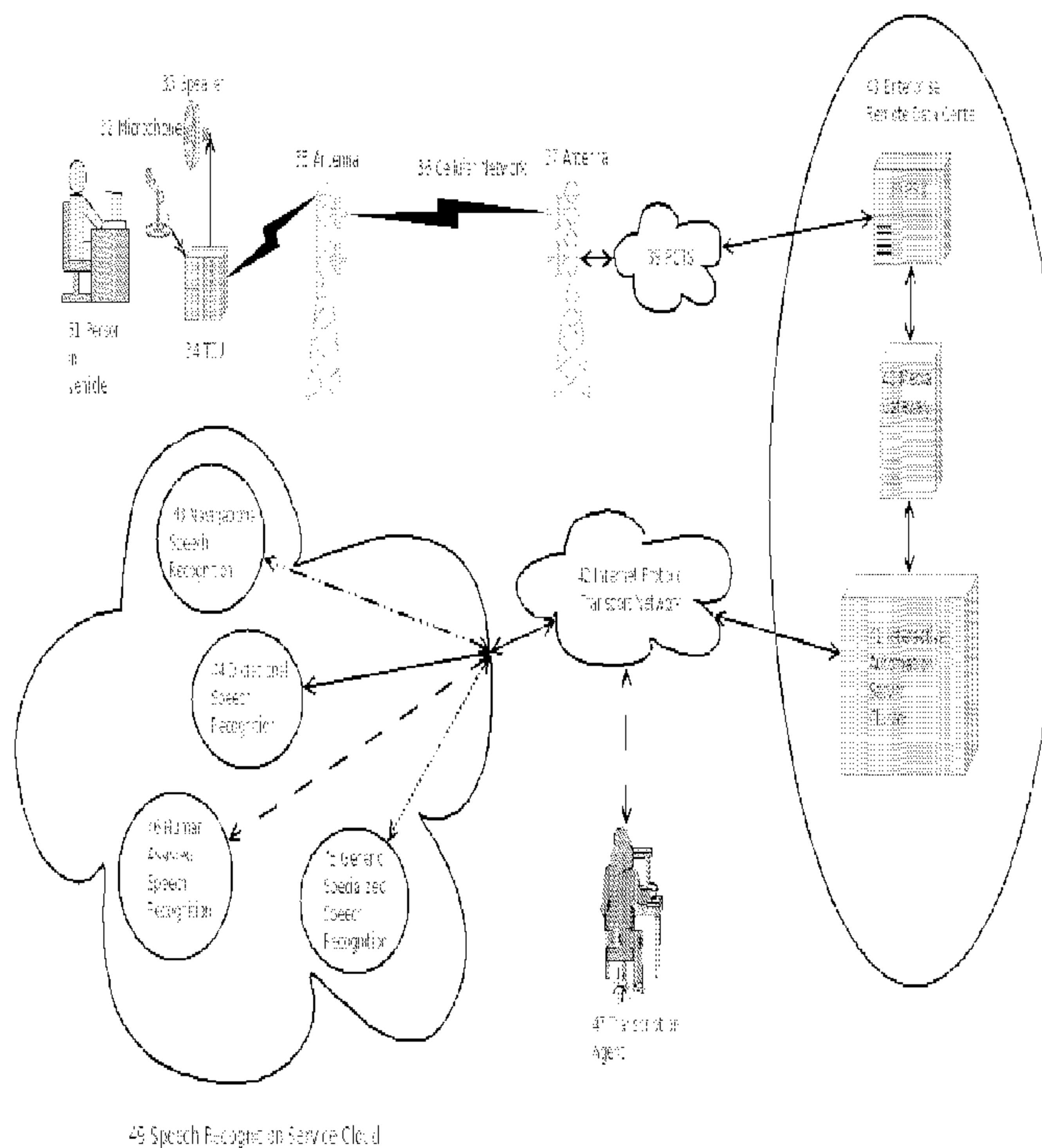


FIG. 2

(57) **Abrégé/Abstract:**

A system and method for implementing a server-based speech recognition system for multi-modal automated interaction in a vehicle includes receiving, by a vehicle driver, audio prompts by an on-board human-to-machine interface and a response with



(57) **Abrégé(suite)/Abstract(continued):**

speech to complete tasks such as creating and sending text messages, web browsing, navigation, etc. This service-oriented architecture is utilized to call upon specialized speech recognizers in an adaptive fashion. The human-to-machine interface enables completion of a text input task while driving a vehicle in a way that minimizes the frequency of the driver's visual and mechanical interactions with the interface, thereby eliminating unsafe distractions during driving conditions. After the initial prompting, the typing task is followed by a computerized verbalization of the text. Subsequent interface steps can be visual in nature, or involve only sound.

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

(19) World Intellectual Property Organization
International Bureau



(10) International Publication Number
WO 2012/174515 A1

(43) International Publication Date
20 December 2012 (20.12.2012)

- (51) International Patent Classification:
G10L 21/00 (2013.01)
- (21) International Application Number:
PCT/US2012/042916
- (22) International Filing Date:
18 June 2012 (18.06.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/497,699 16 June 2011 (16.06.2011) US
13/524,695 15 June 2012 (15.06.2012) US
- (71) Applicant (for all designated States except US): **AGERO CONNECTED SERVICES, INC.** [US/US]; 8550 Freeport Parkway, Irving, TX 75063-2547 (US).

- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **SCHALK, Thomas, Barton** [US/US]; 6637 Muirfield Circle, Plano, TX 75093 (US). **SAENZ, Leonel** [US/US]; 1502 Iroquois Circle, Carrollton, TX 75007 (US). **BURCH, Barry** [—/US]; 3007 Allister Street, Dallas, TX 75229 (US).
- (74) Agents: **MAYBACK, Gregory, L.** et al.; Mayback and Hoffman, P.A., 5722 South Flamingo Road #232, Fort Lauderdale, FL 33330 (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, 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, KM, 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, 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,

[Continued on next page]

(54) Title: HYBRID DIALOG SPEECH RECOGNITION FOR IN-VEHICLE AUTOMATED INTERACTION AND IN-VEHICLE USER INTERFACES REQUIRING MINIMAL COGNITIVE DRIVER PROCESSING FOR SAME

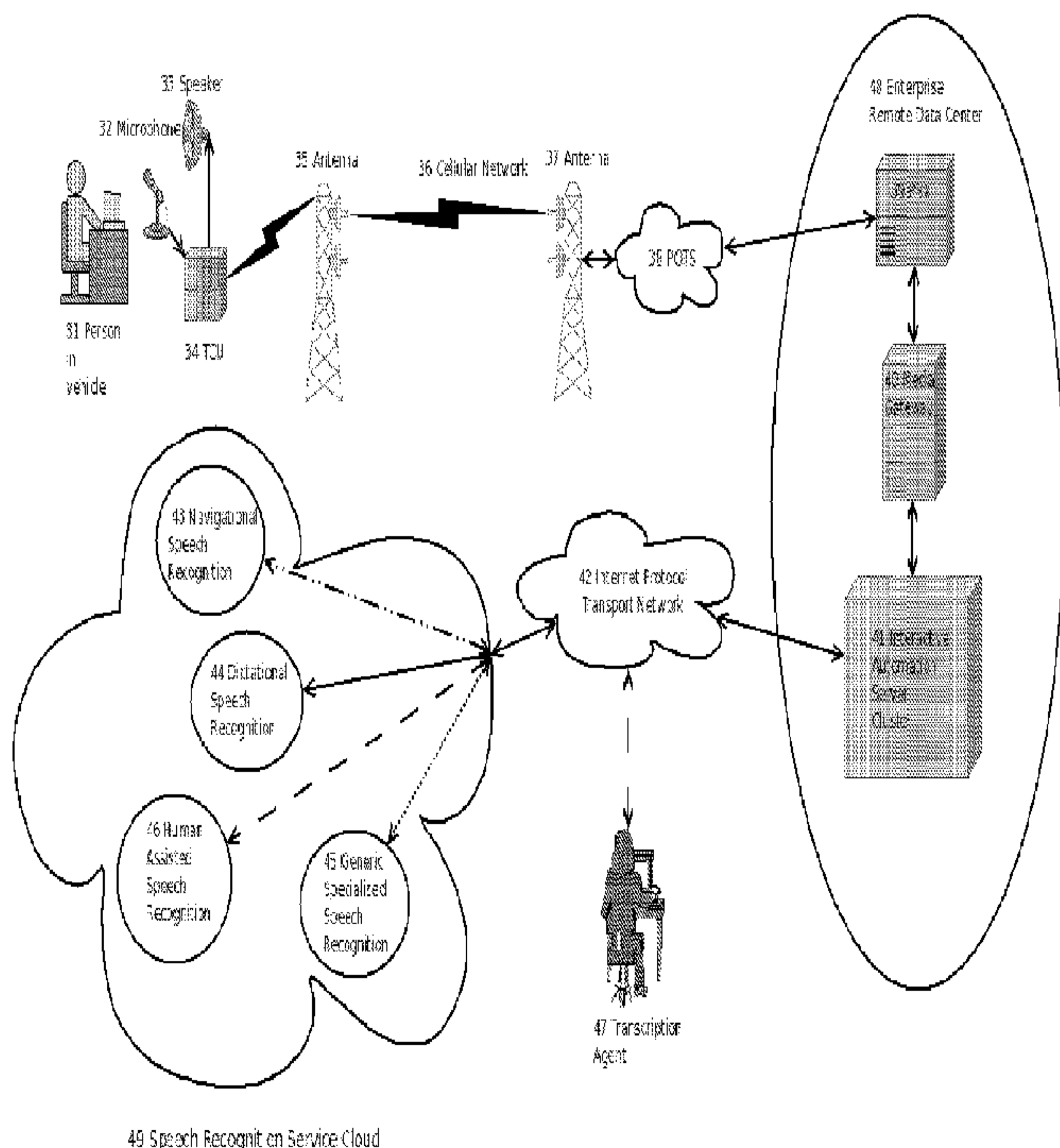


FIG. 2

(57) Abstract: A system and method for implementing a server-based speech recognition system for multi-modal automated interaction in a vehicle includes receiving, by a vehicle driver, audio prompts by an on-board human-to-machine interface and a response with speech to complete tasks such as creating and sending text messages, web browsing, navigation, etc. This service-oriented architecture is utilized to call upon specialized speech recognizers in an adaptive fashion. The human-to-machine interface enables completion of a text input task while driving a vehicle in a way that minimizes the frequency of the driver's visual and mechanical interactions with the interface, thereby eliminating unsafe distractions during driving conditions. After the initial prompting, the typing task is followed by a computerized verbalization of the text. Subsequent interface steps can be visual in nature, or involve only sound.

WO 2012/174515 A1

WO 2012/174515 A1

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, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *with amended claims (Art. 19(1))*
- *with information concerning incorporation by reference of missing parts and/or elements (Rule 20.6)*

**HYBRID DIALOG SPEECH RECOGNITION FOR IN-VEHICLE AUTOMATED
INTERACTION AND IN-VEHICLE USER INTERFACES REQUIRING MINIMAL
COGNITIVE DRIVER PROCESSING FOR SAME**

5

Technical Field

The present invention relates generally to a system and method for implementing a server-based speech recognition system for multi-modal interaction that may be applied to any interactive automated system, such as an interactive automated system that is being used inside a motor vehicle. More particularly, the present invention pertains to a system and method of utilizing multiple speech recognizers and an associated human-to-machine, in-vehicle interface to create an efficient, safe, reliable, convenient, and enjoyable experience for the motorist under driving conditions while simultaneously achieving high automation rates.

The present invention provides call center enterprises with highly effective automation to reduce costs without sacrificing the quality of service for the customer. Interactive automation should be a preferred measure of interaction by the customer, or motorist, to achieve tasks that could otherwise be handled through human/agent interaction through a call center. In the present invention, a service oriented architecture (SOA) is utilized to selectively leverage specialized speech recognizers in a uniquely adaptive fashion. The benefits of such an approach are to provide a safe and enjoyable user interface and to improve a call center's efficiency, as described herein.

The advent of telematics services, which were introduced over a decade ago, brought with it a trend to incorporate the ability of a vehicle to communicate with remote data centers and transmit location data and vehicle information related to safety, security, and emergency breakdown. "Telematics," as it is referred to in the art, includes the integration of wireless communications, vehicle monitoring systems, and location devices. Such technologies in automotive communications combine wireless voice and data capability for management of information and safety applications.

Most of the early telematics communication was achieved through wireless voice channels that were analog in nature. By law in 2008, all analog connectivity became digital and, consequently, data connectivity, such as "3G" technology, became a readily available measure for mobile devices to "connect" to the Internet. As a result of these advances, the vehicle is also being adapted to leverage data connectivity in combination with voice channel connectivity in what is referred to as the "connected car" concept.

The "connected car" concept has continued to evolve over the past few years and commercial launches of rather sophisticated vehicle services are becoming a reality. These services often rely on

vehicle location and “cloud computing,” defined as web services accessed over a data channel. Examples of these services include off-board routing, destination capture, remote-vehicle diagnostics, music downloads, traffic reporting, local searches, access to concierge services, connecting to a vehicle dealer, and roadside assistance. The term “off-board” as used herein refers to a location away from and outside the vehicle. The term “local search” as used herein refers to a point-of-interest (POI) search based on proximity to a specific location. The examples given above are regarded as being vehicle-centric in nature and many invoke some form of vocal communication with a live agent or an off-board interactive automation system.

Recently, a trend has emerged whereby motorists operate personal devices while in a vehicle, such as mobile devices, in a way that makes it unsafe while driving. Built-in user interfaces are now being added to the inside of vehicles to provide these mobile functionalities as a component of the vehicle itself. However, a number of concerns about the safety and practicality of these built-in components still exist. It is difficult to enable personal device functionality in a vehicle in a way that makes it safe while driving. The user interfaces are not at all practical for a vehicle driver to use while driving. Not only are the screens of the devices rather small, but, more significantly, the primary input modalities to operate and use a typical mobile device include some form of typing or mechanical interaction by the user with the device. Driver distraction can occur when a driver’s cognitive processing is allocated to any task that is not focused on driving a vehicle safely. Making phone calls and entering data into mobile devices are examples of tasks that can be highly distracting while driving. Conventional typing while driving is extremely dangerous because both vision and touch are involved, making it impractical to drive safely. For example, while driving a car, it does not make sense to type a message by twisting and nudging a knob until each target letter is highlighted, followed by a push of the knob (“knobbing”). However, even though it is a very awkward experience, there are cases for which “knobbing” is the only way to enter a destination into a vehicle navigation system. To reduce safety problems, some existing built-in systems attempt to purposefully limit the use of the interface only when the vehicle is stationary. Unfortunately, this stationary requirement adversely compromises the range of capabilities that may be possible with in-vehicle systems.

Accordingly, it would be beneficial to use effective speech interfaces that limit, or completely eliminate, the need for the motorist to use his or her hands to operate the interface. In addition to navigation and dialing of telephone numbers, other applications such as browsing and texting could also benefit from using speech-enabled typing. Thus, speech recognition can play a

critical role in enabling personal device functionality inside a vehicle. As a result, effective multi-modal interfaces are needed that are simple and safe to use under driving conditions.

Still, implementing speech-enabled functionalities in an environment inside a vehicle presents a unique and difficult challenge. For example, the microphone must be hands free and, therefore, may be at a distance from the speaker's mouth. Also, road noise can be harsh and non-stationary. Furthermore, there may be multiple people inside of the vehicle who are also talking, thereby making it difficult for the system to decipher the speech of one person among several different voices. Because the vehicle presents such a difficult speech recognition environment, a considerable amount of speech recognition optimization is required to achieve reasonable speech recognition performance.

A need exists to overcome the problems with the prior art as discussed above. In essence, what is needed is a speech recognition engine that is capable of complex speech tasks in a harsh environment. In addition, it would be beneficial to provide a practical system and method for an enterprise to design its speech-enabled applications, host the applications, and maintain the applications without the need for in-house expertise to support advanced speech recognition.

Furthermore, effective multi-modal interfaces are needed that are simple and safe to use under driving conditions. Unless effective speech interfaces are available, enabling personal device functionality in the vehicle will not be safe while driving. Accordingly, it would be beneficial to provide a human-to-machine, in-vehicle interface that enables safely completing a text input task while driving a vehicle.

Disclosure of Invention

The present invention provides safe measures for completing tasks that involve typing under driving conditions. Safety is maintained because the interface is designed to be extremely simple and quick to use. Simplicity to the driver is achieved by leveraging speech and hearing as primary input/output modalities during interactions within the vehicle while, at the same time, minimizing the need for visual and mechanical interactions that relate to completing tasks. Accordingly, in the present invention, an advanced human-like speech recognition system as described above is used to enable the process of typing short text strings.

More particularly, the present invention pertains to a method of prompting that begins with the speaking task and follows with a computerized verbalization of the text. Subsequent user interface steps can be visual in nature, or only involve sound. In terms of the use case, the vehicle

driver hears audio prompts and responds with speech to complete a task such as creating a text message. As a result, the present invention makes it practical for vehicle drivers to use their speech to enter text strings. By leveraging an on-premise speech-recognition solution that connects to a remote (or hosted) speech recognition system, the SOA, an asynchronous approach can be used to recognize speech. The dialog is always forward moving and the user is not asked to repeat utterances, even though the user can ask to repeat a phrase, if the application includes an appropriate query. The benefits of such an approach provide a safe and enjoyable user interface that is compelling to use while driving a vehicle.

Embodiments of the present invention provide a method for implementing an interactive automated system, comprising processing spoken utterances of a person using a processing system located in proximity to the person, transmitting the processed speech information to a remote data center using a wireless link, analyzing the transmitted processed speech information to scale and end-point the speech utterance, converting the analyzed speech information into packet data format, selecting at least one optimal specialized speech recognition engine to translate the converted speech information into text format, transporting the packet speech information to the at least one selected specialized speech recognition engine using an internet-protocol transport network, retrieving the recognition results and an associated confidence score from the at least one specialized speech recognition engine, continuing the automated dialog with the person if the confidence score meets or exceeds a pre-determined threshold for the best match, and selecting at least one alternative specialized speech recognition engine to translate the converted speech information into text format if the confidence score is low such that it is below a pre-determined threshold for the best match.

In accordance with another feature, the at least one alternative specialized speech recognition engine is agent-assisted.

In accordance with another feature, the at least one selected optimal specialized speech recognition engine is not local.

In accordance with another feature, the at least one selected optimal specialized speech engine is selected based on a given intent of the person.

In accordance with yet another feature of the present invention, the automated dialog is continued with the person prior to, or subsequent to, receiving the recognition results in an asynchronous manner.

In accordance with yet another feature of the present invention, the automated dialog is continued with the person subsequent to receiving the recognition results in a synchronous manner.

In accordance with yet another feature, the packet data and recognition results are logged for subsequent analysis.

In accordance with yet another feature of the present invention, the processing system is located on-board a vehicle.

5 In accordance with yet another feature of the present invention, the vehicle location information is also transported with the packet speech information to the at least one selected specialized speech recognition engine.

In accordance with yet another feature, the vehicle location information is logged for subsequent analysis.

10 In accordance with yet another feature of the present invention, the intent of the person includes at least one of texting, browsing, navigation, and social networking.

Embodiments of the present invention also provide an interactive automated speech recognition system comprising a processing system located in proximity to a person wherein the processing system processes spoken utterances of the person, a remote data center, a wireless link
15 that transmits the processed speech information from the processing system to the remote data center wherein the transmitted processed speech information is analyzed to scale and end-point the speech utterance and converted into packet data format, at least one optimal specialized speech recognition engine selected to translate the converted speech information into text format, an internet protocol transport network that transports the converted speech information to the at least one selected
20 optimal specialized speech recognition engine, and wherein the at least one specialized speech recognition engine produces recognition results and an associated confidence score, and based upon the confidence score, the automated dialog is continued with the person if the confidence score meets or exceeds a pre-determined threshold for the best match, or at least one alternative specialized speech recognition engine is selected to translate the converted speech information into
25 text format if the confidence score is low such that it is below a pre-determined threshold for the best match.

With the foregoing and other objects in view, there is provided, in accordance with the invention, a method for providing dynamic interactive voice recognition (IVR) over a wireless network comprises establishing a connection with a telematics control unit via the wireless network,
30 configuring a directed dialog application of at least one of a remote data center and a vehicle to provide IVR for use with expected spoken user commands, and using an open dialog application

separate from the remote data center to provide IVR for use with unexpected spoken user commands.

In accordance with another mode of the invention, the process switches back and forth between the directed dialog application and the open dialog application in accordance with pre-
5 defined criteria.

In accordance with a further mode of the invention, the directed dialog application is executed to present questions corresponding to a limited subset of possible spoken user commands.

In accordance with an added mode of the invention, a teaching mode of the directed dialog application is entered before using the open dialog application when an unexpected spoken user
10 command is received.

In accordance with an additional mode of the invention, a reduced subset of possible choices is presented in the teaching mode in order to obtain a valid spoken user command. In particular, the reduced subset is less than the limited subset.

In accordance with yet another mode of the invention, a further subset of possible choices is
15 presented in response to a user selection.

In accordance with yet a further mode of the invention, the process switches to the open dialog application absent a selection of the further subset of possible choices.

In accordance with yet an added mode of the invention, the process switches to the open dialog application when a further unexpected spoken user command is received.

In accordance with yet an additional mode of the invention, the process switches to the open
20 dialog application absent a selection of one of the reduced subset of possible choices.

In accordance with again another mode of the invention, resources for the open dialog application are provided with a speech recognition service cloud. In particular, the speech recognition service cloud provides different speech recognition systems in parallel.

In accordance with again a further mode of the invention, the different speech recognition
25 systems are selected from a group consisting of at least two of a navigation speech recognition system, a dictation speech recognition system, an audio information recognition system, and a human assisted speech recognition system.

In accordance with again an added mode of the invention, after receiving an unexpected user
30 command, the unexpected user command comprising natural user language, and the open dialog application provided by the speech recognition service cloud is used to provide information based on the natural user language.

In accordance with again an additional mode of the invention, the information comprises possible spoken user commands.

In accordance with still another mode of the invention, suggested user commands are provided in response to receiving an invalid spoken user command.

5 With the objects of the invention in view, there is also provided a method for minimizing task completion time using dynamic interactive voice recognition (IVR) over a wireless network comprises the steps of establishing a connection with a telematics control unit via the wireless network, configuring a directed dialog application of at least one of a remote data center and a vehicle to provide IVR for use with expected spoken user commands, receiving one of a plurality of
10 expected spoken user commands, depending on the user command received, prompting a user for further spoken information in order to complete one or more actions required by the user command received, and completing the one or more actions upon receipt of the further spoken information.

In accordance with still a further mode of the invention, the expected spoken user command received comprises a shortcut.

15 With the objects of the invention in view, there is also provided a system for providing dynamic interactive voice recognition (IVR) over a wireless network, comprising a remote data center comprising a communications device operable to establish a connection with a telematics control unit via the wireless network, a directed dialog application operable to provide IVR for use with expected spoken user commands, and a communications subsystem, and an open dialog
20 application separate from the remote data center, communicatively connected to the remote data center through the communications subsystem, and operable to provide IVR for use with unexpected spoken user commands.

In accordance with a concomitant mode of the invention, the remote data center is configured to use resources for the open dialog application that are provided by a speech recognition service
25 cloud.

Additional advantages of the present invention will be set forth in the Detailed Description which follows and may be understandable from the Detailed Description or may be learned by practice of exemplary embodiments of the invention. Still other advantages of the invention may be realized by any of the instrumentalities, methods, or combinations particularly pointed out in the
30 claims. Although the invention is illustrated and described herein as embodied in one or more exemplary embodiments of systems and methods for providing hybrid dialog speech recognition for automated interaction and in-vehicle user interfaces requiring minimal cognitive driver processing, it

is, nevertheless, not intended to be limited to the details shown because various modifications and structural changes may be made therein without departing from the spirit of the invention and within the scope and range of equivalents of the claims. Additionally, well-known elements of exemplary embodiments of the invention will not be described in detail or will be omitted so as not to obscure the relevant details of the invention. The system and method of operation of the invention, however, together with additional objects and advantages thereof, will be best understood from the following description of specific embodiments when read in connection with the accompanying drawings.

Brief Description of Drawings

10 The accompanying figures, where like reference numerals refer to identical or functionally similar elements throughout the separate views, and which together with the detailed description below are incorporated in and form part of the specification, serve to further illustrate various embodiments and to explain various principles and advantages all in accordance with the present invention.

15 **FIG. 1** is a system architecture diagram illustrating components of a speech recognizer according to an exemplary embodiment of the present invention.

FIG. 2 is a system architecture diagram illustrating components of a service-oriented architecture for in-vehicle speech recognition according to an exemplary embodiment of the present invention.

20 **FIG. 3** is a system architecture diagram illustrating components of a connected vehicle speech recognizer according to an exemplary embodiment of the present invention.

FIG. 4 is a graph illustrating usability and flexibility of the system according to an exemplary embodiment of the present invention utilizing directed and open dialogs.

25 **FIG. 5** is a flow diagram illustrating the system of processes that comprise a multi-modal user interface design and how commonalities are shared among a number of exemplary user interfaces according to an exemplary embodiment of the present invention.

FIG. 6 is a process flow diagram of a synchronous speech recognition approach aimed at showing the limitations of the user experience.

30 **FIG. 7** is a process flow diagram of an asynchronous speech recognition approach aimed at showing the advantages of the asynchronous approach according to an exemplary embodiment of the present invention.

FIG. 8 is a process flow diagram of a method for providing dynamic interactive voice recognition (IVR) over a wireless network according to an exemplary embodiment of the present invention.

FIG. 9 is a process flow diagram of a method for minimizing task completion time using dynamic interactive voice recognition (IVR) over a wireless network according to an exemplary embodiment of the present invention.

Other features that are considered as characteristic for the invention are set forth in the appended claims.

10 Best Mode for Carrying out the Invention

As required, detailed embodiments of the present invention are disclosed herein; however, it is to be understood that the disclosed embodiments are merely exemplary of the invention, which can be embodied in various forms. Therefore, specific structural and functional details disclosed herein are not to be interpreted as limiting, but merely as a basis for the claims and as a representative basis for teaching one skilled in the art to variously employ the present invention in virtually any appropriately detailed structure. Further, the terms and phrases used herein are not intended to be limiting; but rather, to provide an understandable description of the invention. While the specification concludes with claims defining the features of the invention that are regarded as novel, it is believed that the invention will be better understood from a consideration of the following description in conjunction with the drawing figures, in which like reference numerals are carried forward.

Before the present invention is disclosed and described, it is to be understood that the terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting. The terms “a” or “an”, as used herein, are defined as one or more than one. The term “plurality”, as used herein, is defined as two or more than two. The term “another”, as used herein, is defined as at least a second or more. The terms “including” and/or “having”, as used herein, are defined as comprising (i.e., open language).

As used herein, the term “about” or “approximately” applies to all numeric values, whether or not explicitly indicated. These terms generally refer to a range of numbers that one of skill in the art would consider equivalent to the recited values (i.e. having the same function or result). In many instances these terms may include numbers that are rounded to the nearest significant figure. The terms “program,” “software,” “software application,” and the like as used herein, are defined as a

sequence of instructions designed for execution on a computer system. A “program,” “software,” “computer program,” or “software application” may include a subroutine, a function, a procedure, an object method, an object implementation, an executable application, an applet, a servlet, a source code, an object code, a shared library/dynamic load library, and/or other sequence of instructions
5 designed for execution on a computer system.

Referring now to **FIG. 1** of the drawings in detail, there is shown a system architecture diagram representing the basic components of a speech recognizer in connection with a remote data center that require special optimization under conditions in which the environment is harsh and the recognition task is complex (e.g., recognition of dictation or a street address). Even when a speech
10 recognizer is highly tuned, accuracy can be unpredictable simply because it is virtually impossible to model every sound that a person can make when speaking into a microphone. However, when the user is cooperative and possesses some experience in using the system, acceptable results can be achieved.

AUTOMATED SPEECH RECOGNITION

15 Generally speaking, there are a number of complicated components to any automatic speech recognition engine that include acoustic models, grammars, dictionaries, and language models. In simple terms, “acoustic models” represent how speech sounds in a target environment, “grammars” represent what can be recognized during an application, “dictionaries” represent the way that words are to be pronounced, and “language models” govern the allowable sentence structure. In what
20 follows, a brief review of selected components of a speech recognition engine is made to gain an appreciation of the speech technology required by the invention as it is disclosed here forth. A detailed discussion of the fundamental components of spoken language processing and speech recognition systems is found in “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” by Xuedong Huang, et al., May 5, 2001, the contents of which are fully
25 incorporated herein by reference as though fully set forth.

In detail, “acoustic models” **15** are statistical representations of phonetic sounds that are produced under specific environmental conditions. Phonetic sounds can be thought of as sub-units of spoken words that are to be recognized by an automated speech recognition (ASR) system. The environmental conditions are characterized by numerous components, including the microphone
30 type and its placement, the surrounding acoustic media, audio transmission properties, background noise, signal conditioning software, and any other variable that influences the quality of the sound that the speech recognizer processes. Acoustic models **15** are needed for high accuracy speech

recognition, and, the more highly tuned the acoustic model, the more accurate is the speech recognition. Speech data collections form the basis of acoustic models. However, live adaptation is used for in-the-field tuning. Thousands of recordings that represent environmental extremes of a target recognition environment constitute a “good” base of speech data.

5 “Grammar” or “Grammars” **17** are a set of rules that define the set of words and phrases (i.e., a vocabulary) that may be recognized during voice applications. An application may have several grammars such as “yes/no,” numerical digits, street names, action menu items, etc. To maximize accuracy, only the necessary vocabulary should be active at any point of an application call flow. For example, numerical digits would not be a part of the active vocabulary for recognition during a
10 “yes/no” query unless there is a special reason, as there is a risk, for example, that the sound “oh” for the number zero may be confused with the sound of the word “no.” It is important to take into consideration that grammars containing too many short words are more prone to exhibiting low accuracy because short words are more difficult to recognize than long, multi-syllabic words. In general, the longer the word, the more phonetic content there is available for distinguishing it from
15 other words. For purposes of speech recognition, a difficult vocabulary is the alphabet in which there are short sounds that, in many instances, rhyme with or sound familiar with one another.

Grammars **17** rely on “dictionaries” for pronunciation information. Dictionaries are commonly referred to as “lexicons.” A “lexicon” **16** is a collection of words and their associated pronunciations in terms of phonetic transcriptions. Similar to a common dictionary, pronunciation is
20 specified by a standard symbol set.

“Language models” **18** are designed to assist the recognition matching process for multiple words in a phrase or a sentence. Common languages are statistical in nature and attempt to assign a probability to a sequence of allowable words by a probability distribution. Language modeling can be used in many natural language processing applications such as speech recognition, machine
25 translation, part-of-speech tagging, parsing, and information retrieval. In speech recognition, to predict the next word in a speech sequence, a language model can be used to capture the properties of a language.

In summary, for any given speech recognition technology, acoustic models **15**, grammars **17**, lexicons **16**, and language models **18** are optimized to reach a high level of accuracy. Ideally, if a
30 human can understand a command or a structured utterance, then a properly tuned speech recognizer should be able to recognize it too. Accordingly, using real-world recordings for adaptation purposes is one way to improve accuracy.

A key feature of the present invention lies in a particular division of duties -- the performance of the complex speech recognition tasks are separated from the system that is hosting the application. The base speech application contains a dialog structure that relies on its own recognizer for basic command and control. For complex speech recognition tasks, accessible specialized recognizers can
5 be used.

From a system perspective, latency, as perceived by the user, must be low to avoid user frustration. That is, the recognition system must respond quickly to spoken utterances. When an application connects to specialized speech recognizers through an Internet data connection, the connection time is extremely fast, thereby yielding a reasonable amount of time between the end of a
10 spoken utterance and the resulting action by the application (e.g., a subsequent prompt or a vehicle display change). The time to convert a wave file (i.e., a waveform audio file that can be compressed to minimize data size) into a packeted data format for Internet transmission is negligible. However, if a voice channel is used to pipe audio wave files to a remote recognizer, the connect time may prevent a good user experience from being possible because a typical telephone connection time is
15 approximately 10 seconds. Speech recognition that is server-based leverages the remote device's Internet connection to transmit packeted audio and to have returned recognition results almost instantaneously. The remote device acts as the client and the recognition is performed off-board by way of a data channel.

The present invention is unique as it viably mixes recognition engines in real-time with a
20 real-time dialog with humans. The present invention deals with an enterprise automated system having its own speech recognition resources and an actual dialog occurs (i.e., audio prompting occurs). The Internet is not accessed wirelessly -- a telephone voice channel serves as the means of communication between the person, or motorist, and the enterprise automated system. The present invention provides an automatic interactive system at an enterprise remote data center (ERDC) that
25 leverages multiple specialized speech recognizers over a data channel (i.e., the Internet) and allows, by way of a wireless voice communication channel, a person, such as a motorist, to interact in a hands-free environment with the automated system, the automated system being capable of understanding complex speech requests. The primary advantages of hosting the application on-premise at the ERDC include ease of back-end integration, control of application design and support,
30 improvement of application maintenance, and cost-effective implementation due to server sharing. Because the platform is off-board, the application dialog design can easily be modified without changing any remote, or in-vehicle, hardware or software.

As a result, the ERDC can prototype and launch automated interactive applications that are designed internally. This means that complete application ownership is possible even though sophisticated speech recognition is used within the application and candidate speech recognition engines can be evaluated without making application changes. Also, multiple-language speech recognition is easily accommodated through the use of specialized speech recognition services.

In terms of cost, the ability to share internal server-based speech recognition resources and the web-accessed server-based speech recognition resources across a large spectrum of different vehicles is beneficial. For example, each channel of a server-based, interactive, automation system could accommodate numerous vehicles simultaneously.

Locating an automated interactive automation service cluster within the ERDC provides substantial benefits over an embedded speech system inside a vehicle. For example, this architecture provides increased operational flexibility and control from the call center. Efficiency is increased because content can be added or modified with centralized hardware and/or software. Calls from the vehicles can be monitored and improvements can be made at the centralized locations, rather than inside each vehicle. Also, there is an improved scalability as computer resources can be shared across a large number of vehicles. To supplement these benefits provided by the invention, another advantage arises because a “thin” client can be located in the vehicle using standard telematics control units (TCUs), rather than a specialized on-board computer. Furthermore, the inventive system provides the ability to manage personalization in terms of customer preferences.

The present invention is directed to a system and method that leverages multiple specialized speech recognizers that are accessed on-premise through an Internet-protocol transport network. The ERDS is equipped with highly available connectivity for high-speed Internet access, thereby eliminating wireless coverage from being an issue. The speech application is hosted on an automated interactive system located within the ERDC (or call center). All application maintenance and updating can be managed by the enterprise remote data center (ERDC) without the need of costly subject-matter experts in speech recognition. For purposes of providing an illustrative non-limiting example, this particular embodiment is shown as being used in connection with motor vehicles in **FIG. 2**. However, the system and method of the present invention is applicable to all interactive systems.

Referring to **FIG. 2**, in one exemplary embodiment, after the motorist or vehicle driver **31** initiates a telematics connection, the vehicle’s telematics control unit (TCU) **34** connects to the ERDC **48** by way of a wireless communications link that includes antennas **35, 37** and the cellular

network **36**. The final stage of the telematics connection includes the telephone network (e.g., POTS) **38**, which terminates at a communications device, e.g., the PBX **39**, within the ERDC **48**. In this exemplary embodiment, the ERDC **48** is comprised of a media gateway **40** and an interactive automation service cluster **41**. The media gateway **40** manages the communications media session
5 between the PBX **39** and the interactive automation service cluster **41**. The interactive automation service cluster **41** is the central point of intelligence of the inventive systems and methods as described in the following text.

A telematics request can be accomplished, for example, by the vehicle driver **31** pressing a button, in response to which the TCU **38** initiates a connection with the ERDC **48** as described
10 above. After the connection is established, the vehicle driver **31** is able to hear audio prompts, for example, through the in-vehicle speaker **33** located in proximity to the vehicle driver **31**. With the in-vehicle speaker **33** and a microphone **32**, an automated interaction takes place with the vehicle driver **31**. The interaction could begin, for example, with the audio prompt “How may I help you?” Simultaneously and in a seamless fashion, when the telematics connection is established to the
15 ERDC **48**, data information such as the vehicle’s location, vehicle model information, vehicle driver information, diagnostic information, etc. is captured and communicated via a data channel to the interactive automation service cluster **41**.

In response to the initial audio prompt (e.g., “How may I help you?”), the vehicle driver may then respond out loud with a request and say, for example, “I need to find an Italian restaurant
20 nearby” or “I want to text my friend Bob.” Depending on the type of request made by the vehicle driver **31**, the interactive automation server cluster **41**, which is comprised of a group of servers interconnected together to form the ERDC-located speech system, automatically selects the appropriate speech recognition engine. The appropriate speech recognition engine could be located internal to the interactive automation server cluster **41** within the ERDC **48** or could be externally
25 available to the interactive automation server cluster in a speech recognition service cloud **49** that may be accessed through the world-wide-web (referred to as “cloud computing”) from one or more speech vendors that offer a URL access point to their speech server farm. The speech engine that is selected depends on the type of request made by the vehicle driver **31**. For example, simple “yes/no” queries or action menu selections may be handled by the recognition engine hosted within
30 the interactive automation server cluster **41**. More complex speech recognition tasks, such as recognizing a spoken sentence, could be handled by a remote dictation engine **44**. The Internet protocol transport network **42** is highly available to the interactive automation server cluster **41** and

operates at a high speed, making it practical to recognize complex speech utterances in just a matter of seconds from the time the vehicle driver utters the directive. As will be described below in further detail, the process for selecting the appropriate recognition engine and whether or not to direct the user to a live operator can be dependent upon the complexity of the speech to be
5 recognized.

When handling a complex speech recognition task, such as recognizing a navigational destination, a remote navigational engine **43** by way of the Internet protocol transport network **42** can perform the handling. The speech application is executed within the interactive automation server cluster **41** and waits for a response from the remote navigational engine **43** before proceeding
10 with the next step (e.g., a subsequent prompt, visual display of the destination information, or an end to the interactive session).

For each spoken utterance, a recognition process occurs and, as part of the process, the recognizer creates an “n”-best list of the top hypotheses, or “recognition results.” In other words, if “n” equals five, the recognizer generates up to five text representations of what was possibly spoken,
15 each with an associated probability of correct recognition. The variable “n” may be a pre-defined, limited number and/or is dependent upon the number of results returned that meet or exceed a certain probability of correct recognition. Each recognition hypothesis is assigned a confidence score (or probability) that is typically normalized to 1. If the top choice is assigned a confidence score above a specified threshold (e.g., 0.85), the spoken input is considered to be a final recognition result
20 without requiring further processing. Exemplary alternative processes for performing such speech recognition are detailed in U.S. Patent Nos. 7,373,248 and 7,634,357 and in U.S. Patent Application Serial No. 12/636,327, filed December 11, 2009, the disclosures of which have been incorporated herein.

It may be that the result provided by the remote navigational engine **43** is of low confidence,
25 meaning that the spoken speech was not automatically recognized with a certain level of confidence. To perform the recognition in such an instance, the corresponding audio wave file could be passed over the web to a live transcription agent **47**. Should this situation occur, the speech application, executed within the interactive automation server cluster **41**, waits for a response from the transcription agent **47** before proceeding to the next step (e.g., a subsequent prompt, a visual display
30 of the destination information, or an end to the interactive session). Exemplary processes for making the decision of transferring back and forth between the voice recognition engines of the interactive

automation server cluster **41** of the ERDC **48** and the speech recognition service cloud **49** are described in further detail below.

By accessing the speech recognition service cloud **49** in real time, the user experience is much improved over the prior art. The interactive automation server cluster **41** serves as the intelligence behind the automation experienced by the vehicle driver, or other users. The fact that the systems and methods are agnostic (i.e., not tied to one particular technology vendor) in choosing the speech recognition engine means that the system and method highly robust and flexible because multiple best-in-class recognizers can be leveraged. In addition, because the system and method of the present invention incorporates agent assistance into the implementation model, the user experience can also “feel” human in the wake of noise and heavy accent characteristics. Finally, the vehicle driver or other user **31** can be connected to a live agent (see, e.g., **FIG. 7**) to handle extreme circumstances. Such a connection is easily managed by the interactive automation server cluster **41** if the application is configured accordingly, examples of which are described in the following section.

15 HYBRID DIALOG

Dialog designs of the invention improve the connected car speech process. The methodologies discussed herein can be applied to any number of tasks including managing search results, requesting information, and texting through voice, to name a few. As driver distraction is desired to be minimized, the inventive system utilizes the positive characteristics of both limited dialog and expansive dialog, the latter of which can include live voice exchanges. The inventive system employs dynamic interactive voice recognition that switches seamlessly between a limited dialog having minimal vocabulary to an open dialog having a virtually unlimited vocabulary. Throughout the process, the user is able keep his/her hands on wheel because the inventive systems and methods and their human-machine interfaces enable the user to get information to/from the remote data center. The inventive dialog processes and interfaces are configured such that communication flow should work to make it easier for the user to obtain the information desired and avoid a situation where the user does not like the process or quits learning the process because the process is too complicated to learn. The systems and methods are configured to react in different ways depending on whether the user is speaking within the defined vocabulary or the user is generating instructions that are outside the defined vocabulary. A power user presumably limits uttered speech to the defined vocabulary, while a novice user utters instructions that are generally outside the defined vocabulary.

To accomplish dynamic interactive voice recognition, the invention provides a hybrid system that moves back and forth between a structured/directed dialog mode, e.g., power mode, and an unstructured/open dialog mode, e.g., novice mode. To explain this process, reference is made to **FIGS. 3** and **4**. When having an interactive voice exchange, many processes start with a limited grammar. This means that the user is asked questions that have a small subset of possible valid responses. Herein, this limited process is referred to as a “directed dialog.” Some examples of a directed dialog include binary responses such as yes/no, up/down, left/right, or a set of numbers that is limited between 1 and 10. In such cases, the ERDC **48** can easily handle expected responses when correct. When the user’s utterances are incorrect or invalid, the system recognizes this as an indication that the user needs help. At this point, the system transitions into a teaching mode.

When the unexpected (i.e., non-matching) utterance is received, the teaching mode does not try to figure out what was said by the user. Instead, the system utilizes a process that tries to make it easier for the user to say a valid response. In attempting to obtain a valid response from the user, the system provides a small or reduced subset of possible choices. Assume, for example, that the user has thirty possible valid responses in a directed dialog but that the user does not give a valid response. Instead of trying to determine what was actually said, which would require use of off-board, process-intensive, voice recognition, the system provides a second level of choices that is much smaller than the thirty possible valid responses. The system is configured to give the user broad categories from which to choose a valid command. For example, the system says “please say traffic, new destination, or gas prices.” In this example, each of these three categories broadly describe a number of different possible valid responses within the set of thirty. The traffic sub-category, if selected, could provide six different choices. Likewise, the destination sub-category, if selected, could provide five different choices and the gas sub-category, if selected, could provide four different choices. As such, instead of listing thirty items to the user, which is unworkable, only three are listed for selection. If one is not appropriate, the user can be prompted for “more choices.” When one is selected, the small, manageable subset is uttered to the user.

The strategy of the invention is not just to skip steps but leverage the low-vocabulary mode (i.e., directed dialog) and teach user the available categories. The system will not say to the user: “I’m sorry I don’t understand.” If no categories are selected or if the user’s utterance is still not understood with confidence, then the system switches to a high-vocabulary mode by having an open dialog with the user (i.e., natural language). (Typically, a dialog session would end because the user has shown an inability to say simple commands – like someone failing to enter their password

correctly on a web page.) If natural language is actually needed (e.g., the user gets through the directed dialog interaction), the inventive system proceeds to the speech recognition service cloud **49**. This process is illustrated in FIG. 3 where the directed dialog mode with the limited vocabulary occurs at the ERDC **48** and the open dialog mode with the infinite vocabulary occurs in the cloud **49** where the resources are outside the IVR server of the ERDC **48**. The cloud **49** is able to provide different speech recognition systems in parallel. If navigation information is desired, a navigation speech recognition system **102** can take over and quickly provide the ERDC **48** with the information to be sent to the user. If dictation is desired (e.g., text or email), a dictation speech recognition system **104** can take over and quickly provide the ERDC **48** with the information to be sent to the user. If audio information is desired (e.g., music), a voice search speech recognition system **106** can take over and quickly provide the ERDC **48** with the information to be sent to the user. Finally, if human assistance is desired, a human assisted speech recognition system **108** can take over and quickly provide the ERDC **48** with the information to be sent to the user. When assistance is needed, the system can use the confidence scoring process set forth in co-pending U.S. Patent Application Serial No. 12/636,327 or U.S. Patent Nos. 7,373,248 and 7,634,357, for example, to address the invalid user response.

Because the system specifically designs choices for the user, the system handles out-of-vocabulary utterances in a way that engages the user instead of alienating the user. **FIG. 4** is a graph illustrating the problem associated with use of the directed and open dialogs. The limited directed dialog is usable only to limited extent because of the small vocabulary. It is flexible to a point (dashed vertical line A), but starts to become unusable and inflexible as the user utters responses that are unintelligible or are outside the vocabulary. Open dialog is very flexible because it can interpret virtually all responses from a user. However, open dialog requires time and outside processing and, possibly, requires live operator assistance and, therefore, there is a point at which open dialog is undesirable (dashed vertical line B). From this, there exists a range in which both directed and open dialogs are disadvantageous. The system of the invention guides the user and utilizes the directed and open dialogs to prevent use of either dialog in a disadvantageous situation.

USER INTERFACE

In conjunction with the system and method of the speech recognition solution that is described above, the present invention also provides a user interface that enables functionality in a vehicle in a way that makes it safe while driving. The user interface of the present invention allows

navigation, dialing, web browsing, text messaging (texting), and other applications for mobile devices by speech-enabled typing.

Generally, the primary objective of the user interface is to make it practical for a vehicle driver to access a set of applications that are controlled and managed by user interfaces that share a strong degree of commonality. For example, the user interface for texting shares commonality with the user interface for web browsing. Likewise, the user interface for web browsing shares commonality with the user interface for local search and full address entry for navigation. By design, there is virtually no learning required by the vehicle driver. The invention utilizes a three-step approach for completing tasks that normally require conventional typing. The three steps are:

- intent initiation;
- speaking a phrase; and
- managing the result.

No typing is required.

In use, the vehicle driver initiates the task by indicating intent. Intent can be communicated through a specific button push, touching a specific icon, or saying a specific speech command such as "I want to send a text message." Once the user indicates intent, the user is prompted by speech to say a phrase that will match the intended text for a text message or the intended text to enter in a search box, or a destination category, name, or address. Most significantly, the invention makes it practical for vehicle drivers to use their own speech to enter text strings. The recognized result is, then, managed by the user in a way that depends on the task. Web browsing would entail a simple glance at a screen. Texting would entail saying the name of the recipient and then speaking the content of the text message. Destination entry could entail touching a screen to download a specific destination to an on-board navigation system. Other examples follow the same pattern: input intent; speak a phrase; and manage the result. As set forth above, the user interface of the inventive systems and methods requires advanced speech recognition that allows free-form dictation in the highly challenging environment of a vehicle's interior.

It is noted that the present invention also encompasses asynchronous speech recognition, which means that the user interface can step forward in execution before recognition results are obtained. For example, a user could speak a text message and be prompted to say the name of the recipient *before* the spoken text message is actually recognized. The user interface can include playing back the later-recognized text message along with the target recipient. Longer latencies associated with obtaining recognition results can be managed by sending the message without

confirmation but subsequent to the user interaction within the vehicle. For example, the message may be recognized and sent twenty (20) seconds later, without the user knowing exactly when the message was sent. However, some tasks, such as web browsing or local search, are sensitive to timing, and a synchronous approach is only practical when the latencies are controlled to be within
5 several seconds, analogous to delays typically experienced with conventional web browsing.

The asynchronous speech recognition approach of the inventive systems and methods has advantages that extend beyond the vehicle. For example, a conventional interactive voice response system (IVR) typically includes error handling dialogs that slow down the interactive process and often cause user frustration when recognition failures occur. However, for purely asynchronous
10 speech recognition, the dialog is always forward moving (i.e., the next-level prompts occur immediately after a user speaks even if the speech is not recognized) and the user is not asked to repeat spoken utterances. Furthermore, a portion of the dialog can be synchronous and thereby allow for the system to ask a user to confirm a phrase that was spoken by the user or the user can cause the system to repeat a result by responding to a yes/no query (e.g., answering “no” to the
15 system’s query of “did you say...?”).

According to an exemplary embodiment of the present invention, **FIG. 5** shows a representation of the inventive in-vehicle, user-interface solution, based on a system of user interface processes or applications that involve or require the same basic steps albeit accomplishing the steps by different methods and producing different results or providing different functions to the user. The
20 user interface is multi-modal in nature and is based on three steps that are common among a variety of applications including, but not limited to, texting **210**, browsing **213**, navigation **216**, and social networking **219**, as well as other applications **222**. Step one **225** involves establishment of intent, or selecting the application intended to be used. Application selection may be achieved by touching an icon on a display, pushing a particular button, or by saying a speech command such as “web search”
25 or “text-by-voice.” The second step **226** involves speaking the phrase to be converted to text, which can be referred to as speech-enabled typing. The nature of the phrase to be converted to text depends on the user intent. The type(s) of phrases to be converted include, but are not limited to, text messages **211**, search string entries **214**, target destinations **217**, or brief announcements **220**, as well as other phrases **223**, depending on the intent **225**. The recognized phrase is played through
30 audio (text-to-speech, for example) and the user then decides how to manage the result **227**. Step three **227**, or the management of the result, can entail such actions as saying the name of a target text recipient **212**, glancing **215** at search results such as a weather report on a display, touching a

displayed destination to enter **218** the destination into a navigation system, or speaking a group **221** name for a social networking communication. It is noted that steps one and three can involve input modalities other than speech, but step two entails speech-enabled typing. A key to the present invention is the simplicity of a single user interface and method that can be applied across a variety of different applications. The simplicity of the resultant user interface is highly appealing under driving conditions because very little cognitive processing is required by a driver to learn and use many applications. Because there are so few steps, task completion is fast and distraction is thereby minimized.

FIG. 6 is a process flow diagram of a synchronous speech recognition approach. The user starts **300** and experiences an IVR prompt **301** and, typically, utters a verbal response. The recognition engine **302** processes the verbal response and, based upon matching scores that are referred to as recognition confidence levels **303**, either moves on to the next prompt after processing is completed within the enterprise back-end **304** or re-prompts **301**. When all of the prompting steps are deemed successful, the interactive process ends. The potential issue with a synchronous approach is that the user can get stuck in an error loop when successive low confidence levels (**303**→**low**) occur. Those experienced in the science of automatic speech recognition attribute unexpected audio input as a major cause of recognition errors, even though humans can typically understand such unexpected audio input, hence the evolution of human-assisted speech recognition. Thus, synchronous speech recognition solutions often are associated with poor user experiences. For example, a conventional interactive voice response system (IVR) typically includes error handling dialogs that increase the duration of the interactive process and, often, cause user frustration when recognition failures occur.

As depicted in **FIG. 7**, for asynchronous speech recognition, the user starts **310** and experiences an IVR prompt **312**. The IVR captures the user utterance, transfers the audio to a speech recognition engine **313** that can be queued, and executes the next prompt **312** (if any remain) in the series. Processing **315** of the user utterances occurs in parallel to the prompting **312** for user input; that is, the two activities are asynchronous. As a result, the user prompting **311** process will not be interrupted due to low recognition confidences scores **314** or excessive recognition latencies. As shown in **FIG. 7**, low confidence utterances can be transcribed by a human **316**, thereby assuring high accuracy, but at a cost that is greater than fully automated speech recognition. For asynchronous speech recognition as performed by the instant invention, prompting is a forward moving process whether a valid recognition result is obtained or not. The potential issue of a user

getting stuck in a prompting error loop **312** is eliminated and there is some guarantee of a good user experience. Those experienced in the science of automatic speech recognition attribute unexpected audio input as a major cause of recognition errors. Involving humans within the systems and processes of the invention allow these errors to disappear because those humans can usually still
5 transcribe such “infected” audio. Thus, human-assisted speech recognition employed by the inventive methods and systems is very practical when combined with the asynchronous speech recognition solutions. If the system detects silence on the user side (i.e., no utterance is spoken), then prompting could end early by design, the assumption being that the user is not participating in the automated dialog. For purely asynchronous speech recognition, the dialog is always forward
10 moving when the user cooperates, which has a benefit of preventing the user from repeating spoken utterances. It is noted that a portion of the dialog can be synchronous and a portion can be asynchronous. In fact, for some applications, a portion of the dialog may be required to be synchronous to, perhaps, allow for a user-requested repetition of a phrase (a scenario in which a user is prompted with “Did you say <_____>? Please say yes or no.”) More importantly, certain
15 prompting may depend on a recognition result thereby implying the need for synchronous speech recognition in a particular circumstance. The approach described here provides a compelling and reliable user interface that is safe to use and reliable, even while driving a vehicle.

The inventive systems and methods can be purely synchronous, purely asynchronous, or a combination of both. Conventional speech applications utilize prompting schemes within which, for
20 each prompt, prompting is continued after a recognition result is obtained. Certain applications must be implemented with limits on latency between the time an utterance is finished being spoken and the time the recognition result is utilized (such as dialing a phone number by voice); these applications generally require a synchronous approach. However, certain applications can be implemented with less stringent limits on latency between the time an utterance is finished being
25 spoken and the time the recognition result is utilized (for example, a text message can be sent several minutes after a driver has spoken a text message); these applications generally require a synchronous approach, but can tolerate asynchronous speech recognition for part of the dialog. For example, a driver may request to send a text message (intent); the user is prompted and speaks the text message (which could be recognized asynchronously); the user is prompted and speaks the name of the text
30 message recipient, which is recognized synchronously, or asynchronously; the test message is sent after all recognition results are determined. Some applications, such as form-filling, can be completely asynchronous. Form-filling applications can include, for example, capturing a user name,

address, credit card number, and service selection; the form can be filled out with text after the recognition results are determined, perhaps hours after the user dialog is complete. As a further example, part of a form-filling dialog can include having a user describe something like an automobile accident where an application simply records the description for subsequent recognition, possible though human-assisted speech recognition.

FIG. 8 illustrates a diagram of a method for providing dynamic interactive voice recognition (IVR) over a wireless network. At item 805, a connection with a telematics control unit is established via the wireless network. At item 810, directed dialog application is configured to provide IVR for use with expected spoken user commands. The directed dialog application may be implemented at either the vehicle via TCU 34 or ERDC 48 via interactive automation server cluster 41. At item 815, an open dialog application separate from the remote data center is used to provide IVR for use with unexpected spoken user commands.

The inventive hybrid dialog employs systems to accommodate Power and Novice User Modes simultaneously. The system is either listening for a user to speak a specific utterance, e.g., an expected spoken user command from a limited set of grammar items (e.g., a limited vocabulary mode) or listening and trying to interpret anything it hears (unlimited vocabulary mode or unexpected spoken user command mode). Technically, the unlimited vocabulary mode does have boundaries, in that only certain domains are covered (e.g., destinations, song names, dictation). In the limited vocabulary mode, the system is very capable of detecting that a user has spoken an invalid command (e.g., out-of-vocabulary).

In one exemplary embodiment, the remote data center switches back and forth between the directed dialog application and the open dialog application in accordance with pre-defined criteria. In one exemplary embodiment, the pre-defined criteria may be a confidence score as described above. The present hybrid system moves back and forth between a structured/directed dialog mode, e.g., power mode, and an unstructured/open dialog mode, e.g., novice mode.

The directed dialog application presents questions that correspond to a limited subset of possible spoken user commands. This means that the user is asked questions having a small subset of possible valid responses. When the user's utterances, e.g., spoken user commands, are incorrect or invalid, the system recognizes this as an indication that the user needs help. At this point, the system transitions into a teaching mode.

A teaching mode, e.g., learning mode, of the directed dialog application is entered before using the open dialog application when an unexpected spoken user command is received. The

directed dialog application seamlessly switches to the learning mode, in which the user is prompted with exactly what to say (e.g., “Please say traffic, weather, or navigation”) as opposed to prompting the user to say the name of a service. When an unexpected (i.e., non-matching) utterance or command is received, the teaching mode does not try to figure out what was said by the user. Instead, the system utilizes a process that tries to make it easier for the user to say a valid response. In attempting to obtain a valid response from the user, the teaching mode of the system provides a small or reduced subset of possible choices. If the user has thirty possible valid responses, e.g., a limited subset of possible spoken user commands, in a directed dialog but does not give a valid response, the system provides a second level of choices that is much smaller than the thirty possible valid responses. The system is configured to give the user broad categories, e.g., a reduced subset, from which to choose a valid command. For example, the system says “please say traffic, new destination, or gas prices.” In this example, each of these three categories broadly describe a number of different possible valid responses within the set of thirty. The traffic sub-category, if selected, could provide six different choices. Likewise, the destination sub-category, if selected, could provide five different choices and the gas sub-category, if selected, could provide four different choices. As such, instead of listing thirty items to the user, which is unworkable, only three are listed for selection, as such, the reduced subset is less than the limited subset of possible spoken user commands.

In one exemplary embodiment, switching to the open dialog application occurs when a further unexpected spoken user command is received. In one exemplary embodiment, switching to the open dialog application occurs absent a selection of one of the reduced subsets of possible choices.

In response to a user selection, a further subset of possible choices can be presented. If a sub-category is not appropriate, the user can be prompted for “more choices,” e.g., a further subset of possible choices. When one is selected, the small, manageable subset is uttered to the user. In one embodiment, switching to the open dialog application occurs absent a selection of the further subset of possible choices.

Resources for the open dialog application may be provided by a speech recognition service cloud. The speech recognition service cloud can provide different speech recognition systems in parallel. The different speech recognition systems may be a navigation speech recognition system, a dictation speech recognition system, an audio information recognition system, and a human assisted speech recognition system.

In one exemplary embodiment, an unexpected user command comprising natural user language is received. A user may utter an unexpected user command asking, for example, for “best directions based on traffic.” The open dialog application provided by the speech recognition service cloud is used to provide information based on the natural user language. In one exemplary embodiment, the information provided is a list of possible spoken user commands. In one exemplary embodiment, suggested user commands can be provided in response to receiving an invalid spoken user command. In one exemplary embodiment, the present hybrid system treats “dynamic grammars”, i.e., open dialog or natural user language, the same way as “limited grammars”, e.g., limited vocabulary, so that on-the-fly vocabularies accommodate the novice and power users. In the unlimited vocabulary mode, e.g. open dialog mode, invalid utterances are not readily detected and the user is asked to confirm a recognition result if a low confidence score occurs, or they are re-prompted.

FIG. 9 discloses a method 900 for minimizing task completion time using dynamic interactive voice recognition (IVR) over a wireless network.

At item 905, a connection is established with a telematics control unit via the wireless network. At item 910, a directed dialog application is configured to provide IVR for use with expected spoken user commands. The directed dialog application may be implemented at either the vehicle via TCU 34 or ERDC 48 via interactive automation server cluster 41. At item 915, one of a plurality of expected spoken user commands is received. Depending on the user command received, at item 920, a user is prompted for further spoken information in order to complete one or more actions required by the received user command. At item 925, the one or more actions is completed upon receipt of the further spoken information.

The hybrid dialog minimizes the number of steps in the dialog to minimize task completion time. Task completion time from start to finish becomes very short and, therefore, very safe. The user is prompted to speak a minimal number of times for any given task. For example, if a user wants to send a message to someone, the hybrid design prompts the user to speak the message first, and follows this query by asking to whom the message is to be sent. If the user utters a valid recipient (e.g., from the user’s contact list), then it is assumed that the message was recognized correctly. In addition, the system also recognizes that the user wants to send the message and that the user wants to text a specific contact. This recognition by the system minimizes the number of steps in the dialog and, in addition, minimizes time spent by the user on the activity. Returning to the text message example, once a user has determined that a text message is to be sent, the user will

only need to complete two steps: 1) uttering the message for inclusion in the text message; and 2) uttering the contact(s) to whom the message is to be sent. If the system receives an answer that is not expected, a learning mode can be entered in order to provide other suggestions. For example, if the user mistakenly utters “Mary Smithers” as a contact instead of “Mary Smith”, the system may ask the user to try again or select another contact.

In one exemplary embodiment, the received expected spoken user command comprises a shortcut. The user is enabled to use verbal shortcuts by speaking commands that are specific to a particular service, such as saying “new destination” instead of saying “navigation” first. If a user utters an incorrect shortcut command, the system may enter a learning mode in order to teach the user the proper shortcut to use.

Although specific embodiments of the invention have been disclosed, those having ordinary skill in the art will understand that changes can be made to the embodiments without departing from the spirit and scope of the invention. The scope of the invention is not to be restricted, therefore, to the embodiments described, and it is intended that the appended claims cover any and all such applications, modifications, and embodiments within the scope of the present invention. The speech recognition systems and methods and the in-vehicle user interface and processes that minimize driver cognition described according to the present invention have been applied to a vehicle example. The above-described embodiments, however, should be regarded as illustrative rather than restrictive. The invention should not be construed as being limited to these particular embodiments discussed above. Additional variations of the embodiments discussed above will be appreciated by those skilled in the art as well as for applications, unrelated to vehicles, which require minimizing driver cognitive actions.

AMENDED CLAIMS

received by the International Bureau on 25 October 2012 (25.10.2012)

1. A method for providing dynamic interactive voice recognition (IVR) over a wireless network, which comprises:
- 5 establishing a connection with a telematics control unit via the wireless network;
- configuring a directed dialog application of at least one of a remote data center and a vehicle to provide IVR for use with spoken user commands; and
- 10 using an open dialog application separate from the remote data center to provide IVR for use with natural user language.
2. The method according to claim 1, which further comprises switching back and forth between the directed dialog application and the open dialog application in accordance with pre-defined
- 15 criteria.
3. The method according to claim 2, which further comprises executing the directed dialog application to present questions corresponding to a limited subset of possible spoken user commands.
- 20
4. The method according to claim 3, which further comprises entering a teaching mode of the directed dialog application before using the open dialog application when an out of grammar spoken user command is received.
- 25
5. The method according to claim 4, which further comprises presenting a reduced subset of possible choices in the teaching mode in order to obtain a valid spoken user command.
6. The method according to claim 5, wherein the reduced subset is less than the limited subset.
- 30
7. The method according to claim 5, which further comprises presenting a further subset of possible choices in response to a user selection.

8. The method according to claim 7, which further comprises switching to the open dialog application absent a selection of the further subset of possible choices.
9. The method according to claim 5, which further comprises switching to the open dialog application when a further out of grammar spoken user command is received.
10. The method according to claim 5, which further comprises switching to the open dialog application absent a selection of one of the reduced subset of possible choices.
11. The method according to claim 1, which further comprises providing resources for the open dialog application with a speech recognition service cloud.
12. The method according to claim 11, wherein the speech recognition service cloud provides different speech recognition systems in parallel.
13. The method according to claim 12, wherein the different speech recognition systems are selected from a group consisting of at least two of a navigation speech recognition system, a dictation speech recognition system, an audio information recognition system, and a human assisted speech recognition system.
14. The method according to claim 11, which further comprises:
receiving the natural user language; and
using the open dialog application provided by the speech recognition service cloud to provide information based on the natural user language.
15. The method according to claim 14, wherein the information comprises possible spoken user commands.

16. The method according to claim 14, which further comprises providing suggested user commands in response to receiving an invalid spoken user command.

17. A method for minimizing task completion time using dynamic interactive voice recognition
5 (IVR) over a wireless network, which comprises:

establishing a connection with a telematics control unit via the wireless network;

10 configuring a directed dialog application of at least one of a remote data center and a vehicle to provide IVR for use with spoken user commands;

receiving one of a plurality of spoken user commands;

15 minimizing a number of steps in a dialog by prompting a user for further spoken information;

and

completing one or more actions required by the received user command upon receipt of the further spoken information.

20 18. The method according to claim 17, wherein the expected spoken user command received comprises a shortcut.

19. A system for providing dynamic interactive voice recognition (IVR) over a wireless
25 network, comprising:

a remote data center comprising,

a communications device operable to establish a connection with a telematics control unit
30 via the wireless network;

a directed dialog application operable to provide IVR for use with spoken user commands;
and

a communications subsystem; and

an open dialog application separate from the remote data center, communicatively connected to the remote data center through the communications subsystem, and operable to provide IVR for
5 use with natural user language.

20. The system according to claim 19, wherein the remote data center is configured to use resources for the open dialog application that are provided by a speech recognition service cloud.

10

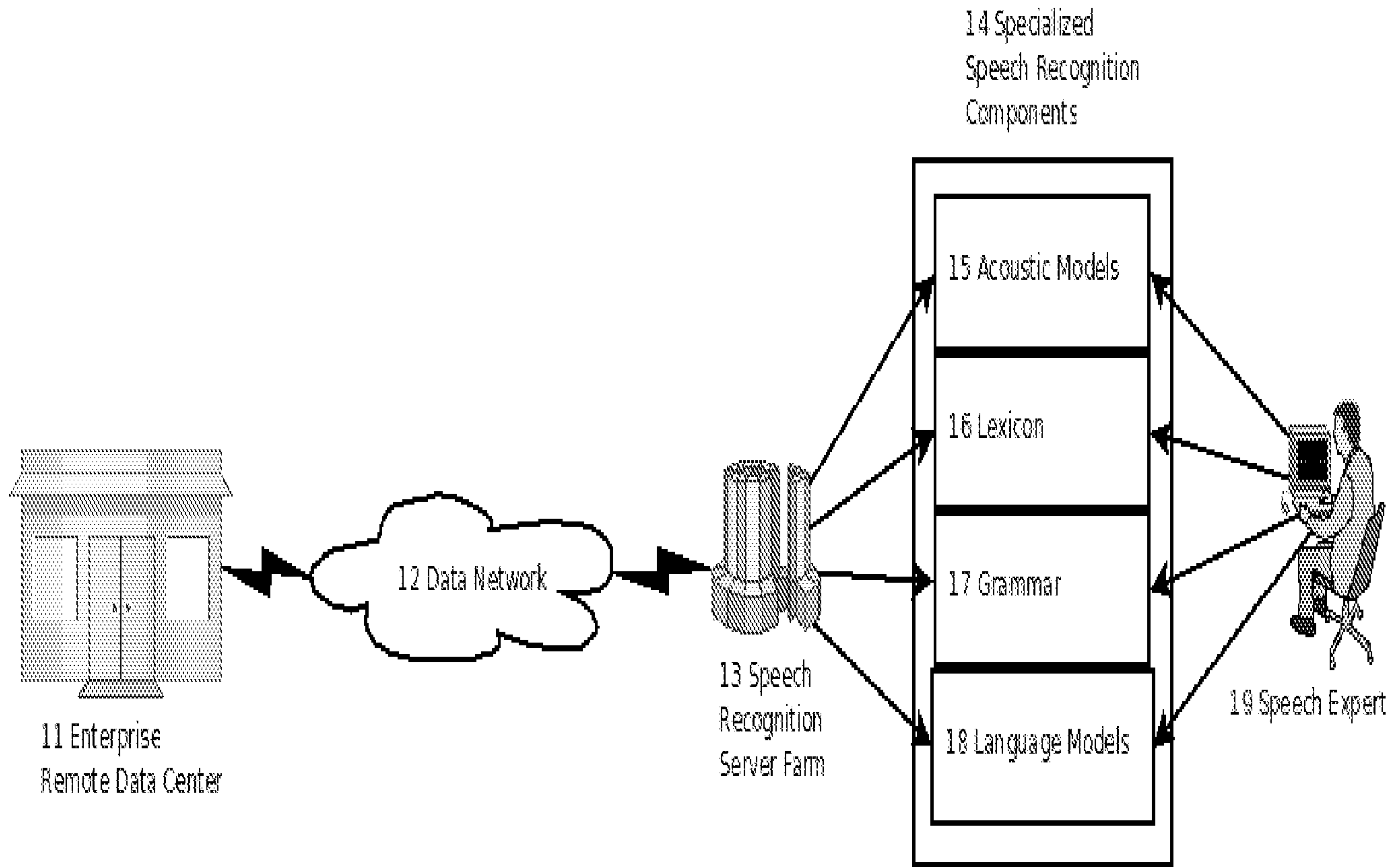


FIG. 1

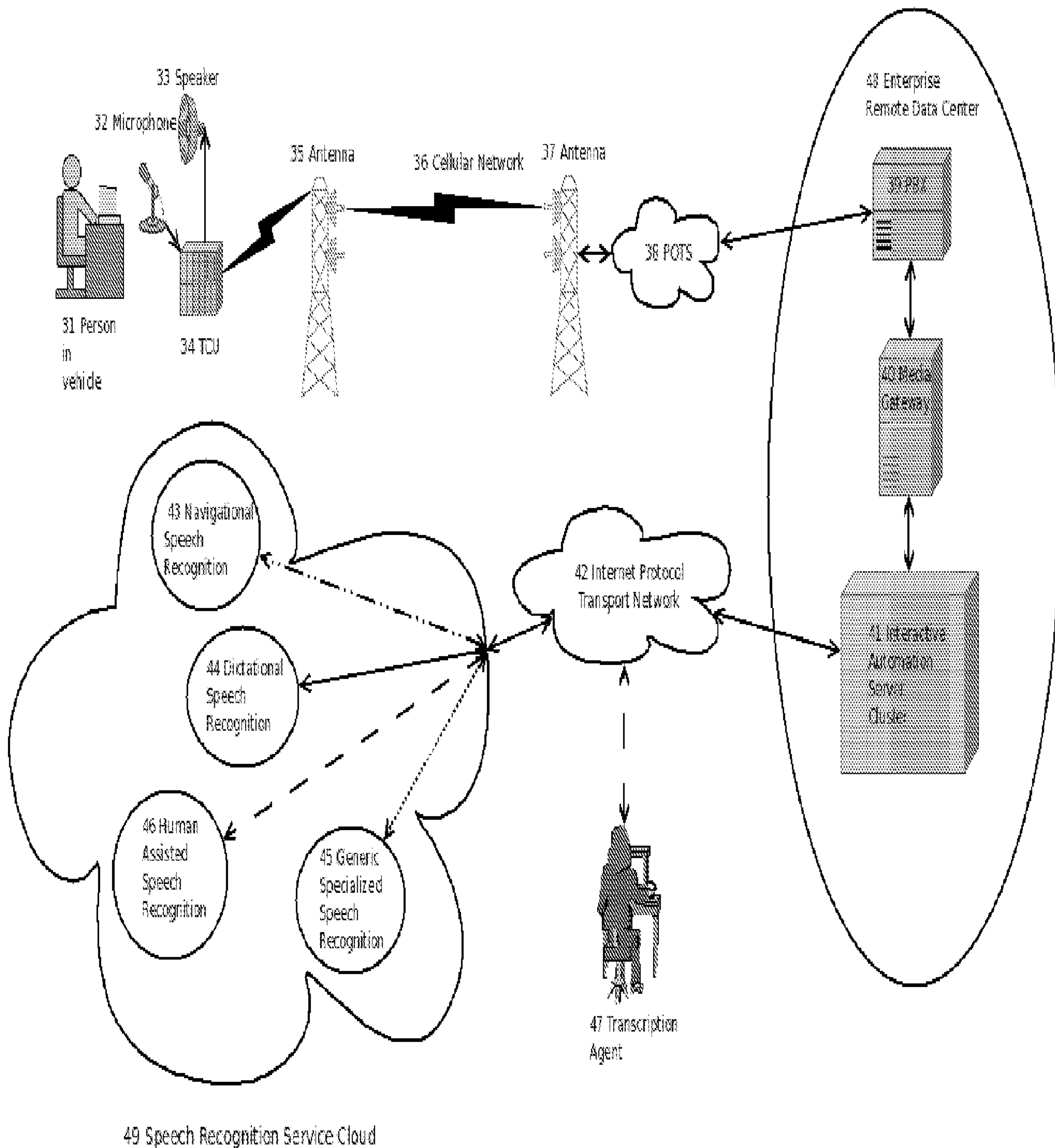


FIG. 2

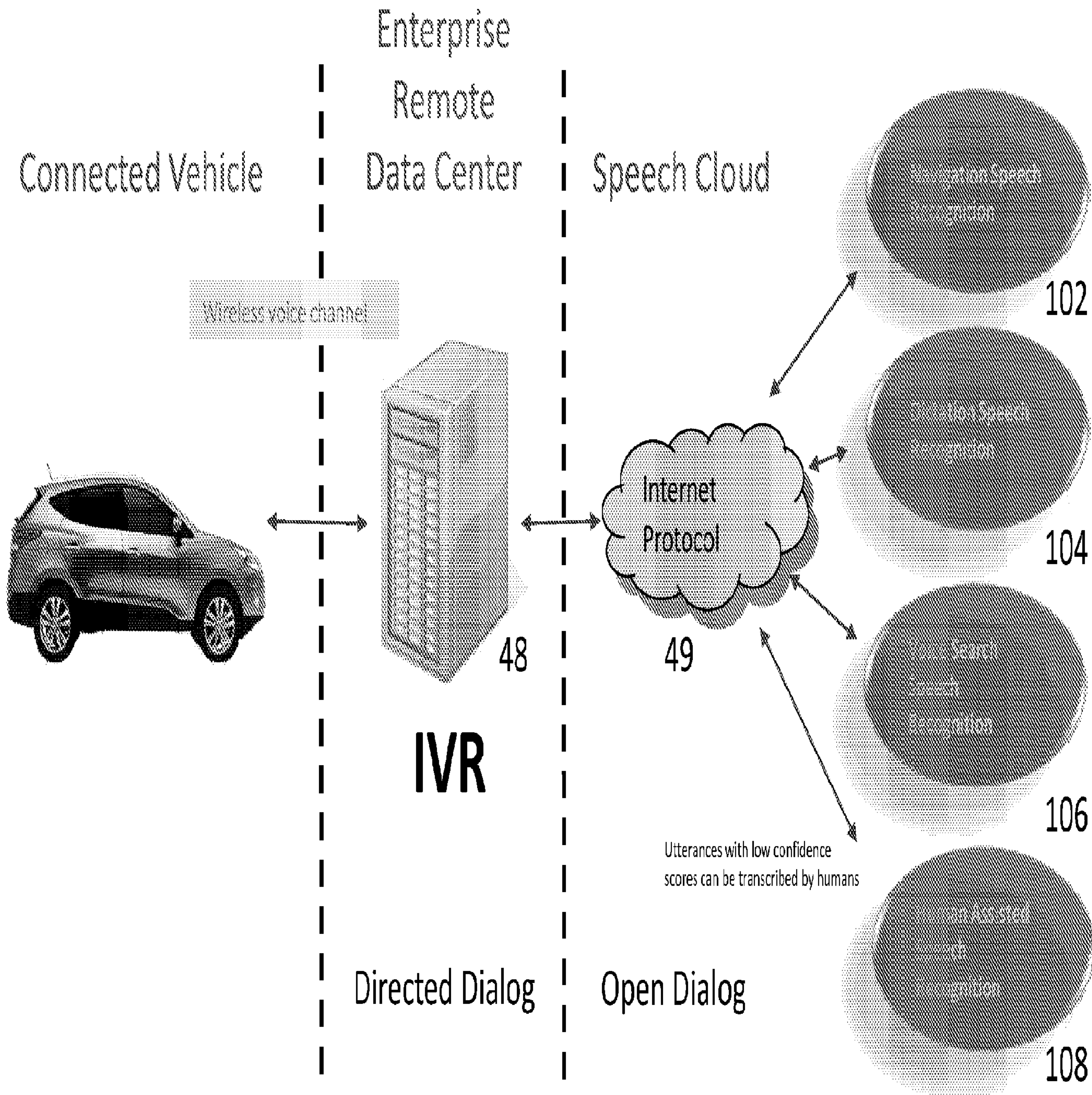


FIG. 3

Usability strategy for IVR

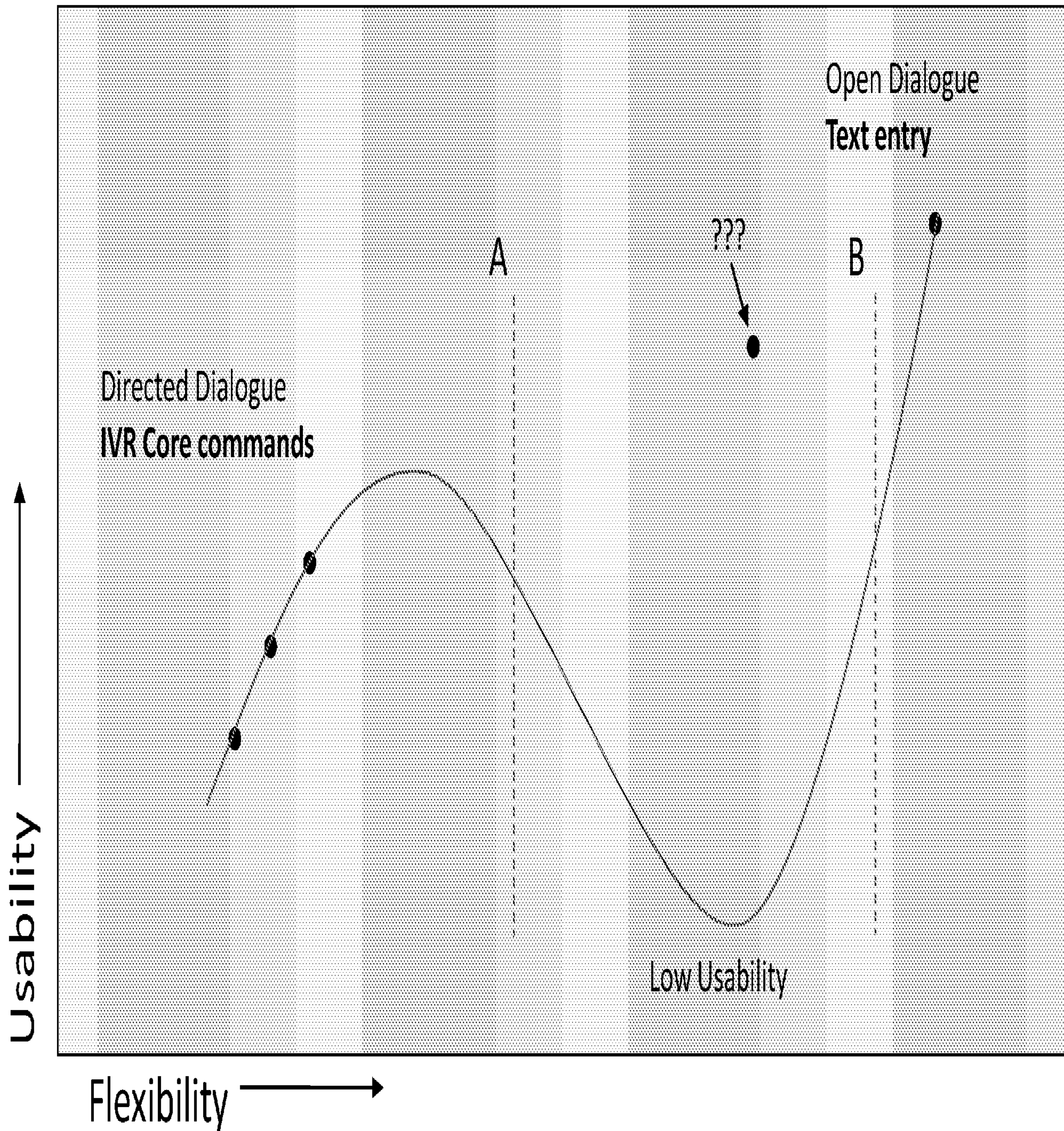


FIG. 4

Multimodal UI Commonalities

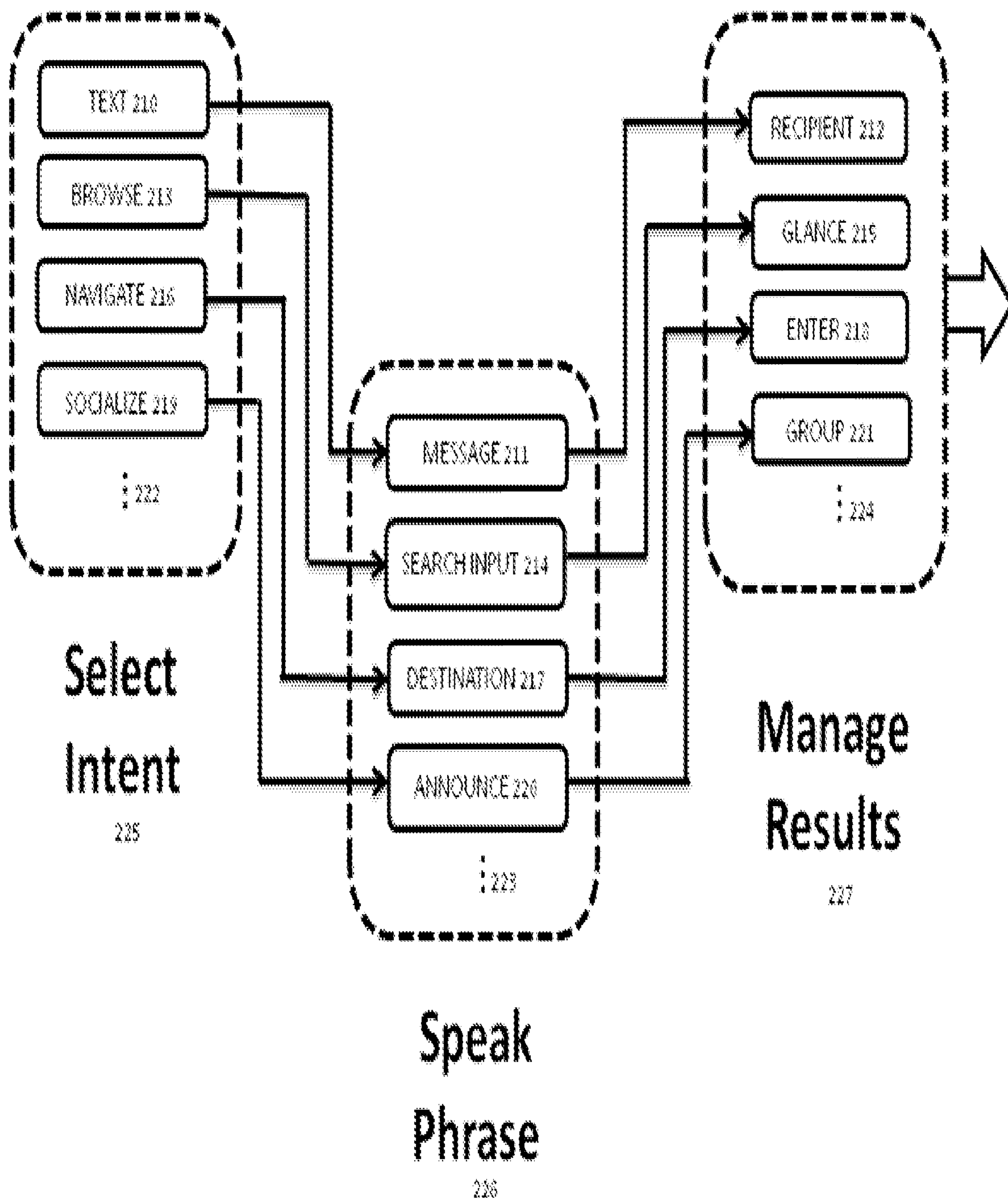


FIG. 5

Synchronous speech recognition approach

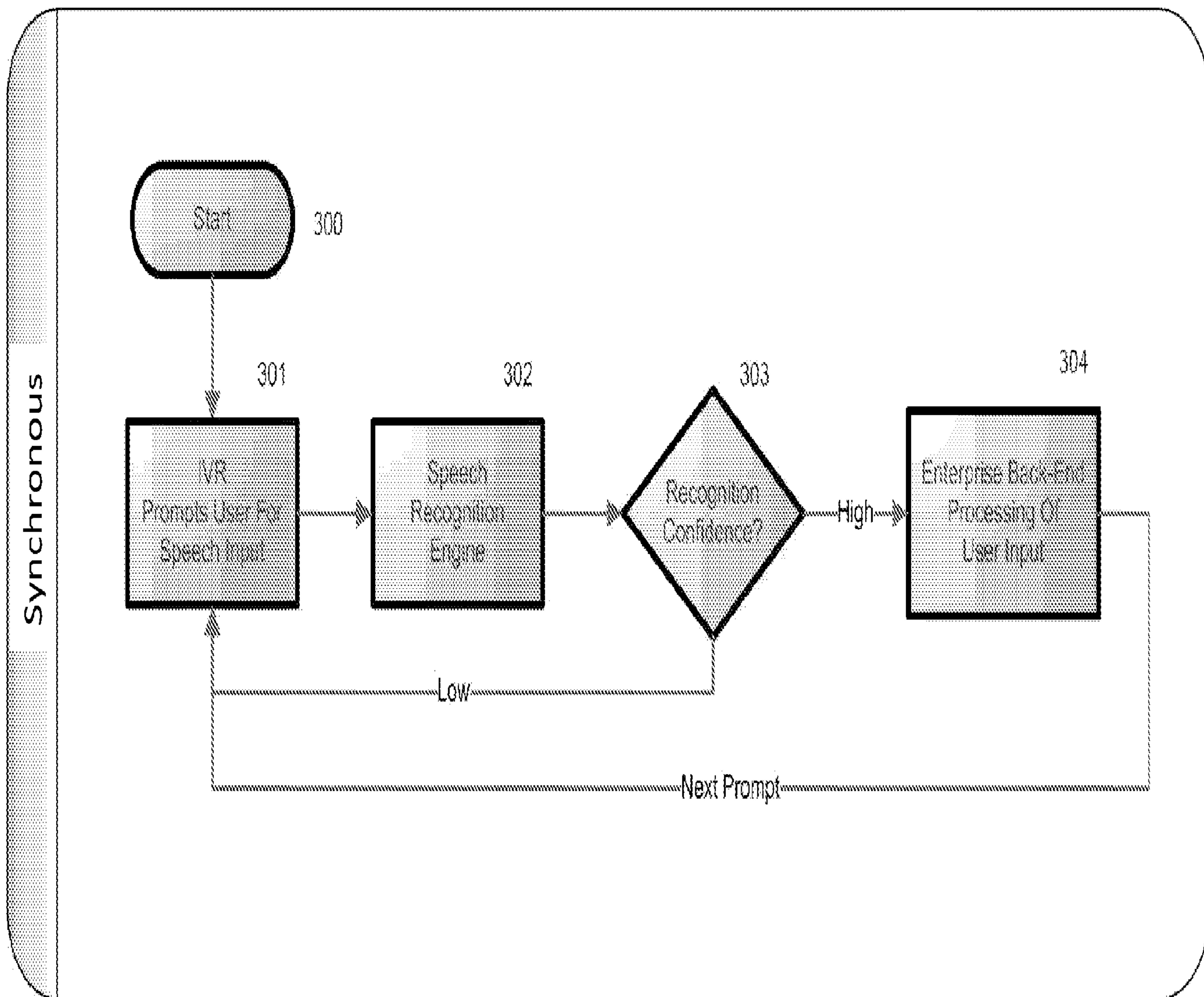
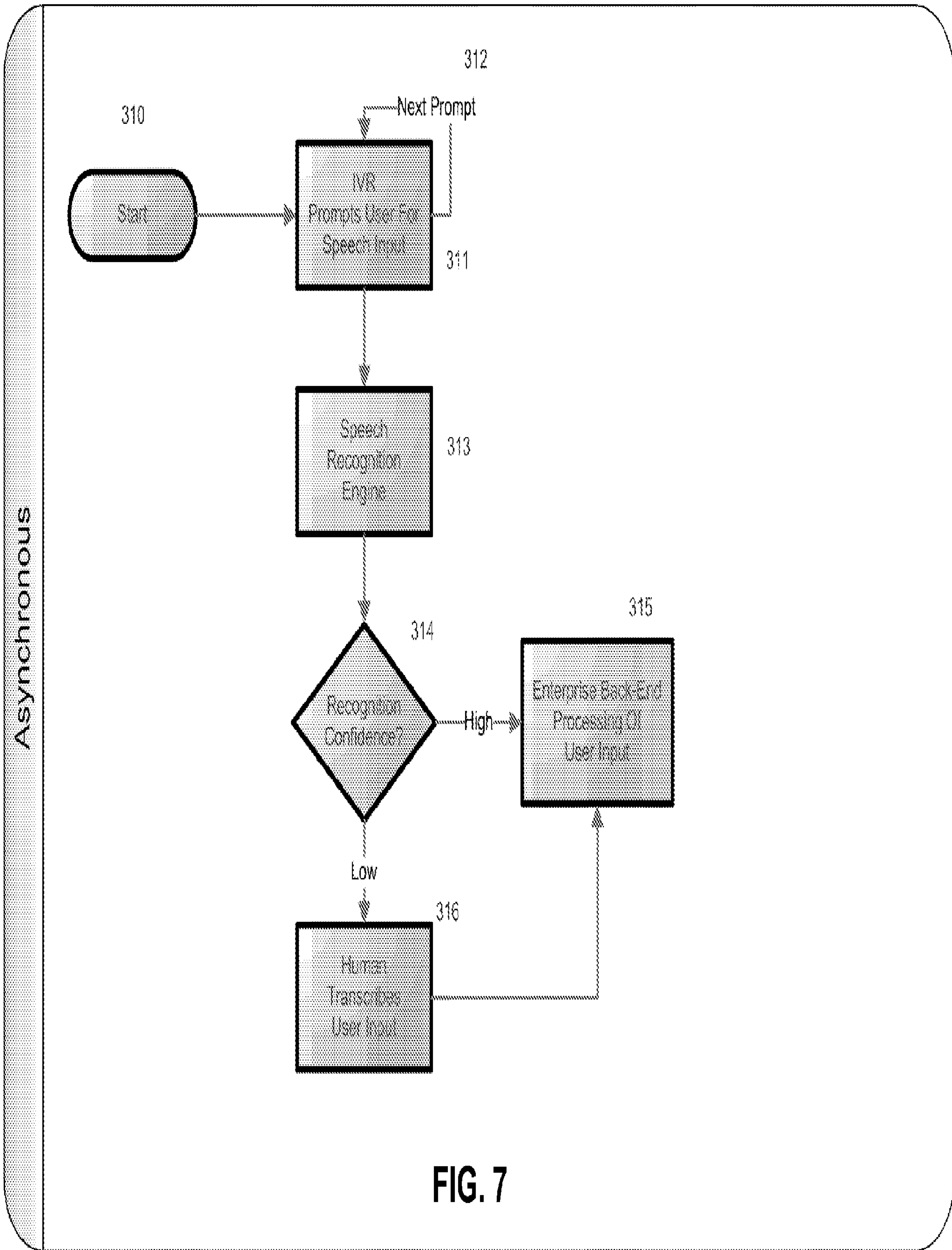


FIG. 6

Asynchronous speech recognition approach



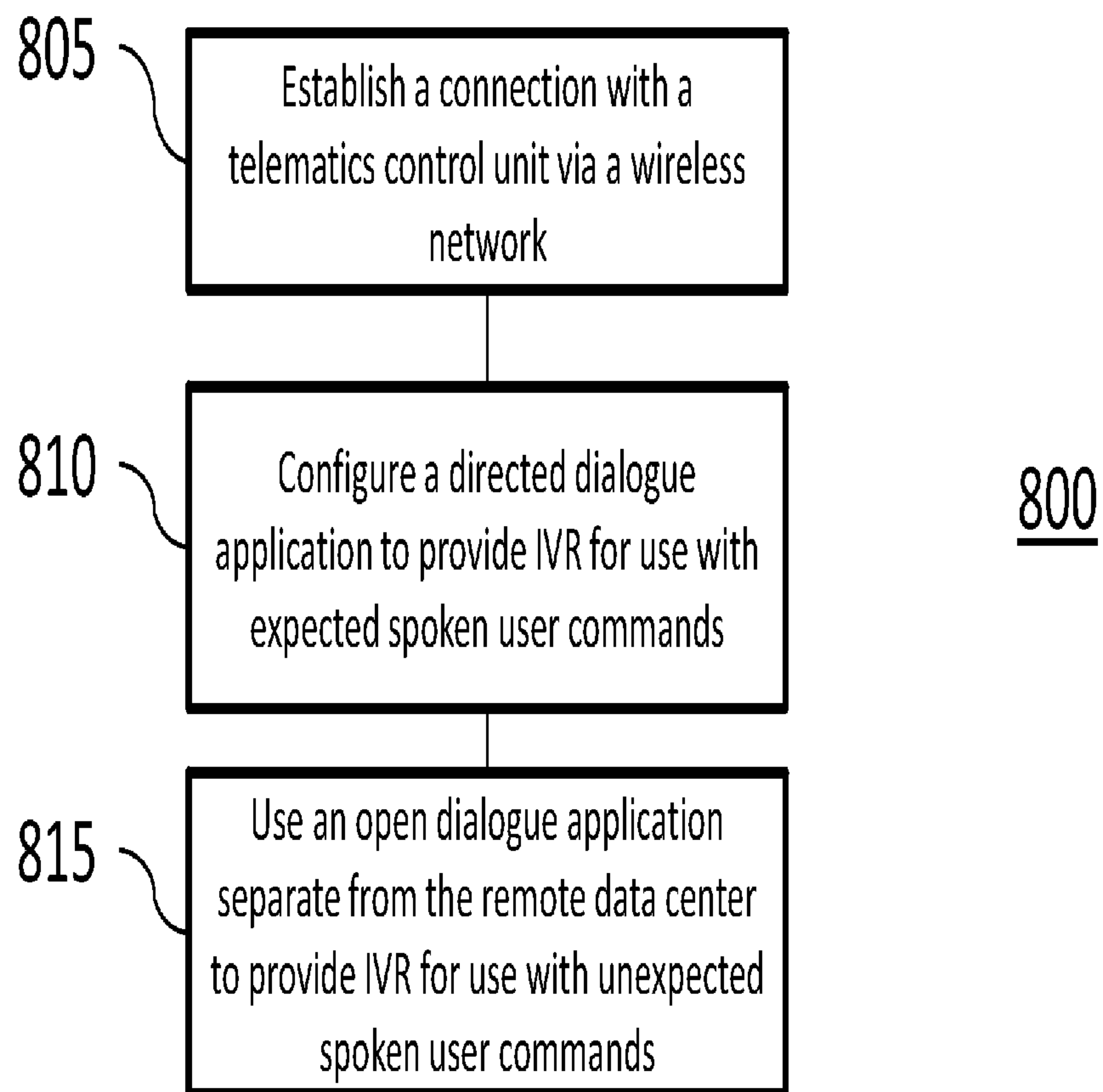


FIG. 8

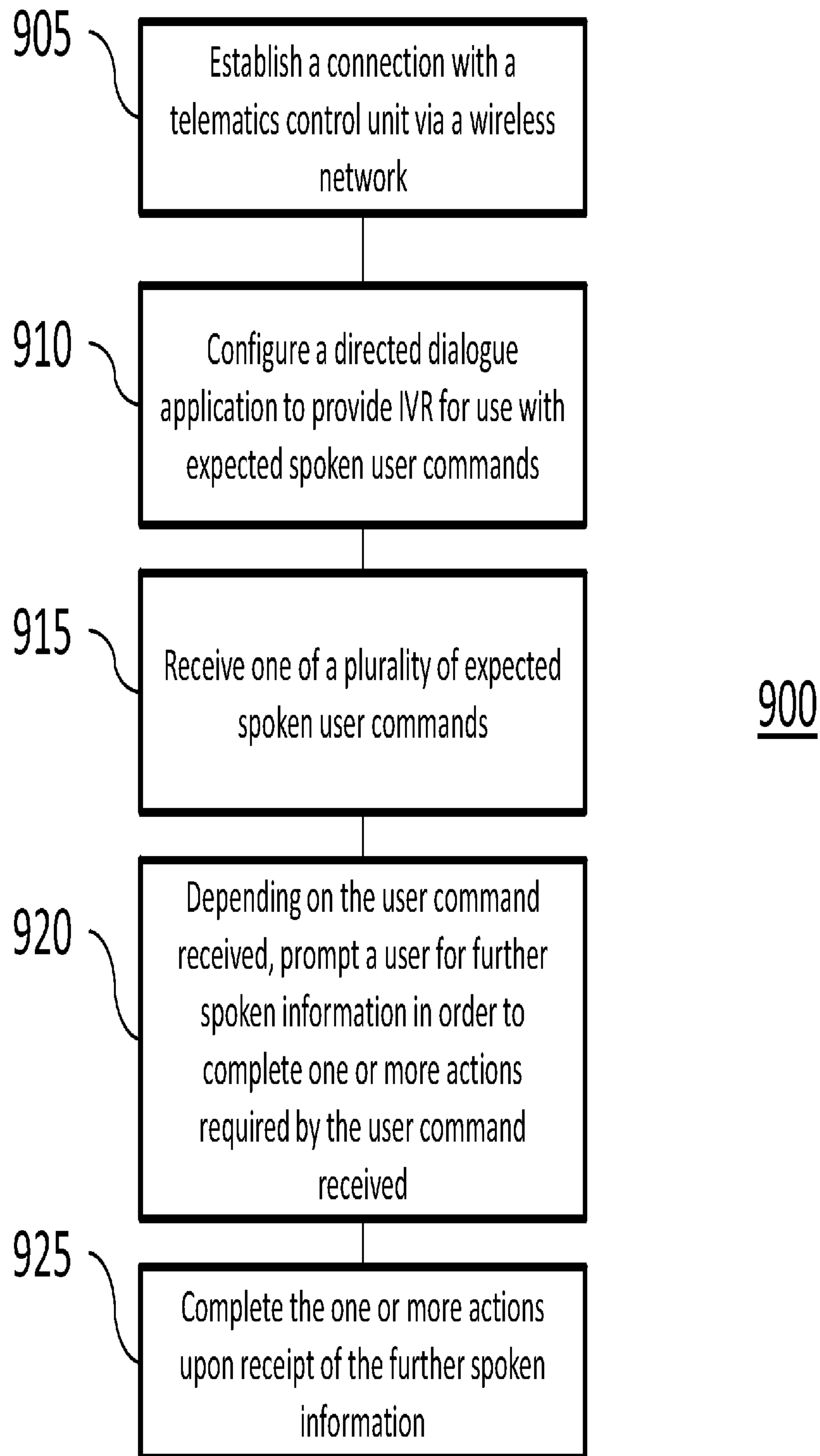


FIG. 9

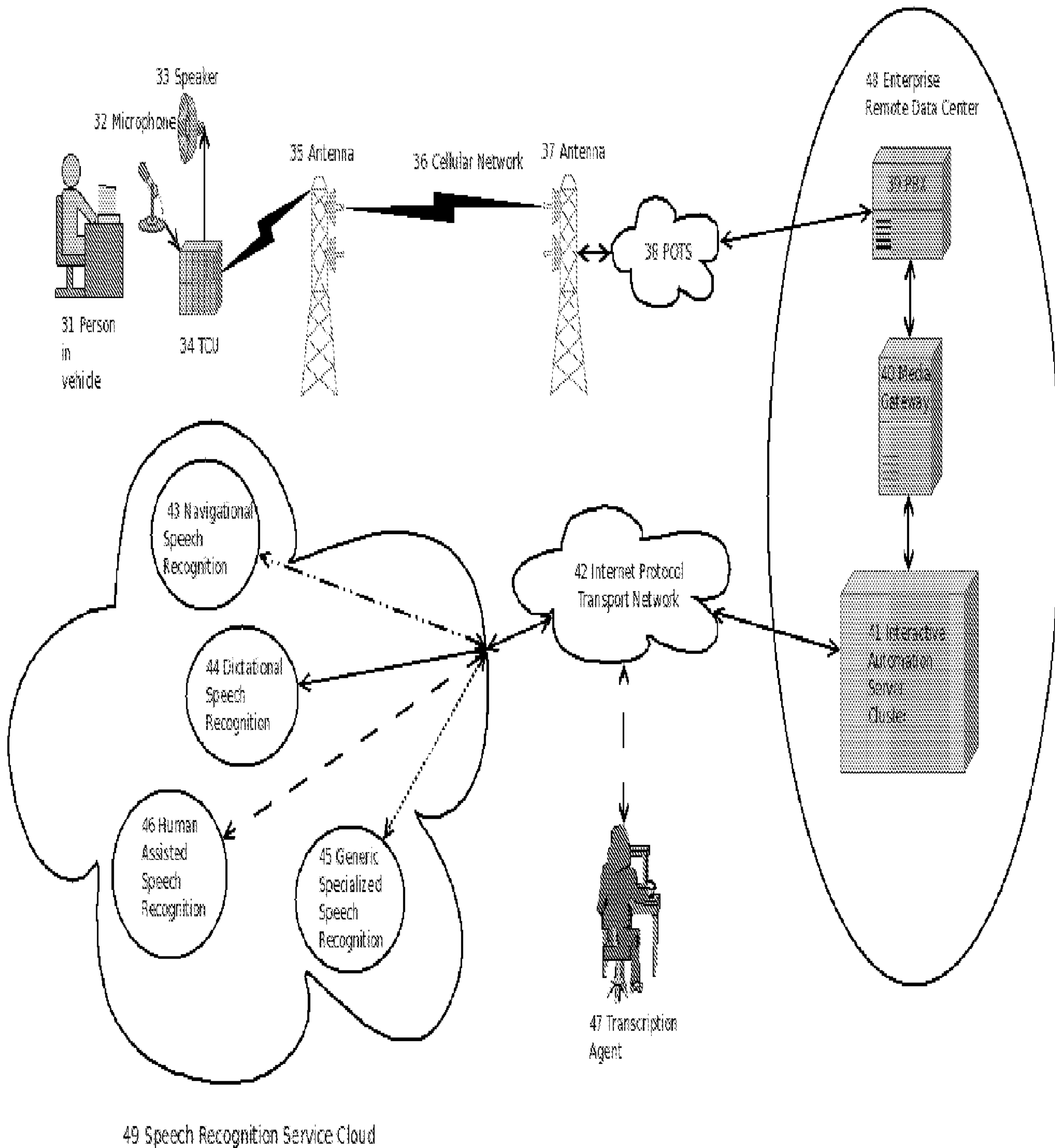


FIG. 2