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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2002 (27.12.2002)

PCT

(10) International Publication Number
WO 02/103675 A1

(51) International Patent Classification[7]: G10L 15/00

(21) International Application Number: PCT/CN01/01030

(22) International Filing Date: 19 June 2001 (19.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicants *(for all designated States except US)*: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052 (US). **INTEL CHINA LTD.** [CN/CN]; Beijing Kerry Center, 6/F, North Tower, 1 Guanghua Road, Chaoyang District 100020 (CN).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ZHAO, Qingwei** [CN/CN]; Room 205, Building 938, ZhongGuanCun, Haidian District, Beijing 100086 (CN). **ZHANG, Xiangdong** [CN/CN]; #6 Floor, North Office Tower, 06-01 Beijing Kerry Center, 1 Guanghua Road, Chaoyang District, Beijing 100020 (CN). **YAN, Yonghong** [CN/CN]; 20756 NW Amber View Lane, Beaverton, OR 97006 (CN). **YUAN, Baosheng** [SG/SG]; BLK515, 52 Jurong West Street, #08-15, Singapore S(640515) (SG).

(74) Agent: **CCPIT PATENT AND TRADEMARK LAW OFFICE**; 10/F, Ocean Plaza, 158 Fuxingmennei Street, Beijing 100031 (CN).

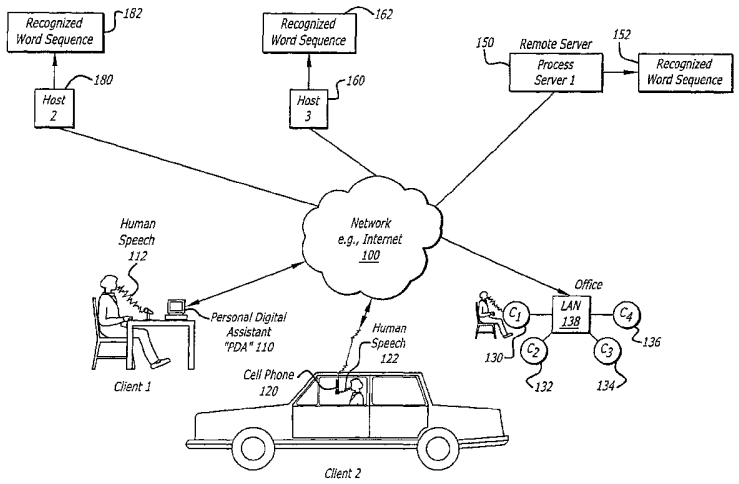
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

[Continued on next page]

(54) Title: CLIENT-SERVER BASED DISTRIBUTED SPEECH RECOGNITION SYSTEM ARCHITECTURE



(57) Abstract: In general, the new client-server based Distributed Speech Recognition (DSR) system provides an effective method of recognizing speech made by a human at a client device and transmitted to a remote server over a network. The system distributes the speech recognition process between the client and the server so that a speaker-dependent language model may be utilized yielding higher accuracy as compared to the tradition DSR systems. Accordingly, the client device is configured to generate a phonetic word graph by performing acoustic recognition using an acoustic model that is trained by the same end-user whose speech is to be recognized. The resulting phonetic word graph is transmitted to the server which will handle the language processing and generate a recognized word sequence. When compared to a design that uses the traditional DSR, the new DSR method and system produces a word error rate that is at least 2-3 times lower, resulting in a higher accuracy recognition system.

WO 02/103675 A1

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

CLIENT-SERVER BASED DISTRIBUTED SPEECH RECOGNITION SYSTEM ARCHITECTURE

FIELD

[1001] The invention relates to Distributed Speech Recognition (DSR) systems and architectures. More particularly, the invention relates to a new DSR system and method of performing the acoustic processing part of speech recognition at a client device, and the language processing part at a server device.

BACKGROUND

[1002] Ever since the conception of the modern computer, engineers and linguists have worked together to perfect the art of human speech recognition by a machine. A goal of automatic speech recognition is for a system that receives as input human speech, converts it into recognizable form and performs a useful function with the recognized speech.

[1003] Today, there exists a variety of commercial applications for speech recognition technology. Dictation machines, for example, can "listen" to a person dictating a text, and in "real time" deliver the 'heard' text on a monitor. Other applications involve machines capable of receiving and executing control commands by human speech, instead of through a mouse or a keyboard. For example, a person may utter at a computer, "read my e-mails." The application can use speech recognition technology to recognize the word sequence uttered by the speaker. A series of commands to execute the requested tasks may then be issued resulting in the computer reading the person's e-mail.

[1004] Still other applications have been developed for client-server based speech systems and architectures. Traditionally, the task of speech recognition is distributed between a client and a server. For example, mobile phones or Personal Digital Assistants (PDAs) may be used as clients which capture the speech, obtain speech features, and transmit the features to a server located at a central location. The communication could happen over a network such as the internet. Once the speech features are

received by the server, they are then processed for acoustic recognition and language processing for the given human language used.

[1005] More specifically, the human speech is captured at the client side by a device such as a microphone. The speech signal is converted into digital form so it can be analyzed by a digital computer. The digitized signal is passed through a feature extractor module which will extract acoustic features of the speech signal such as energy concentrations at periodic sampling points. The extracted features, are then quantized, e.g., by a mathematical model such as the Mel Frequency Cepstral Coefficients. The quantized features are organized into a data packet for transmit ion to a server.

[1006] The server then receives the data packets containing the quantized features and performs acoustic and linguistic processing to provide a word sequence. The acoustic processing is modeled by a speaker-independent (SI) model since the server serves multiple clients.

[1007] Among the shortcomings of the traditional DSR method is its inability to take advantage of the improved word error rate (WER) afforded by utilizing speaker-dependent (SD) recognition models. The difference between the two models lies in the fact that an SD model, has been trained by a specific person's voice resulting in a lower WER for that specific person. This is because humans from different linguistic backgrounds produce significantly different acoustic signals for the same word. Persons from different regions may have different accents and pronunciations.

[1008] Conversely when the system is used by a variety of speakers, such as an Automatic Teller Machine, a SI model is utilized and it is designed to handle any speaker regardless of the speaker's speech characteristics such as the speaker's pronunciation, speech variations due to gender and age and the robustness of the speaker's sound. Not surprisingly, SD models deliver 2-3 times lower WER than SI models. Because the traditional DSR method handles acoustic processing at the server, not at the client, it is impractical and inefficient to allow system architects to employ SD acoustic recognition models with lower WER to improve overall recognition accuracy.

BRIEF DESCRIPTION OF THE DRAWINGS

[1009] Figure 1 illustrates a block diagram of an exemplary communications network applying the new DSR methodology pursuant to an embodiment of the invention.

[1010] Figure 2 illustrates a block diagram of an exemplary client-server based DSR system applying the new DSR methodology pursuant to an embodiment of the invention.

[1011] Figure 3 illustrates an exemplary schematic representation of the new DSR methodology pursuant to an embodiment of the invention.

[1012] Figure 4 illustrates an exemplary schematic representation of the new DSR methodology pursuant to an embodiment of the invention at a client node of a client-server based system.

[1013] Figure 5 illustrates an exemplary phonetic word graph generated at a client node according to the methodology pursuant to an embodiment of the invention.

[1014] Figure 6A illustrates an exemplary flow chart of the transmission process of a phonetic word graph according to an embodiment of the invention.

[1015] Figure 6B illustrates an exemplary datagram used to transmit an exemplary phonetic word graph according to the methods pursuant to an embodiment of the invention.

[1016] Figure 7A illustrates an exemplary schematic representation of the new DSR methodology pursuant to a method pursuant to an embodiment of the invention at a server node of a client-server based system.

[1017] Figure 7B illustrates an exemplary phonetic word graph generated at a server node of a client-server network system according to a methodology pursuant to an embodiment of the invention.

[1018] Figure 7C illustrates an exemplary phonetic word graph expanded from the phonetic word graph shown in Figure 5.

[1019] Figure 8 illustrates a block diagram of an exemplary system applying the new DSR methodology pursuant to an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

[1020] In the following detailed description of the invention, numerous specific details are set forth in order to provide a thorough understanding of the invention. However, it will be obvious to one skilled in the art that the methodology pursuant to the invention may be practiced without these specific details. In other instances, well known methods, procedures, components, and circuits have not been described in detail so as not to unnecessarily obscure aspects of the invention.

[1021] The methods pursuant to the invention includes various steps, which will be described below. The steps may be performed by hardware components or may be embodied in machine-executable instructions, which may be used to cause a general-purpose processor programmed with the instructions to perform the steps. Alternatively, the steps may be performed by a combination of hardware and software.

[1022] The invention reveals a new DSR methodology that is different from traditional DSR methods, and thus results in improved recognition accuracy. The new DSR methodology takes advantage of the lower WER associated with SD acoustic recognition models. This is accomplished by dividing and distributing the speech recognition process into acoustic recognition at a client device, and language processing at a server device. Thus, after a speech is captured at a client device, it is processed acoustically in accordance with a SD personalized acoustic model. This process results in an N-best hypothesis as to what was most likely said. Next, a word graph packet is formed and transmitted over a network to a server device. Finally, the server device receives the word graph packet, decodes it, and processes it linguistically resulting in a recognized word sequence.

[1023] Once the human speech is captured at a client device, it will be analyzed acoustically. Acoustic recognition involves extracting features of the captured speech signal and searching an acoustic model for one or more possible matches between the extracted features of the captured speech and known previously recorded speech features stored in the acoustic model. In a preferred embodiment, a phonetic word graph may be used to represent the speech. The acoustic model may be a personalized

SD model which is personally trained by the user, for example the owner of a mobile phone or a PDA. Subsequently, the word graph is packetized and transmitted over a network to a central server. The server may then utilize the selected language model to process the phonetic word graph linguistically and produce a recognized word sequence.

[1024] Referring now to Figure 1, an exemplary DSR system is illustrated. Client 1, Client 2, and C1-C4 are examples of different client devices which employ the new DSR methods pursuant to one embodiment of the invention. Client 1 is a Personal Digital Assistant 110. Client 2 is a mobile phone 120, and C1-C4 are computer terminals serving as nodes in an exemplary LAN 138. In each case, the client devices are configured to capture human speech. For example, human speech 112 with regards to PDA 110, or human speech 122 with regards to the mobile phone 120. The captured signals are then acoustically processed and a resulting data packet containing a sequence of phonetic word graphs are generated at the client device.

[1025] Subsequently, the data packets are transmitted over the network 100 e.g., the internet. With regards to the speech made at the PDA 110, process server 150 receives the data packet (not shown) and after performing linguistic processing on the data contained in the packet, generates the recognized word sequence 152. Similarly, host 180 and host 160 receive data packets from mobile phone 120 (client 2) and C1 130 and generate the recognized word sequences 182 and 162, respectively. Client 1 can train PDA 110 by initially uttering a series of words and training the PDA. Similarly Client 2 can train mobile phone 120 by uttering a series of words and providing the mobile phone with a text of what is said. Once the client device has been trained, it can form a personalized acoustic model which can be used by the client device as a basis for a search and comparison of what has been said.

[1026] Referring now to Figure 2, a block diagram of an exemplary DSR system is shown. The Figure illustrates how a human speech signal 200, spoken at client device 220, is converted into the recognized word sequence 260 at the server device 252. The signal 200 is captured at a client 220 node of a client-server based system and acoustic processing

and recognition 230 is performed at the client device 220. As shown in Figure 2, client 220 may be a computer terminal 210. But, client 220, could also be a mobile phone, a PDA, etc.... In fact, the client device is any device having the capability of receiving human speech signal 200; processing it acoustically and recognizing its phonetic make up; and preparing the resulting phonetic data for transmission over a network 242, such as the communication network 240.

[1027] Referring still to Figure 2, the process server 254, handles the language processing 250 phase of the new DSR system. The process server 254 may be a server computer system 252, capable of receiving phonetic data and analyzing it linguistically in order to arrive at word sequence 260. Once the server 252 has completed the language processing 250, the server 252 generates a recognized word sequence 260, which may then be transmitted back to client 220 over the network 242.

[1028] Referring now to Figure 3, a schematic representation of the new DSR system which is designed according to the methods pursuant to one embodiment of the invention is illustrated. The Figure presents an exemplary sequence of operations that human speech signal 300 is put through when the new DSR methodology is applied.

[1029] At functional block 310, human speech signal 300 is received at the client device 340. For example, a person may speak into a microphone which will capture the human speech signal 300. In an exemplary embodiment, the person may be limited to single word commands to be used as control commands, i.e., an Automatic Speech Recognition (ASR) system. In another embodiment, the system may include Large Vocabulary Continuous Speech Recognition (LVCSR). The methods pursuant to this embodiment apply to both ASR and LVCSR.

[1030] At functional block 310, the captured human speech signal 300 is converted into a digital signal by an analog-to-digital converter. In functional block 320, the features of the resulting digitalized signal are extracted, for example by a feature parameter extraction module 820 (See Figure 8). Functional block 320 may further be further broken down into: end-point detection as shown at functional block 322, pre-emphasizing filtration as

shown in functional block 324, and feature computation as shown at functional block 326. During end-point detection, a determination is made as to the beginning and ending of a speech feature. In other words, a determination is made as to when one feature ends and another feature begins. During pre-emphasizing filtration, the speech signal is filtered in order to amplify the speech signal's important features. Finally, during the feature computation, features of the speech signal are computed to develop a sequence of possible candidates.

[1031] Accordingly, after the speech features have been extracted, acoustic processing of the captured human speech signal takes place at functional block 342. Acoustic processing is a process to provide a match of the speech features identified at functional block 320 with known phonetic units. Thus, acoustic processing comprises receiving a human speech signal and using an acoustic model to recreate a sequence of sounds which most closely represent the input speech. The acoustic model may be organized by sub-word units such as phonetic-level, demi-syllable, or syllable-level units. However, acoustic models using other phonetic units may also be implemented.

[1032] One method to perform acoustic processing is by utilizing Hidden Markov Models (HMMs). HMMs, which are well known in the art, are stochastic finite-state automata which comprise of a Markov chain of acoustic states. These states model the temporal structure of speech, i.e., how the states vary with time. A probabilistic function for each of these states, modeling the emission and observation of acoustic vectors are what is represented by HMMs.

[1033] Once an HMM is used to represent the speech features, a search space is defined and a search may be conducted against a previously formed HMM within an acoustic model. The HMM may be formed during a training phase of a client device which may occur the first time a person uses the client device 340. For example, when a person purchases a mobile phone, the phone may have a button which when pressed may place the phone in training mode. During this mode, the person may be asked to utter words, phonemes, or other units of speech as they appear on the screen. The mobile phone can then capture the sounds produced by

the user and run it through functional blocks 322-326 of Figure 3 in order to extract the features associated with the sounds and form an HMM. Since, during the training phase, the client device 340 knows exactly what words the sounds represents, it is able to store the two pieces of information (the word uttered and its extracted features) and create an acoustic model which is personalized for the owner of the mobile phone.

[1034] By creating a personalized acoustic-phonetic model, the mobile phone can take advantage of an SD acoustic model which has a 2-3 times better WER than an SI acoustic model.

[1035] In functional block 334, an optimization procedure is deployed. The decision on the spoken word may be made using any number of knowledge sources. For example, an acoustic-phonetic model of single phonemes as trained by the owner of client device might be used either alone or in combination with other knowledge sources such as a pronunciation lexicon. However, if the owner is not the person actually using the client device, then the person who does actually use the device should train the device because it is this person's voice characteristics which results in a more accurate recognition process.

[1036] At functional block 336, an N-best hypothesis is determined after a search of the acoustic model is completed. However, instead of an N-best hypothesis, a single-best hypothesis strategy may be utilized. In functional block 338, a phonetic word graph (Pword graph) is generated. The main idea of a pword graph is to come up with phonetic alternatives in regions of speech signal, where the ambiguity about the actually spoken phoneme is high. The expected advantage is that the acoustic recognition process is decoupled from the application of a complex language model. This language model can be applied in subsequent post processing which takes place at a server computer according to the methodology pursuant to this embodiment of the invention. The number of word alternatives is a design parameter which can vary according to the level of ambiguity or accuracy that the user desires.

[1037] Once a Pword graph has been generated at functional block 338, the Pword graph may be packetized and transmitted to the server device. Any transmission medium and any packeting scheme may be used to

transmit the Pword graph to the server. For example, an Internet Protocol datagram may be generated as illustrated at functional block 354 by packetizing the Pword graph onto the datagram. This datagram can then be transmitted over the network 350, as shown at functional block 352. In a preferred embodiment of the invention, network 350 may be the internet, but any other type of networks such as a Local Area Network will suffice.

[1038] In functional block 356, the datagram containing the Pword graph is received by a server and the Pword graph is removed from the datagram. At functional block 382, language processing may take place on the Pword graph. Language processing involves taking the sequence of sounds as organized in a Pword graph and transforming it into an actual word. The received Pword graph is analyzed node-by-node. For each node the dictionary and grammar rules are checked for the specific language model available and chosen by the user. In one embodiment, the client device may have a language selector key allowing the user to speak in English or Chinese or any other language that the system may support. At functional block 390, a true Pword graph is formed according to the Pword graph transmitted by the client device. (See Figure 5). Finally, at functional block 386 a search algorithm is deployed to determine the recognized word sequence by using a dictionary and grammar lexicon.

[1039] Referring now to Figure 4, a block diagram of a client device is shown. The client device may be a variety of portable devices such as a mobile phone, a PDA, a portable computer, or any other device which may be used by a person to communicate with another device located at a different geographical location.

[1040] Once a person decides to communicate with a remote server, the person will have the option of talking into a receiver module, such as a microphone, of a client device. The client device, however, performs a series of operations on the captured human speech signal 400. These operations are shown at functional blocks 420 and 450 in Figure 4. The operations performed on the human speech signal 400 can be generally divided into two types. During the first series of functions indicated at functional block 422, 424, 426, and 428, the human speech signal is put through a process wherein the human speech signal is converted into a

digital signal according to well known methods as is known in the prior art. Subsequently, the digitalized signal at functional block 412, is presented to a feature extraction module which extracts the features present in the human speech signal. These features may represent the energy concentrated in the speech signal as measured periodically and can be represented as a sum of acoustic vectors, for example as is shown at functional block 428, as x1, x2, ...xT. However, other features of the acoustic signal, as is well known in the art, may also be extracted.

[1041] At functional block 450, acoustic vectors x1,x2,.....xT are presented to an acoustic processor whose function may be to recognize the speech that caused the x1, x2,.....xT acoustic vectors. To accomplish this task, the acoustic processor may consult an acoustic model which contains acoustic vectors for a variety of speeches previously made by the person using the client device. This model can be easily trained by the person who will be primarily using the client device. For example, the client device can be programmed or trained when the person first purchases it. The device could have a "train me" switch, which when activated by the user, will flash words on its screen and prompt the user to pronounce them. The device may , for example, flash words, phonemes, syllables, demi-syllables or any other units of words according to a specific design parameter. The choice of the phonetic units does not affect the methodologies pursuant to this embodiment of the invention.

[1042] Thus for example, the device flashes the word, "apple" and the user says, "apple". The device will capture the speech signal that is generated by the speech made by the user for example by a microphone. Those skilled in the art know that this signal will be an analog signal which when viewed on an oscilloscope may look like, speech signal 400. After capturing the signal, the acoustic processor may apply functions at functional blocks 412, 422, 424, 426, and 428 in order to extract the features of the signal generated by the user uttering the word, "apple", resulting in the generation of a set of acoustic vectors. This representation is then stored at a data base together with an indication of the word apple. This process can continue for word after word. The more words presented to the device the more complete of an acoustic model for the user or owner

of the device. Once this model is complete, the device is ready for acoustic recognition which occurs at functional block 450.

[1043] The acoustic processor is now charged with the task of recognizing the uttered speech. It accomplishes this task by conducting a search of the data base containing the trained acoustic model. At functional block 446, a search is done to find one or more matches for the speech. The decision on the spoken word may be done by an optimization procedure. Several searching schemes have been developed and are well known in the art. For example, a beam search strategy with a pruning option may be used. Alternatively, a tree lexicon or a one-pass algorithm may be applied. The choice of the specific search strategy does not influence or alter the methods pursuant to this embodiment of the invention.

[1044] At functional block 442 the data base containing the acoustic-phonetic model is gathered. The training phase of the speech recognition system of this embodiment takes place at this functional block. At functional block 444, a language model is consulted in connection to the search strategy deployed at functional block 446. However, the addition of a language model at the client side may be a design choice. It is not necessary to include a language model in order to implement the methods pursuant to this embodiment.

[1045] The result of the search is generated at functional block 448. Here, an N-best hypothesis is generated. Although, a single best hypothesis could also be used, in a preferred embodiment, an N-best hypothesis yields a higher accuracy since it provides not just a single guess as to what was uttered, but a multiple of guesses. From this information, at functional block 452, a word graph may be generated. The main idea of a word graph is to come up with word alternatives. Word graphs have proven to be effective where there is a need for high accuracy. In effect, a word graph as illustrated in Figure 5, presents words that have similar sounds, or features, or acoustic vectors. This similarity can cause confusion. For example, the words, "duo" and "dao" and "yao" in Mandarin Chinese look nearly identical on a spectrum analyzer. Similarly, with reference to Figure 5, the words, "dai" , "nai" and "mai" are alike but for a single letter or phoneme. These similarities which are prevalent in most languages, can be

further analyzed by use of a grammar lexicon presented in a language model as will be discussed with reference to Figure 7A.

[1046] Referring back to Figure 4, once the word graph representing word alternatives of what has been uttered, is generated, the device may transmit this information as a binary file to a remote server. The word graph can be represented in a datagram as shown in Figure 6B. However, any other form of packetizing of this data may be implemented.

[1047] Referring now to Figure 5, an example of a word graph having a two-level alternative capacity is illustrated. The actually uttered words in this example are in Chinese Mandarin: "wo yao mai zhong ke jian" which means, "I want to buy zhong ke jian (China Science Healthy)" (zhongkejian is the name of a stock in Chinese stock market). This word graph is the output of the acoustic processor as was illustrated in Figure 4, at functional block 452. The acoustic processor, compares the acoustic vectors of what the device captured with the acoustic model and presents the acoustic processor with three alternatives for each word. Words 512, 511, and 510 represent "yao" and its alternatives. Words 514, 515, and 516 represent "mai" and its alternatives. Accordingly, the word graph depicted in Figure 5, may be used in conjunction with a language model which comprises a dictionary lexicon and a grammar lexicon in order to determine the single best sentence represented by the word graph. The application of a language model to the word graph may take place at a server node since the language processing is a rather complex process and is independent of the acoustic recognition process. Thus, the methods pursuant to this embodiment takes advantage of the SD acoustic model by generating a word graph with a two-level word alternative. The word graph will be transmitted to a server which will then complete the recognition process and determine the single best sentence.

[1048] Referring now to Figure 6A, the transmission process according to one embodiment of the invention is illustrated. At functional block 602. a phonetic word graph is generated by the client device. At functional block 604, the word graph is transformed into a binary file, and packetized for transmission over the network. For example, at functional block 604, a TCP/IP datagram is used as a vehicle for transmission. However, any other

method of packetizing the word graph for transmission is possible and the specific choice does not impact the methods pursuant to this embodiment of the invention. At functional block 606, the datagram is transmitted to the server, and at functional block 608, the datagram is received at the server.

[1049] Referring now to Figure 6B, an exemplary Internet Protocol datagram is illustrated. In the header 612 segment of the datagram 600 a client device's logical address and a server device's logical address and any other control information, as is well known in the art, is included. The data area 610 may include a binary representation of the phonetic word graph as generated by the client device.

[1050] Referring now to Figure 7a, an exemplary block diagram of the server node 700 is illustrated. At functional block 710, the TCP/IP datagram as shown in Figure 6B is received by the server 700. At functional block 712, the word graph is decoded from its binary form and a true word graph representation of what was uttered at the corresponding client node (not shown) is formed. At this point the server is equipped with the equivalent of an N-hypothesis representation of the speech. The server may use a language model as at functional block 720, and a dictionary lexicon as shown at functional block 718, to conduct a search and make decisions as to the most likely speech as is well known in the art.

[1051] In this procedure, for each phonetic word graph node (See Figure 7b and 7c) the server 700 finds all those words that correspond to the chosen Phonetic word by checking the dictionary and the grammar lexicon. However, the invention is not limited to the dictionary and grammar model. Any other language model, for example, cache-based language models, trigger-based language models, and long-range trigram language models (lexical zed context-free grammars) may also be used. Regardless of the specific language model used, the result at functional block 720 is a recognized word sequence which can be stored, or may be used as a command to the server 700.

[1052] Referring now to Figure 7b, an exemplary true phonetic word graph is illustrated. The phonetic word graph represents "wo yao mai zhong ke jian" as the uttered word sequence. From this word graph a corresponding word graph can be generated as shown in Figure 7c. In this

procedure, for each phonetic word node (such as "yao") the server searches for those words that sound just like "Yao". As another example, for the phonetic word, "zhong" , the true word may be "china" or "heavy" or "seed" (these are translations of words that sound similar to "zhong" . These words do not sound alike in English, but in Chinese Mandarin they do. The methods pursuant to this embodiment is not limited to English, or Chinese. Any language for which a language model may be constructed can be used.

[1053] Still referring to Figure 7b, once word alternatives are arrived at, for each phonetic word node, the server may generate multi true word nodes based on those words that are checked out. Then the topology relation is duplicated in the phonetic word graph in order to arrive at an expanded phonetic word graph as illustrated in Figure 7c which illustrates the resulting phonetic word graph from the word graph shown in Figure 7b.

[1054] Referring now to Figure 7c, the expanded phonetic word graph is illustrated. Here, the server will consider the different possibilities of the uttered sequence in accordance to a language model. For example, following the word, "I", a subject, the language model may dictate a verb, such as "want" and not a noun. Accordingly, a pruning strategy may take place where nouns that follow "I" are dropped from further consideration, for example, the word "medicine" a noun may not follow the word "I", which is a subject. This way the resulting search space may be considerably reduced. Similarly, a dictionary lexicon may be used to eliminate other similarly sounding words. Here, a language model based on a bigram language model or a trigram language model may be used. The choice of a bigram or trigram language model does not effect the methodologies pursuant to this embodiment.

[1055] Referring now to Figure 8, an exemplary block diagram of a speech recognition system including a client device, a server device and a communication network is illustrated. Speech input 800 may be words of a human user, for example, John. The speech input 800 may be captured by a microphone coupled to a client device belonging to John, for example, John's mobile phone, or PDA. John can use the training mode of his device to train his device to recognize his speech. The acoustic model 824,

located at the client device 810 is utilized in the training mode. As john is prompted to utter different words, phrases, or sentences, the acoustic model collects the data corresponding to each speech. When John is ready to communicate with a remote server 850 over a communication network, 840, he may switch off the training mode, and begin speaking as if he were carrying a normal conversation with another human. The client device 810 will capture John's speech and runs through feature extraction module 822 in order to perform a series of front end processes on the analog human speech signal 800 as is well known in the art. Whereas, in the prior art models, the extracted features are transmitted to the server for linguistic processing, according to one embodiment of the invention, an additional function takes place at the client device, i.e., acoustic processing which results in the generation of a phonetic word graph. By doing so, one embodiment takes advantage of an SD acoustic model, since John is able to train the device personally thus resulting in a SD personalized acoustic model. The prior art, is not able to take advantage of the lower WER associated with SD models since in the prior art, the features gathered at the client are transmitted directly to the server, and the server performs the acoustic recognition and analysis at the server. It is not practical to use a SD model with the prior art because the server services many users without knowing their identities. Thus the prior art is limited to SI models which are prone to higher error rates.

[1056] Once the acoustic processor receives the extracted features, it searches the speaker-dependent acoustic model trained by John's voice. The resulting matches are known phonetic units 830 which can be transmitted in a datagram to the server 850. The package is received at server850 and presented to a language processor 855 which in conjunction with pronunciation lexicon 857, and language model 859 determines a recognized word sequence.

CLAIMS

What is claimed is:

1. A method comprising:

receiving a human speech signal at a client node;

identifying features of said human speech signal;

identifying known phonetic units corresponding to said identified features;

forming a data packet containing at least one of said known phonetic units; and,

transmitting said data packet to a server node.

2. The method of claim 1 wherein said client node is selected from a group consisting of a mobile phone, a Personal Digital Assistant, and a portable computer system.

3. The method of claim 1, wherein identifying features of said human speech signal comprises:

performing end-point detection on said human speech signal;

performing pre-emphasizing filtration on said human speech signal; and

quantizing said human speech signal.

4. The method of claim 1, wherein identifying known phonetic units corresponding to said identified features comprises searching an acoustic model comprising:

a sub-word unit modeled by Hidden Markov Model comprising a Markov chain of acoustic states.

5. The method of claim 1, wherein identifying known phonetic units corresponding to said identified features comprises using a speaker-dependent acoustic model.

6. The method of claim 1, wherein said known phonetic units form a phonetic word graph.

7. The method of claim 1, wherein said data packet comprises a source address, a destination address and a binary representation of said known phonetic unit.

8. The method of claim 1, wherein said data packet is transmitted over the internet.

9. A system comprising:

a client node including:

a feature extraction module to identify features of a human speech signal,

an acoustic processing module coupled to said feature extraction modules, the acoustic processing module to identify known phonetic units from said identified features, and

a sender module coupled to said acoustic processor module, the sender module to form a data packet containing at least one of said phonetic units and transmit said data packet to a server; and

a server comprising:

a receiver module to receive said data packet and remove said at least one of said known phonetic units from said data packet, and

a language processing module to identify a word associated with said at least one of said known phonetic units.

10. The system of claim 9, wherein said client node is selected from a group consisting of a mobile phone, a Personal Digital Assistant, and a portable computer system.

11. The system of claim 9 wherein said feature extraction module is also configured to perform end-point detection, pre-emphasizing filtration and quantization on said human speech signal.

12. The system of claim 9, wherein said acoustic processing module includes a speaker-dependent acoustic model.

13. The system of claim 9, wherein said acoustic processing module forms a phonetic word graph according to said known phonetic units.

14. The system of claim 9, wherein said sender module forms a binary representation of said word graph and places said binary representation along with a source address and a destination address onto a datagram before transmitting said word graph.

15. A client device comprising:

a receiver module for receiving a human speech signal;

a feature extraction module coupled to said receiver module to identify features of said human speech signal;

an acoustic processing module coupled to said feature extraction module to identify known phonetic units from said identified features and form a data packet containing at least one of said phonetic units; and,

a sender module coupled to said acoustic processing module to send said data packet to a server node.

16. The client device of claim 15, wherein said feature extraction module is also configured to perform end-point detection, pre-emphasizing filtration, and quantization of said human speech signal.

17. The client device of claim 15, wherein said acoustic processing module includes a speaker-dependent acoustic model.

18. The client device of claim 15, wherein said acoustic processing module forms a word graph from said identified known phonetic units.

19. The client device of claim 18, wherein said word graph is a phonetic word graph.

20. A server comprising:

a receiver module to receive a data packet containing at least one known phonetic unit from a client node and removing said at least one known phonetic unit from said data packet; and

a language processing module coupled to said receiver module to determine a word associated with said at least one known phonetic unit.

21. The server of claim 20, wherein said data packet is received over the internet.

22. The server of claim 20, wherein said data packet is a datagram comprising:

a header section containing an address of said client node, an address of said server, and said known phonetic units.

23. A computer-readable medium including a program executable by a processor, comprising:

a first subroutine for receiving a human speech signal at a client node;

a second subroutine for identifying features of said human speech signal;

a third subroutine for identifying known phonetic units corresponding to said identified features;

a fourth subroutine for forming a data packet containing at least one of said known phonetic units; and

a fifth subroutine for transmitting said data packet to a server node.

24. The computer-readable medium of claim 23, wherein

said third subroutine forms a phonetic word graph from said known phonetic units.

25. The computer-readable medium of claim 24, wherein said data packet is a datagram containing said phonetic word graph, an address of said client node, and an address of said server node.

26. The computer-readable medium of claim 23, wherein said third subroutine also forms a speaker-dependent acoustic model.

27. The computer-readable medium of claim 23, wherein said fifth subroutine transmits said data packet to said server node over the internet.

28. A computer-readable medium including a program executable by a processor, comprising:

a first subroutine for receiving a data packet containing at least one known phonetic unit from a client node;

a second subroutine for removing said at least one known phonetic unit from said data packet; and

a third subroutine for identifying a word associated with said at least one known phonetic unit.

29. The computer-readable medium of claim 28, wherein said data packet is sent by a client node over the internet.

30. The computer-readable medium of claim 28, wherein said data packet is a datagram containing said at least one known phonetic unit along with an address of said client node.

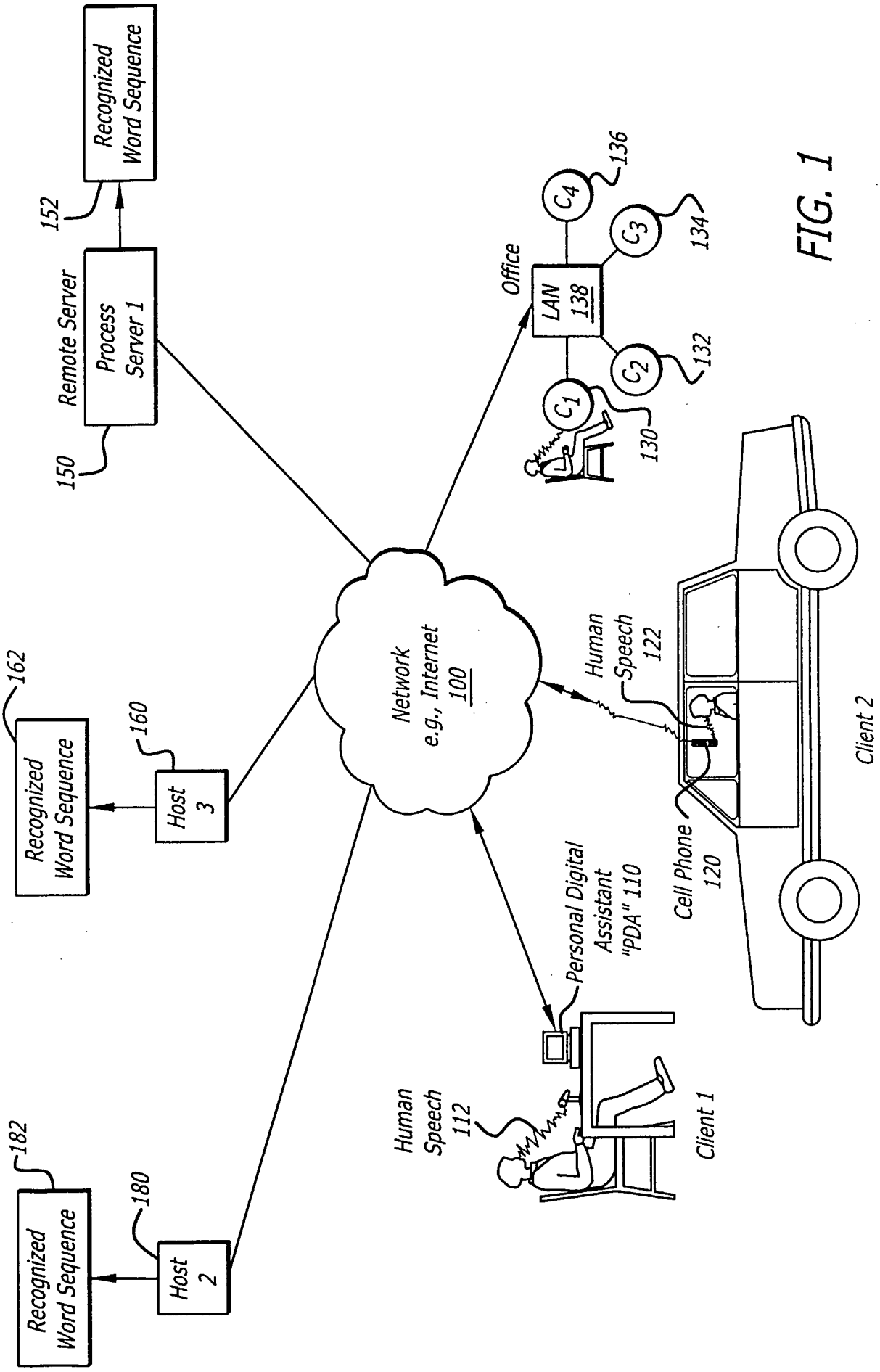

Recognized Word Sequence
152
Remote Server
Process Server 1
150
Recognized Word Sequence
162
Host 3
160
Recognized Word Sequence
182
Host 2
180
Network e.g., Internet 100
Office
LAN 138
C4
136
C3
134
C2
132
C1
130
Human Speech 122
Cell Phone 120
Client 2
Personal Digital Assistant "PDA" 110
Human Speech 112
Client 1

FIG. 1

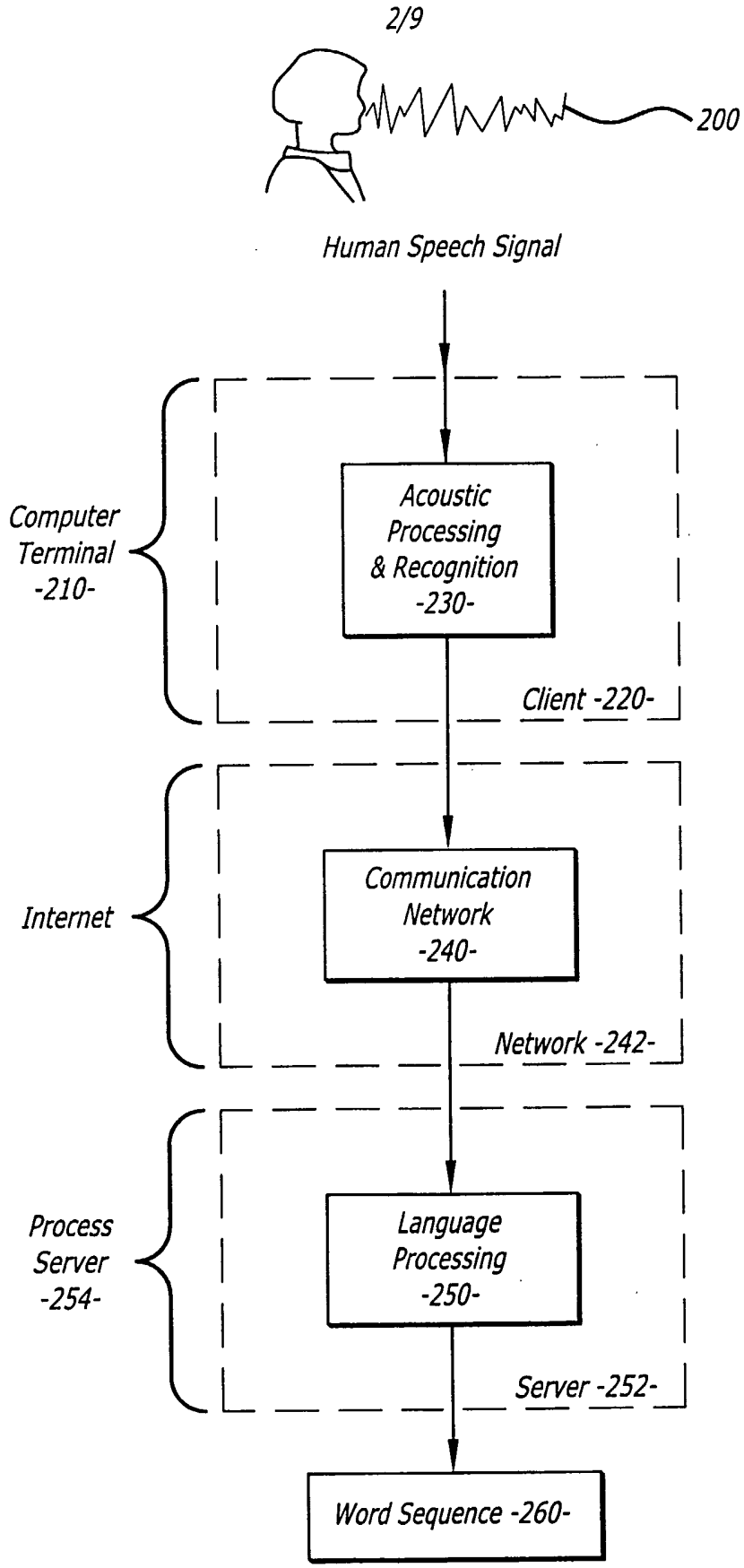
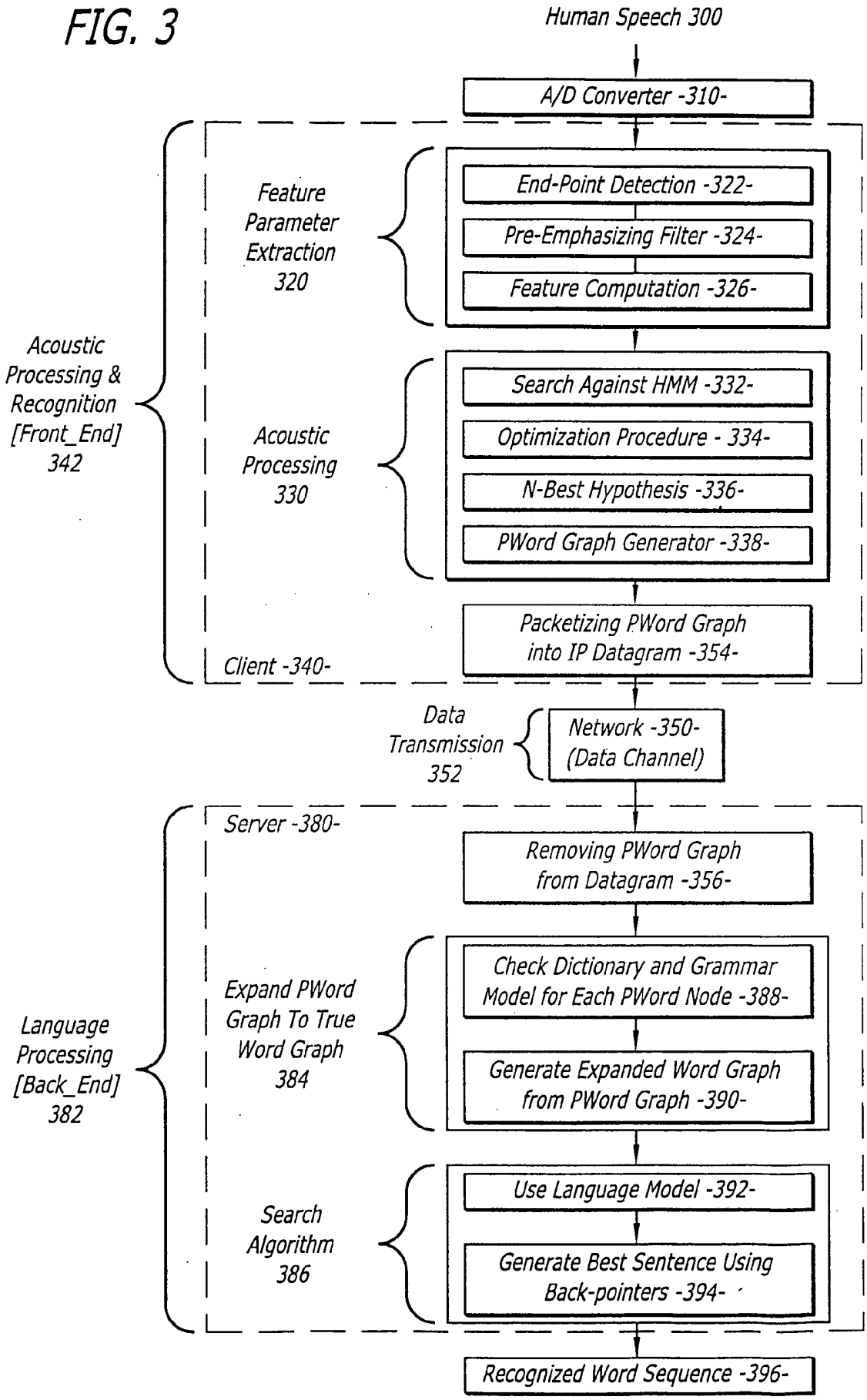

200
Human Speech Signal
Computer Terminal -210-
Acoustic Processing & Recognition -230-
Client -220-
Internet
Communication Network -240-
Network -242-
Process Server -254-
Language Processing -250-
Server -252-
Word Sequence -260-

FIG. 2

FIG. 3


Human Speech 300
A/D Converter -310-
Acoustic Processing & Recognition [Front_End] 342
Feature Parameter Extraction 320
End-Point Detection -322-
Pre-Emphasizing Filter -324-
Feature Computation -326-
Acoustic Processing 330
Search Against HMM -332-
Optimization Procedure - 334-
N-Best Hypothesis -336-
PWord Graph Generator -338-
Packetizing PWord Graph into IP Datagram -354-
Client -340-
Data Transmission 352
Network -350- (Data Channel)
Server -380-
Removing PWord Graph from Datagram -356-
Language Processing [Back_End] 382
Expand PWord Graph To True Word Graph 384
Check Dictionary and Grammar Model for Each PWord Node -388-
Generate Expanded Word Graph from PWord Graph -390-
Search Algorithm 386
Use Language Model -392-
Generate Best Sentence Using Back-pointers -394-
Recognized Word Sequence -396-

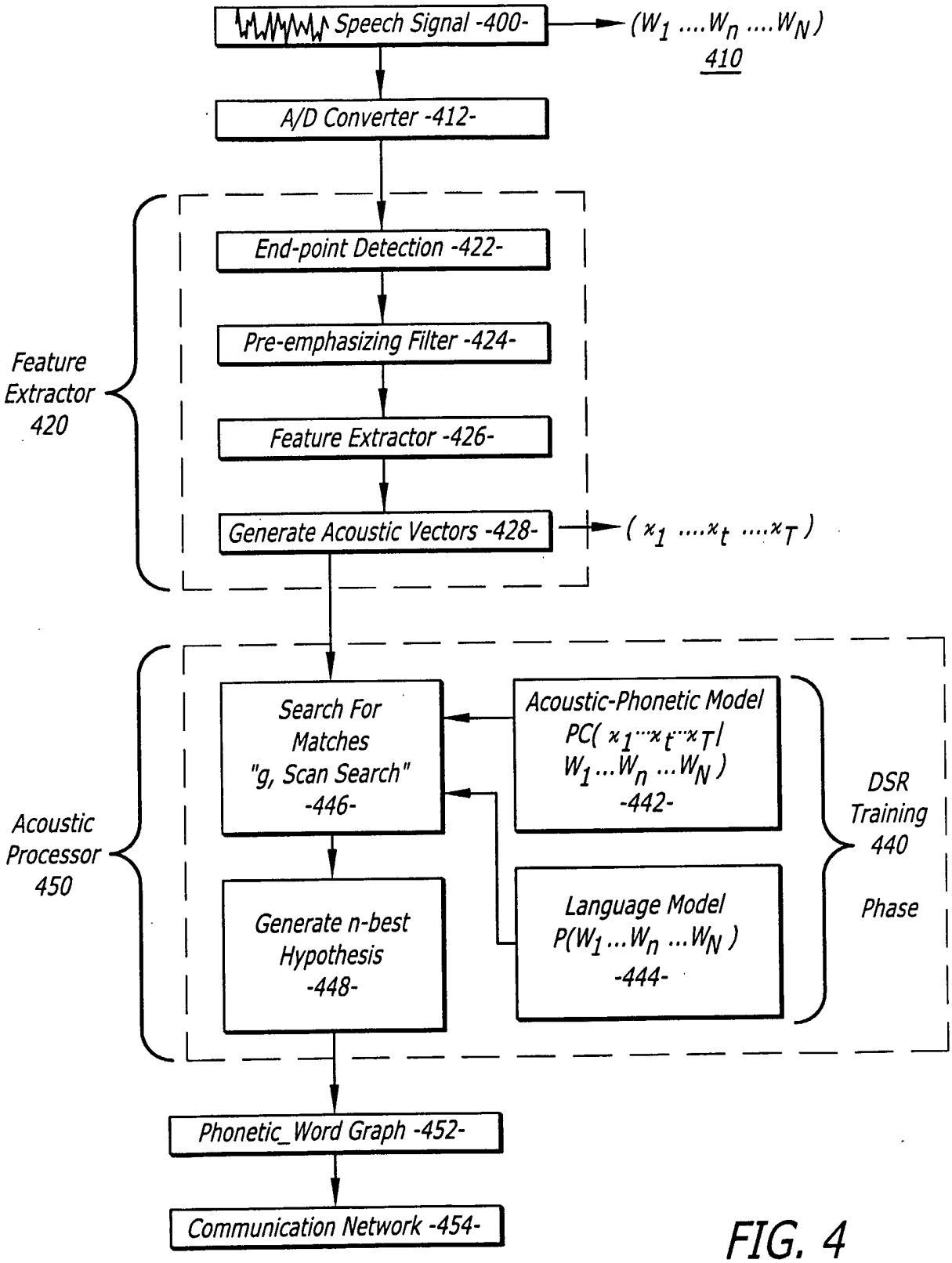

Speech Signal -400-
(W1 ....Wn ....WN)
410
A/D Converter -412-
End-point Detection -422-
Pre-emphasizing Filter -424-
Feature Extractor -426-
Generate Acoustic Vectors -428-
(x1 ....xt ....xT)
Feature Extractor 420
Search For Matches "g, Scan Search" -446-
Acoustic-Phonetic Model PC(x1...xt...xT/ W1...Wn ...WN) -442-
Language Model P(W1...Wn ...WN) -444-
Generate n-best Hypothesis -448-
Acoustic Processor 450
DSR Training 440
Phase
Phonetic_Word Graph -452-
Communication Network -454-

FIG. 4

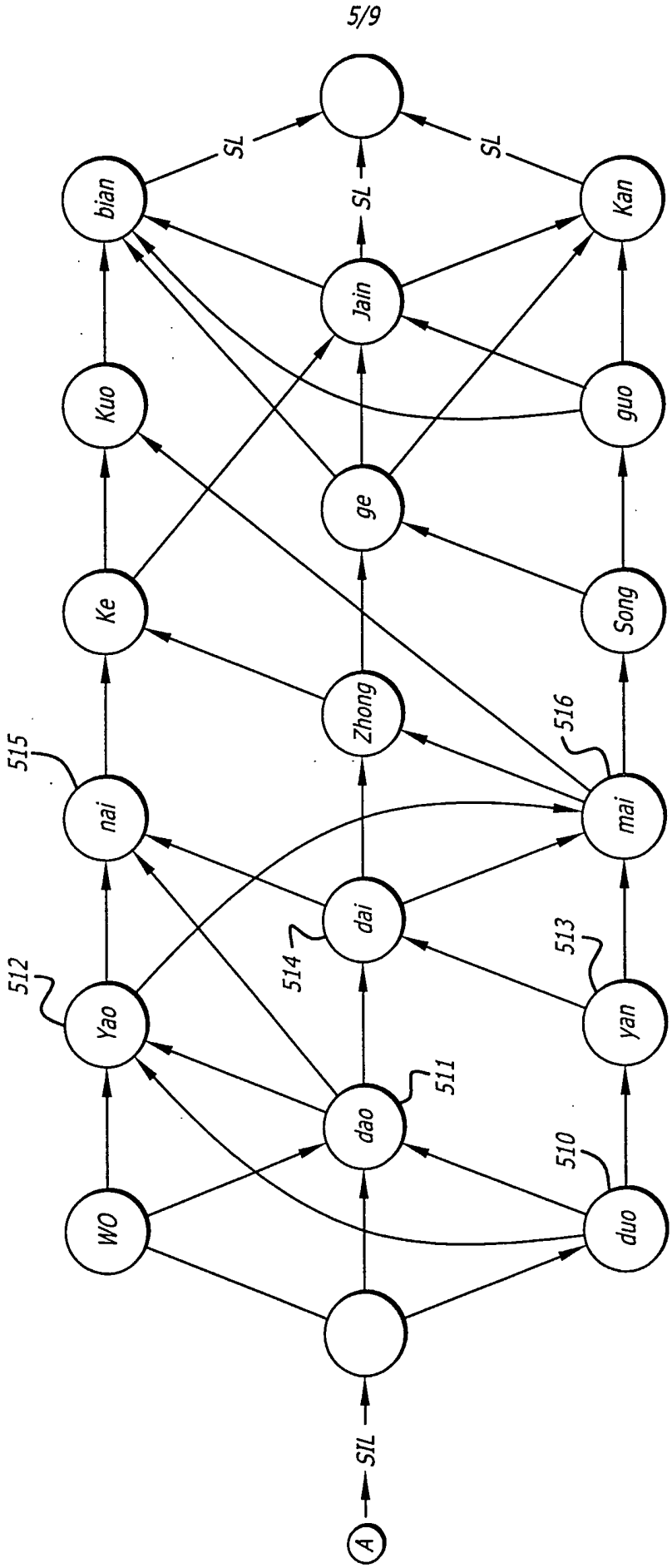

A
SIL
WO
Yao
nai
Ke
Kuo
bian
dao
dai
Zhong
ge
Jain
SL
SL
SL
duo
yan
mai
Song
guo
Kan
510
511
512
513
514
515
516

FIG. 5

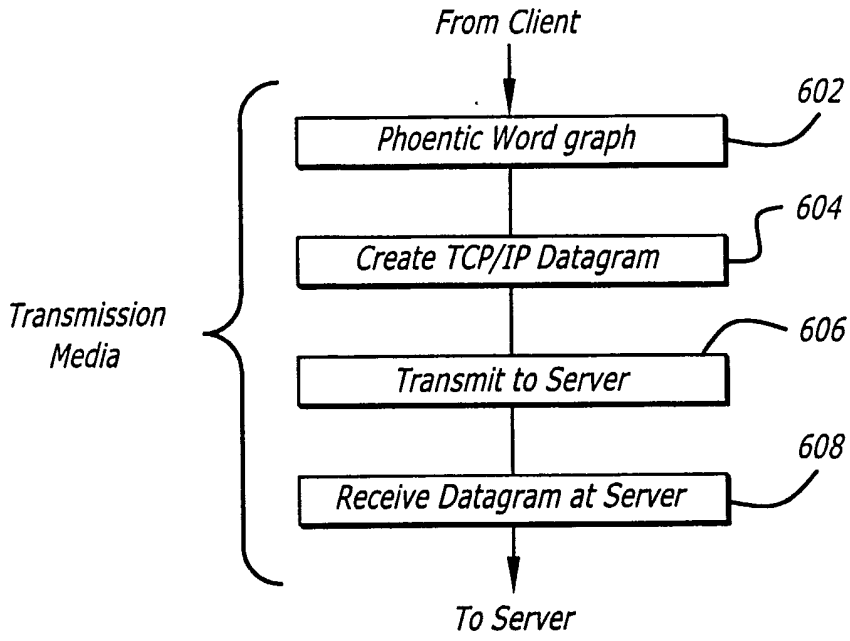

From Client
602
Phoentic Word graph
604
Create TCP/IP Datagram
606
Transmit to Server
608
Receive Datagram at Server
Transmission Media
To Server

FIG. 6A

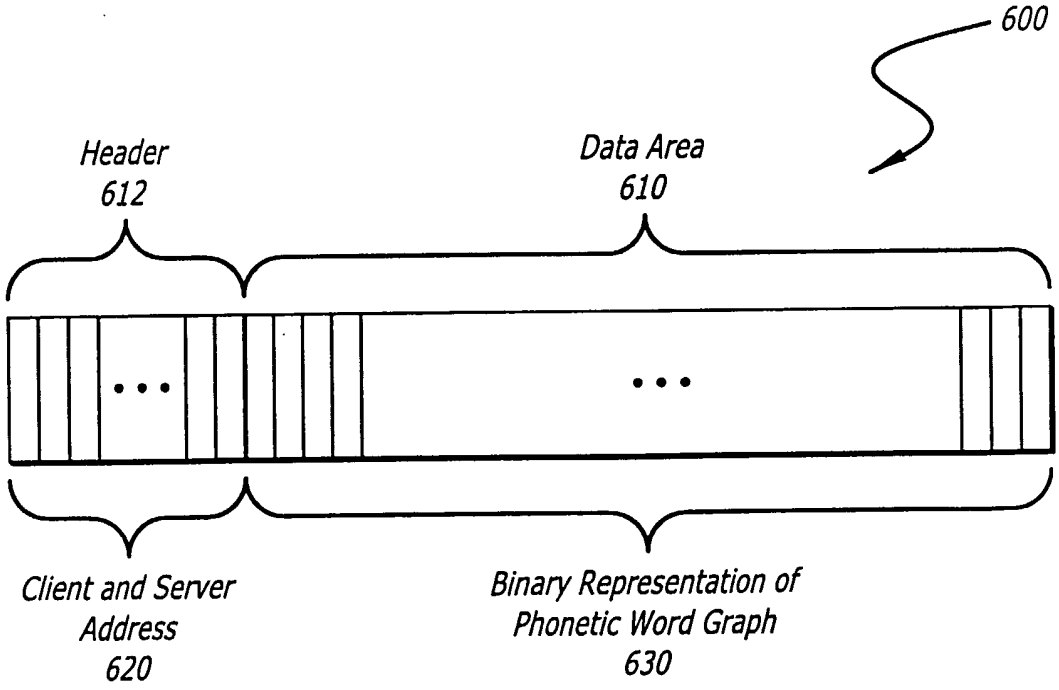

600
Header
612
Data Area
610
Client and Server Address
620
Binary Representation of Phonetic Word Graph
630

FIG. 6B

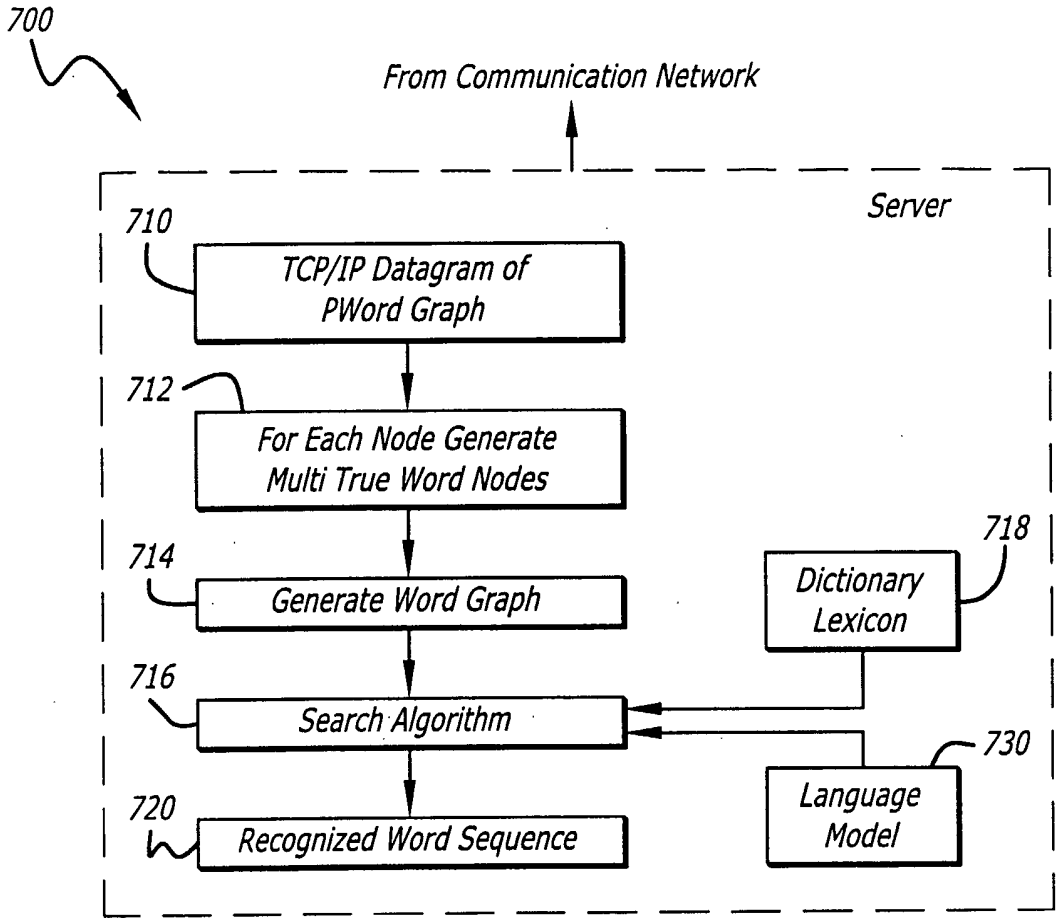

700
From Communication Network
Server
710
TCP/IP Datagram of PWord Graph
712
For Each Node Generate Multi True Word Nodes
714
Generate Word Graph
716
Search Algorithm
720
Recognized Word Sequence
718
Dictionary Lexicon
730
Language Model

FIG. 7A

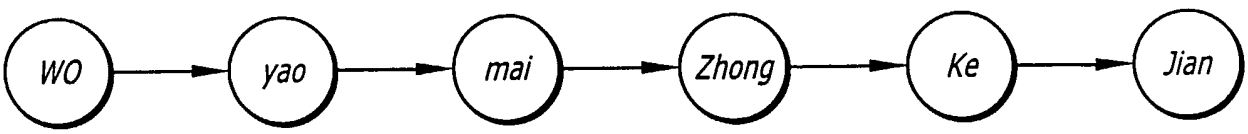

WO
yao
mai
Zhong
Ke
Jian

FIG. 7B

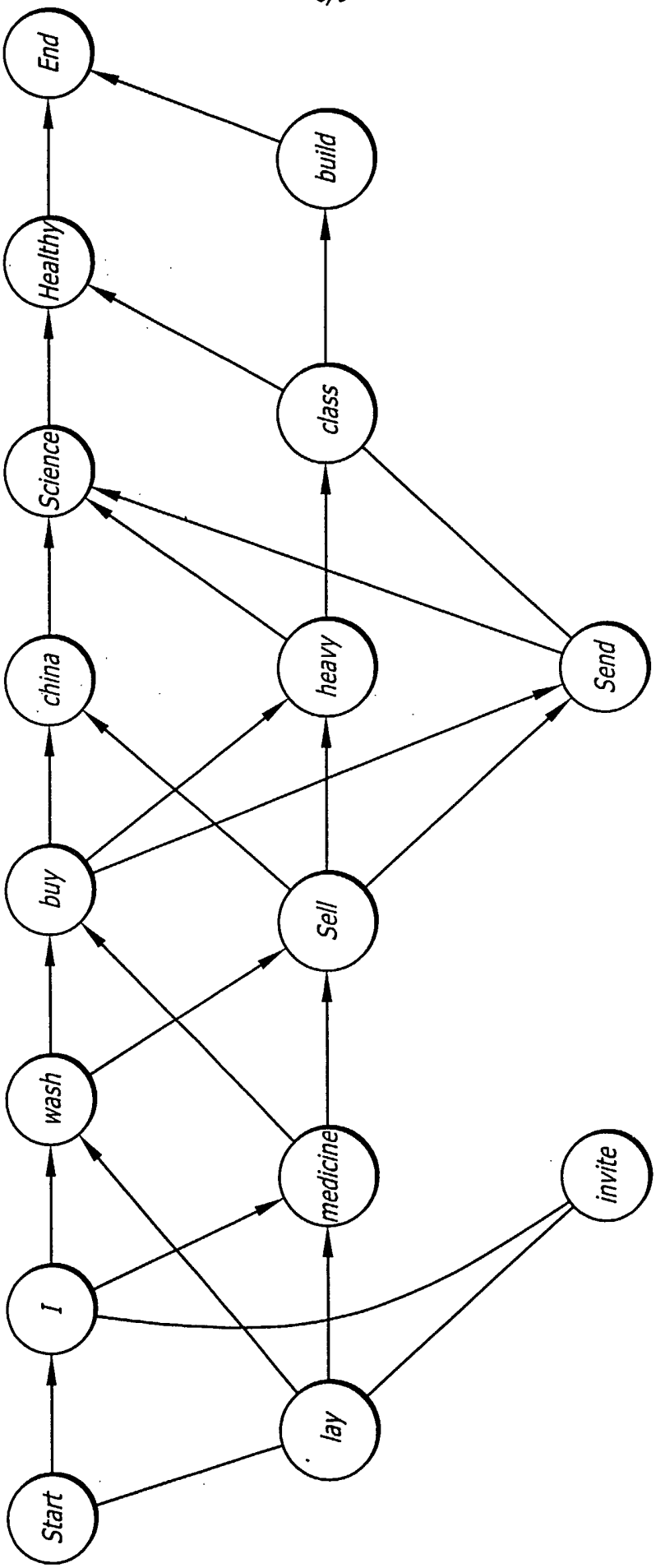

End
build
Healthy
class
Science
china
heavy
Send
buy
Sell
wash
medicine
invite
I
lay
Start

FIG. 7C

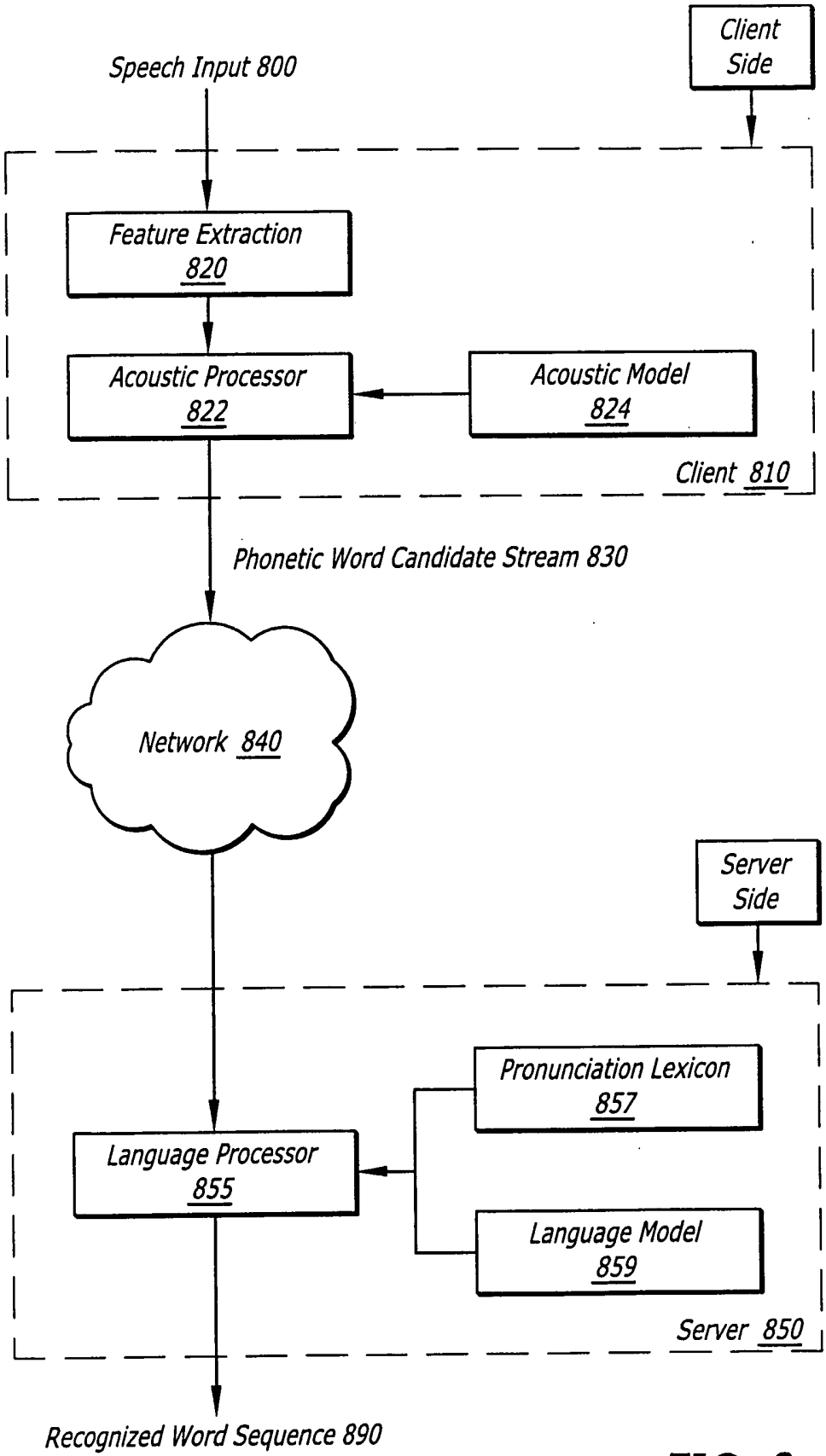

Client Side
Speech Input 800
Feature Extraction 820
Acoustic Processor 822
Acoustic Model 824
Client 810
Phonetic Word Candidate Stream 830
Network 840
Server Side
Pronunciation Lexicon 857
Language Processor 855
Language Model 859
Server 850
Recognized Word Sequence 890

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN01/01030

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/00

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC：G10L 15/00、G10L 15/02、H04L 12/56

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

NONE

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

RS、清华非专利数据库：语音识别、语音通信、分布式、客户机、服务器、网络

WPI、EPODOC、PAJ：speech recognition、distribut+、client、server、network?、speaker dependent

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US5960399 A Sep.28,1999	1-30
A	EP1031963 A2 Aug.30,2000	1-30
A	WO00/54252 A Sep.14,2000	1-30
A	EP0854418 A2 Jul.22,1998	1-30
A	语音识别在电话网络中的应用（伍湘彬译 《电声技术》1996 年 10 月）	1-30
T	CN1315721 A Oct.3,2001	1-30

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
Dec.19,2001	03 JAN 2002 (03.01.02)
Name and mailing address of the ISA/CN 6 Xitucheng Rd., Jimen Bridge, Haidian District, 100088 Beijing, China Facsimile No. 86-10-62019451	Authorized officer 熊婷 Telephone No. 86-10-62093798

Form PCT/ISA /210 (second sheet) (July 1998)