



- (51) **International Patent Classification:**
G10L 15/06 (2006.01) **G10L 15/18** (2006.01)
G10L 15/28 (2006.01)
- (21) **International Application Number:**
PCT/US2012/034261
- (22) **International Filing Date:**
19 April 2012 (19.04.2012)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
61/477,533 20 April 2011 (20.04.2011) US
- (71) **Applicant (for all designated States except US):** **ROBERT BOSCH GMBH** [DE/DE]; Postfach 30 02 20, 70442 Stuttgart (DE).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** **WENG, Fuliang** [CN/US]; 1479 Lloyd Way, Mountain View, CA 94040 (US). **FENG, Zhe** [CN/US]; 490 Thompson Avenue, Mountain View, CA 94043 (US). **XU, Kui** [CN/US]; 777 San Antonio Road, Unit #85, Palo Alto, CA 94303 (US).

ZHAO, Lin [CN/US]; 1032 W. Remington Drive, #6-17, Sunnyvale, CA 94087 (US).

- (74) **Agent:** **MAGINOT, Paul, J.**; Maginot, Moore & Beck LLP, Chase Tower, 111 Monument Circle, Suite 3250, Indianapolis, IN 46204-5109 (US).

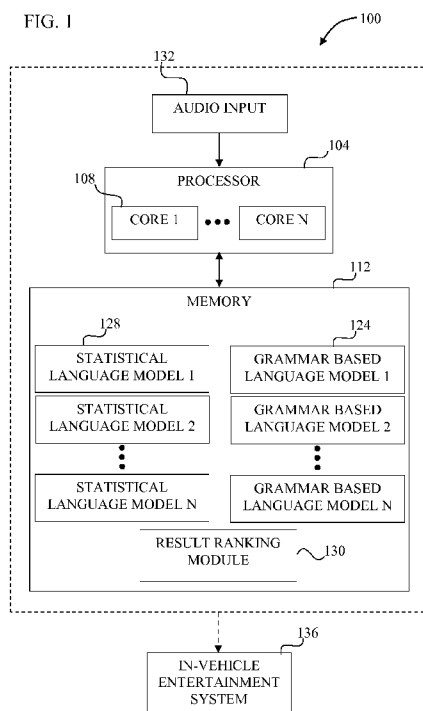
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, 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, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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,

[Continued on next page]

- (54) **Title:** SPEECH RECOGNITION USING MULTIPLE LANGUAGE MODELS

FIG. 1



- (57) **Abstract:** In accordance with one embodiment, a method of generating language models for speech recognition includes identifying a plurality of utterances in training data corresponding to speech, generating a frequency count of each utterance in the plurality of utterances, generating a high-frequency plurality of utterances from the plurality of utterances having a frequency that exceeds a predetermined frequency threshold, generating a low-frequency plurality of utterances from the plurality of utterances having a frequency that is below the predetermined frequency threshold, generating a grammar-based language model using the high-frequency plurality of utterances as training data, and generating a statistical language model using the low-frequency plurality of utterances as training data.

WO 2012/145519 A1



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))*

SPEECH RECOGNITION USING MULTIPLE LANGUAGE MODELS

This application claims the benefit of priority of U.S. provisional application serial no. 61/477,533, filed April 20, 2011, the disclosure which is herein incorporated by
5 reference in its entirety.

TECHNICAL FIELD

[0001] This application relates generally to the field of automated speech recognition, and more particularly to intelligent speech recognitions systems and methods that employ
10 multiple language models.

SUMMARY

[0002] In accordance with one embodiment, a method of generating language models for speech recognition includes identifying a plurality of utterances in training data
15 corresponding to speech, generating a frequency count of each utterance in the plurality of utterances, generating a high-frequency plurality of utterances from the plurality of utterances having a frequency that exceeds a predetermined but adjustable frequency threshold, generating a low-frequency plurality of utterances from the plurality of utterances having a frequency that is below the predetermined frequency threshold,
20 generating a grammar-based language model using the high-frequency plurality of utterances as training data, and generating a statistical language model using the low-frequency plurality of utterances as training data.

[0003] In accordance with a further embodiment, a method of performing speech recognition includes collecting a plurality of utterances, generating a frequency count of

each utterance in the plurality of utterances, identifying a high-frequency segment of the plurality of utterances based upon a predetermined frequency threshold, identifying a low-frequency segment of the plurality of utterances based upon the predetermined frequency threshold, generating a final grammar-based language model based on at least a portion of the high-frequency segment, generating a final statistical language model based on at least a portion of the low-frequency segment, performing a first speech recognition using the final grammar-based language model, performing a second speech recognition using the final statistical language model; and determining a recognized speech based upon the first speech recognition and the second speech recognition. The method further includes the refinement of the threshold to improve the performance of the speech recognizer with multiple models.

[0004] In yet another embodiment, an intelligent speech recognition system includes at least one audio input, at least one memory, a family of grammar-based language models stored within the at least one memory, a family of statistical language models stored within the at least one memory, and at least one processor operably connected to the at least one audio input and the at least one memory and configured to (i) perform a first speech recognition using the family of grammar-based language models, (ii) perform a second speech recognition using the family of statistical language models, and (iii) determine a recognized speech based upon the first speech recognition and the second speech recognition, wherein the family of grammar-based language models is generated based upon a generated frequency count of each utterance in a plurality of utterances identified as being in a high-frequency segment of the plurality of utterances based upon a predetermined frequency threshold, and the family of statistical language models is

generated based upon a generated frequency count of each utterance in a plurality of utterances identified as being in a low-frequency segment of the plurality of utterances based upon the predetermined frequency threshold.

5 BRIEF DESCRIPTION OF THE DRAWINGS

[0005] FIG. 1 is a schematic view of an intelligent speech recognition system that is configured to use a family of grammar-based language models and a family of statistical language models to perform speech recognition on spoken utterances.

10 [0006] FIG. 2 is a block diagram of a process for generating the family of grammar-based language models and the family of statistical language models using selected segments of a training data set.

[0007] FIG. 3 is a schematic diagram of a training system, which in this embodiment uses components in common with the speech recognition system of FIG. 1, which can be used to execute the process of FIG. 2.

15 [0008] FIG. 4 is a simplified graphical representation of an example statistical model identifying probabilities for a word to be spoken based upon a single previously spoken word.

[0009] FIG. 5 is a block diagram of a process that is used to generate, for each of the models within the family of grammar-based language models and the family of statistical
20 language models, a range of confidence scores, generated by a recognizer associated with a respective one of the language models, each point in the range of confidence scores associated with a correctly recognized test utterance.

[0010] FIG. 6 is a block diagram of a process for using the family of grammar-based language models and the family of statistical language models to recognize unknown utterances.

[0011] FIG. 7 is a schematic view of an intelligent speech recognition system that is configured to use a family of grammar-based language models and a family of statistical language models to perform speech recognition on spoken utterances wherein a small footprint family of models is located within a client device and both a large footprint family of models and a small footprint family of models is stored within a cloud computing system connected to the client device by a network.

DETAILED DESCRIPTION

[0012] For a general understanding of the details for the systems and processes disclosed herein, the drawings are referenced throughout this document. In the drawings, like reference numerals designate like elements. As used herein, the term “utterance” refers to any speech spoken by a human including words and phrases. The term “utterance data” refers to data corresponding to one or more utterances. The utterance data may correspond to a direct sound recording of the utterance, or may be processed data generated from a speech recognizer, which typically includes a front-end processor, such as a digital signal processor, acoustic modeler, and a language model.

[0013] FIG. 1 depicts a schematic diagram of a speech recognition system 100. System 100 includes a processor 104, memory 112, and audio input 132. The processor 104 is an electronic processing device such as a microcontroller, application specific integrated circuit (ASIC), field programmable gate array (FPGA), microprocessor

including microprocessors from the x86 and ARM families, or any electronic device configured to perform the functions disclosed herein.

[0014] In the embodiment of FIG. 1, processor 104 includes N cores 108. Each core 108 in the processor 104 is configured to execute programmed instructions, and the cores 108 are configured to execute programmed instructions concurrently. In some embodiments, each of the cores 108 have a substantially identical configuration, while in other embodiments some of the cores are configured for different uses. For example, in one embodiment one set of the cores 108 are general processing cores such as processing cores in the x86 or ARM family of instruction sets, while another set of the cores 108 are configured to perform digital signal processing (DSP) operations that are commonly used in processing of audio data for speech recognition. Some or all of the functions provided by the memory 112 and audio input 132 may be integrated with the processor 104 using either hardware or software in a system on a chip (SoC) configuration.

[0015] Memory 112 is a data storage device configured to store data and provide the stored data to the processor 104, and to write data in response to a write command from the controller 104. Various embodiments of the memory 112 include random access memory (RAM), solid-state devices, magnetic hard drives, optical drives, and the like. The memory 112 is configured to hold one or more stored programs that provide programmed instructions that are executed by the processor 104. The programs include a family of grammar-based language models 124 and a family of statistical language models 128 for analysis of utterances, and a result ranking model 130, all of which are described more fully below.

[0016] Audio input 132 is a device that receives data corresponding to speech. In one embodiment, the audio input 132 is an acoustic transducer that records speech and provides signals corresponding to the recorded speech to the processor 104. For example, the acoustic transducer can be a microphone, a speaker, a dual microphone and speaker system, or other known transducer that converts audio signals to electrical signals and vice versa. The processor 104 may be integrated with the audio input 132 as a single device instead of separate devices. In other embodiments, the audio input 132 records speech and also performs signal processing and acoustic modeling of the speech data before providing the modeled data to the controller 104. In still another embodiment, the audio input 132 is a software program that extracts audio information, including speech, from one or more data sources such as digitally encoded audio and audio-visual sources. In each of the embodiments, audio input 132 provides data corresponding to speech to the processor 104 in a format that enables the processor 104 to perform speech recognition on the data.

[0017] The speech recognition system 100 may be operably coupled to a variety of devices 136 including computing, electronic, and mechanical devices 136 to enable voice operation of the systems. In one embodiment, the speech recognition system 100 is implemented solely on the electronic device 136 to facilitate user interaction with the device. In another embodiment, the speech recognition system 100 is distributed across multiple computing devices. In one such embodiment, the family of models 124/128 having the smallest footprint is stored within the memory 112 while the family of models 124/128 having the largest footprint is stored in a remote memory.

[0018] The device 136 is, in various embodiments, an in-vehicle entertainment system, an end-user computer system, a mobile computing device (such as personal digital assistance, mobile phone, smartphone, tablet, laptop, or the like), a consumer electronic device, a gaming device, a music player, a security system, a network server or server system, a telemedicine system, or any combination or portion thereof. The device 136 may be adapted to communicate with other devices, such as clients and/or servers, over a wired or wireless communications network.

[0019] In general, the system 100 receives spoken user natural language requests corresponding to various operations implemented in the device 136 which in the embodiment of FIG. 1 is an entertainment system 136. For example, the audio input 132 may be configured to receive a spoken request such as “PLAY NEXT SONG” and the processor 104 generates a command signal for the entertainment system 136 to advance playback to the next song in a song playlist. In some embodiments, the spoken user requests may be in the form of predetermined commands.

[0020] The family of grammar-based models 124 and the family of statistical models 128 of FIG. 1 are generated in one embodiment by a process 200 depicted in FIG. 2. The process 200 can be used to generate multiple language models from training data.

Process 200 is described with reference to the speech language model training system 300 of FIG. 3, but the process 200 is also suitable for use with alternative speech

recognition system embodiments. Moreover, while the training system 300 of FIG. 3 uses the same processor 104, audio input 132, and memory 112 of FIG. 1, different components are used to generate the family of grammar-based models 124 and the family

of statistical models 128 with the family of grammar-based models 124 and the family of statistical models 128 then saved into the system 100 in other embodiments.

[0021] The memory 112 in FIG. 3 includes training data 301, an annotation module 302 for classification of utterances in speech data, a grammar-based training module 303 for generation of one or more grammar-based language model(s) for speech recognition, and a statistical training module 304 for generation of one or more statistical language model(s) for speech recognition.

[0022] The training data 301 correspond to a plurality of utterances that are used for generating language models to perform speech recognition. The training data 301 include speech data corresponding to a plurality of utterances. Each utterance is a spoken word, phrase having multiple words, or a sentence with multiple words. In a typical embodiment, the training data 301 include multiple variations of a single utterance, such as speech data for a single phrase as spoken by various people. The contents of the training data 301 are configurable to include utterances that are typical of speech patterns that the system 100 recognizes during operation.

[0023] Process 200 includes generation of both grammar-based language models and statistical language models such as n-gram models and class-based n-gram models.

Grammar-based language models include finite state grammars and context free grammars, among others. These models are typically optimized for accurate and efficient recognition of a comparatively small set of terms that are frequently used in speech. An example of a finite state grammar is represented as one or more graphs with a plurality of nodes representing words or word classes connected to each other by edges. Various rules in the grammar-based model govern the distribution of edges between words and word

classes that are recognized in the grammar. The rules can be generated from observations of transitions between words in frequently used utterances in training data. Some of the grammar rules may be derived recursively, for example, through word classes.

[0024] Statistical language models include n-gram models that are generated from a large corpus of text. In one configuration, n-gram models use the state of one or more known words to provide a conditional probability for what the next word will be. The “n” in n-gram represents a number of given words that are evaluated to determine the probability of the next word. Thus, in a 2-gram or bigram model trained over an English corpus, one word has empirically measured probabilities of preceding another word. This concept is discussed with reference to FIG. 4 which is a highly simplified graphical representation of a statistical language model 310.

[0025] In FIG. 4, an initial word 312 has been previously identified, which in this example is the word “traffic”. Based upon a large corpus of text, a set of possible words for a word which follows the word “traffic” is identified. This set of word is represented in this simplified example by the “following words” 314, 316, 318, 320, and 322. For each of the following words 314, 316, 318, 320, and 322, a respective probability 324, 326, 328, 330, and 332 is assigned based upon the empirical data from the corpus of text. The probabilities 324, 326, 328, 330, and 332, which will add up to a value of “1”, indicate the likelihood that the associated following word 314, 316, 318, 320, or 322, will be spoken. Thus, when the previously identified word is “TRAFFIC”, in the example of FIG. 4, the most likely word to occur next is “LIGHT”.

[0026] While in the above simplified example a single word was used to generate the probabilities for a following word, other n-gram models use two or more known words to generate a conditional probability for the next word in a phrase.

[0027] The grammar-based language models and statistical models such as the n-gram

5 models are each effective in performing speech recognition. The two types of models, however, perform differently for different speech patterns. Grammar-based models typically perform the best when analyzing a narrow scope of speech and the grammatical alternatives have roughly the same probability, or are uniformly distributed. Statistical models perform best when analyzing more complex phrases with a stable but non-flat
10 distribution so that different alternatives can be easily distinguished.

[0028] Process 200 begins by collecting utterance data (block 204). In one embodiment, the utterance data is collected directly by prompting a user to speak predetermined words and phrases. In another embodiment, the utterance data includes a data corresponding to words, phrases, and sentences recorded from multiple sources. In
15 one embodiment, the user simply responds naturally to prompts using the user's own natural expressions or utterances. The collected utterance data is stored as the training data 301 in the memory 112.

[0029] Process 200 continues by annotating the utterance data with class labels (block 208). Annotations include classes that are assigned to one or more utterance entries. The
20 classes include, for example, user natural language requests, commands and named entities that the speech recognition system recognizes. Examples of user natural language requests or commands in an automotive speech recognition system include playing a song, pausing playback, setting a location in a navigation system, dialing a phone

number, texting a message, playing a voice message, activating/deactivating a security system, and requesting a weather report. Annotations may also be applied to class names of entities such as titles of songs, albums, place of interests, and radio channels. In system 300, the processor 104 may annotate the utterance data dynamically using the
5 annotation module 302 prior to generating the language models.

[0030] As described above, each of the grammar-based language models and statistical language models are better suited to deal with a respective type of utterances.

Accordingly, process 200 segments the training utterance data to enable each language model to be generated with a subset of the training utterance data that includes utterances
10 that are best suited to each of the language models. To this end, process 200 identifies the frequency of each utterance in the training utterance data (block 212). The number of utterances that correspond to each class annotation can be used to generate a histogram of the frequency of each type of utterance in the training data.

[0031] Process 200 divides the training utterance data into two segments based on the
15 identified frequency of utterances in the training utterance data (block 216). Training utterance data which occurs with a frequency that is greater than a predetermined threshold frequency is apportioned to a “high-frequency segment”. Training utterance data which occurs with a frequency that is less than a predetermined threshold frequency is apportioned to a “low-frequency segment”. The threshold frequency may be a
20 frequency threshold selected based on various factors including the total number of utterances present in the training utterance data and the distribution of frequencies of the utterances. In one embodiment, the threshold frequency is selected to be the median of the distribution of frequencies in the training utterance data.

[0032] Process 200 generates a first grammar-based language model using the high-frequency utterance data segment (block 220) and generates a first statistical language model using the low-frequency utterance data segment (block 224). Some embodiments perform the generation of the grammar-based language model and the statistical language model concurrently. In system 300, processor 104 may generate the grammar-based language model using the grammar training module 303 on one processor, while the processor 104 executes the statistical training module 304 on a different processor to generate the statistical language model. Since the grammar-based language model is generated using the high-frequency utterance data segment and the statistical model is generated using the low-frequency utterance data segment, the language models are generated independently without requiring dependencies between the utterance data applied to each model.

[0033] The resulting statistical language model “A” trained with low-frequency data may be interpolated with another background statistical language model “B” trained with a set of data which may also contain the high frequency data. In this case, the interpolation weight for the model “A” typically is much higher than the interpolation weight for the model “B”. This interpolation can be used for all the statistical models described in this invention.

[0034] Process 200 tests the initially generated language models and generates additional language models as needed during a tuning process which begins at block 232. Process 200 first performs speech recognition tests using both the grammar-based language model and the statistical language models (block 232) on a set of test utterances. In one embodiment, both language models perform speech recognition over a

plurality of test utterances. Another embodiment performs tests using two sets of previously determined utterance data, one set of utterances being selected to test the grammar-based language model and the other set of utterances being selected to test the statistical language model.

5 **[0035]** Once a respective speech recognition result has been generated by each of the language models for the test utterances, the performance of each language model is measured using one or more factors including the accuracy of recognition and the amount of time that was required by each model to produce a speech recognition result. The performance of each language model is compared to a predetermined performance
10 threshold. Each language model may have a different performance threshold or target. In one embodiment, the grammar-based language model has a performance threshold of a 95% accuracy with an average recognition time of one second, while the statistical language model has a perform threshold of 80% accuracy with an average recognition time of two seconds.

15 **[0036]** If neither of the language models meets or exceeds the performance threshold (block 236), the initial language models are rejected and the process 200 increases the frequency threshold used to divide the utterance data into the high-frequency segment and the low-frequency segment (block 240). Process 200 subsequently segments the utterance data using the increased frequency threshold, generates grammar-based and
20 statistical language models with the revised utterance data segments, and performs speech recognition tests with the revised language models as described above in process blocks 216 – 232 until at least one model reaches the performance target (block 236). Once a language model reaches the performance target, in any of the scenarios described herein,

the language model is saved in the appropriate family of language models 124 or 128 in the memory 112.

[0037] If both the grammar-based and statistical language models exceed the performance threshold (block 244), the tuning process ends and both of the language models are saved and the process 200 continues with a ranking phase that is described in more detail below.

[0038] In the alternative, one of either the grammar-based language model or the statistical language model exceeds the performance threshold, while the other language model falls below the performance threshold (block 244). If only the grammar-based language model exceeds the performance threshold (block 248), the original grammar-based language model is saved. Then, the low-frequency utterance data segment used to generate the statistical language model is subdivided into a high-frequency sub-segment and a low-frequency sub-segment using a second frequency threshold that is lower than the first frequency threshold (block 252).

[0039] The just tested statistical language model is then discarded and process 200 generates new grammar-based and statistical language models. A new grammar-based language model (high frequency model) is generated using the original high-frequency segment and the high-frequency sub-segment (block 220). In an alternative embodiment, a new grammar-based model is generated using only the high frequency sub-segment. Additionally, a new statistical language model (low-frequency model) is generated using only the low-frequency sub-segment (block 224).

[0040] The process 200 then continues at block 232, and performs speech recognition tests with the new language models (block 232). The newly generated grammar-based

language model based on the high-frequency sub-segment in the low-frequency segment can have different grammar rules than the earlier generated grammar-based language model based on the original high-frequency utterance data segment. If both of the language models meets or exceeds the performance threshold, the models are saved and the process stops. If only one of the language models meets or exceeds the performance criteria, then that model is saved and process 200 continues at either block 252 or 256.

[0041] In cases wherein only the statistical language model exceeds the predetermined performance threshold (block 248), the statistical language model is saved in some embodiments, and the high-frequency utterance data segment used to generate the grammar-based language model is subdivided into a high-frequency sub-segment and a low-frequency sub-segment using a third frequency threshold that is higher than the original threshold (block 256).

[0042] The just tested grammar-based language model is then discarded and process 200 generates new grammar-based and statistical language models. A new grammar-based language model (high frequency model) is generated using only the high-frequency sub-segment (block 220). Additionally, a new statistical language model (low-frequency model) is generated using the low-frequency sub-segment (block 224). In alternative embodiments, a statistical language model may be generated using both the low-frequency sub-segment and the original low frequency segment. The process 200 then continues at block 232, and performs speech recognition tests with the new language models (block 232). The newly generated statistical language model based on the low-frequency sub-segment can have different grammar rules than the earlier generated

statistical language model based solely on the original low-frequency utterance data segment.

[0043] Process 200 continues the tuning phase by subdividing the utterance data into high-frequency and low-frequency sub-segments using different frequency thresholds

5 until each of the generated language models (from blocks 220 and 224) exceeds the predetermined performance threshold at block 244 or until a desired number of grammar-based and/or statistical language models have been generated. The pattern of frequency thresholds may be predetermined or may be based upon the performance scores of the two models.

10 **[0044]** Once all of the generated language models exceed the relevant performance threshold (block 244), process 200 continues with a ranking phase. A ranking phase is depicted in FIG. 5 as process 400. The ranking phase, which can be executed using the system of FIGs. 1 or 3, can be executed concurrently with the process of FIG. 2. The ranking process 400 begins at block 402 with obtaining test utterance data. Each of the
15 models in the family of grammar-based language models and the family of statistical language models then analyze the test utterance data at block 404.

[0045] Each of the models in the family of grammar-based language models 124 and the family of statistical language models 128 is associated with a recognizer which at block 406 generates a recognized output and a confidence score for each of the respective
20 models in the family of grammar-based language models and the family of statistical language models. At block 408, the recognized output of each of the models in the family of grammar-based language models and the family of statistical language models

is compared to the actual test utterance transcribed manually (aka, the reference sentence of the same speech utterance).

[0046] If at block 408 the recognized output of a particular model is correct, i.e. it matches the test utterance data, then the process 400 continues at block 410 and the confidence score generated by the associated recognizer for the analysis of the test utterance data is used to define range of confidence scores associated with correctly recognizing test data utterances.

[0047] If at block 408 the recognized output of a particular model is not correct, i.e. it does not match the test utterance data, then the process 400 continues at block 414 and the confidence score generated by the associated recognizer for the analysis of the test utterance data is rejected.

[0048] Process 400 then continues at block 412 and the ranges of confidence scores associated with correctly recognizing test data utterances for each of the models in the family of grammar-based language models and the family of statistical language models is assessed to determine whether or not a desired range of confidence scores associated with correctly recognizing test data utterances has been generated. This assessment may be based upon, for example, a threshold number of correctly recognized test utterances. If a respective range has been adequately defined at block 412, the process terminates at block 416.

[0049] If additional confidence scores are needed in order to establish, for each of the models, a respective desired range of confidence scores associated with correctly recognized test data utterances, then the process continues at block 402. If desired, all of

the models may be used to assess additional test utterance data. Alternatively, only a subset of models may be used to assess additional test utterance data.

[0050] Once a respective desired range of confidence scores associated with correctly recognized test data utterances has been generated for each of the models in the family of grammar-based language models and the family of statistical language models, the ranges
5 are stored in the result ranking module 130 of FIG. 1. The system 100 may then be used in an operational mode to recognize unknown utterance data.

[0051] One process for recognizing new speech utterance is depicted in FIG. 6. The recognition process 500 of FIG. 6 begins when the system 100 receives the speech
10 utterance for recognition (block 260). In system 100, the audio input 132 receives the speech utterance using, for example, a microphone to receive a spoken user natural language request or a command. Once the speech data are received, process 500 performs speech recognition on the speech data using each of the generated grammar-based and statistical language models (block 264). As described above, all the generated language
15 model are used in the speech recognition process. In system 100, processor 104 performs speech recognition using some or all of the generated language models concurrently on the processing cores 108.

[0052] For each of the various grammar-based and statistical language models, the speech recognizer with that model produces a result for the input speech data. If the
20 results generated by all of the language models are same (block 268), then the final speech recognition result is simply the same recognition result (block 272).

[0053] In situations where the results from different language models produce two or more different results, process 200 selects a result using the result ranking module 130

(block 276). Specifically, as noted above, each of the language models has an associate recognizer which produces a recognized output . The recognizers also generate a confidence score associated with the recognized output. The processor 104 executes instructions from the result ranking module 130 to analyze the generated confidence score for each of the models against the range of confidence scores associated with correctly recognized test data utterances (the “acceptable range of confidence scores”) in block 278.

[0054] If the generated confidence score of a result from a particular language model is lower than the lowest confidence score in the acceptable range of confidence scores, then the recognized output of that model is discarded. If all of the confidence scores for all of the language models are lower than the lowest confidence score in the respective acceptable range of confidence scores, then the system flags the result as unrecognized.

[0055] If only one confidence score of all of the language models is higher than the lowest confidence score in the respective acceptable range of confidence scores, then the recognized output associated with that confidence score is identified as the recognized output for the system 100.

[0056] If more than one confidence score of all of the language models is higher than the lowest confidence score in the respective acceptable range of confidence scores, then the system 100 ranks the confidence scores of the results from the remaining models. In one embodiment, the confidence scores of the remaining models are ranked as a percentage of the highest confidence score in the acceptable range of confidence scores for the respective model. Thus, if model M_1 has a confidence score of 75 and an acceptable range of confidence scores of 60-75, then model M_1 would have a 100%

relative confidence score or ranking, derived from $(75-60)/(75-60) = 100\%$. If model M_2 has a confidence score of 90 and an acceptable range of confidence scores of 75-100, then model M_2 would have a relative confidence score or ranking, derived from $(90-75)/(100-75) = 15/25 = 60\%$. Accordingly, the output of model M_1 would be selected as the recognized output for the system 100 because M_1 has a higher relative confidence score. Ranking percentages of greater than 100% are possible. In another embodiment, the relative confidence score can be calculated by a weighted formula, where the weights are proportional to the counts of the confidence scores or the counts of the confidence score bins in the confidence score range.

[0057] In the embodiment, of FIG. 1, the family of grammar-based language models 124 and the family of statistical language models 128 are stored within the same memory 112 and executed by the same processor 104. In another embodiment, one or more models in a family of language models 124/128 may be stored within a device while the models in the other family of models 124/128 is stored remotely in a cloud computing system and accessed over a network. In some of these embodiments, two separate systems which may be identical to the system 100, but including only a single family of models 124/128, are used to generate recognized outputs from the models along with the associated confidence scores.

[0058] FIG. 7, by way of example, depicts a cloud computing system 600. System 600 includes a remote system 602 and a client device 604. The remote system 602 includes a processor 606, memory 608, and audio input 610. The memory 608 is configured to hold one or more stored programs that provide programmed instructions that are executed by the processor 606. The programs include a family of grammar-based language models

612 and a family of statistical language models 614, and a result ranking module 616.

The remote system 602 is thus substantially identical to the system 100. The system 600 in some embodiments, however, has substantially greater computing power and resources (e.g., memory) than the corresponding components in system 100. In some

5 embodiments, the remote system 602 does not include one or more of the audio input 610, grammar-based language models 612, and result ranking module 616.

[0059] The client device 604 is operably connected to the remote system 602 by a network 620. The client device 604 includes a processor 626, memory 628, and audio input 630. The memory 628 is configured to hold one or more stored programs that
10 provide programmed instructions that are executed by the processor 626. The programs include a family of grammar-based language models 632, and a result ranking module 636.

[0060] The client device 604, which in this embodiment is an in-vehicle entertainment system, is thus substantially identical to the system 100. The system 604, however, has
15 substantially less computing power and resources (e.g., memory) than the corresponding components in the remote system 602.

[0061] The system 600 operates in much the same manner as the system 100. Some differences, however, include the manner in which speech recognition is divided. Specifically, in the system 600, the client device 604 receives an audio input through the
20 audio input 630. A digital form of the input is then sent to the processor 606 which performs speech recognition using the family of statistical language models 614. Concurrently, the processor 626 performs speech recognition using the family of grammar based language models 628. Because of the greater computing power of the

remote system 602, both speech recognition analyses are completed in less time than would be necessary for the processor 626 to complete both speech recognition analyses.

[0062] The remote system 602 then passes the results and confidence scores from the analysis using the statistical language models 614 to the processor 626 and the processor
5 626 performs a result ranking analysis as described above.

[0063] Accordingly, a small footprint family of models, typically grammar-based language models 632, may be used by the speech recognition engine in a thin client device such as, an in-vehicle entertainment system, an end-user computer system, a mobile computing device (such as personal digital assistance, mobile phone, smartphone,
10 tablet, laptop, or the like), a consumer electronic device, a gaming device, a music player, a security system, a network server or server system, a telemedicine system, or any combination or portion thereof. At the same time, the other language models, such as the large statistical language model 128, may be used by a speech recognition engine in one or more powerful servers, for example, in a cloud computing system accessed over a
15 network. The output results from both speech recognition engines can be combined based on confidence scores as described above.

[0064] While the invention has been illustrated and described in detail in the drawings and foregoing description, the same should be considered as illustrative and not restrictive in character. It is understood that only the preferred embodiments have been
20 presented and that all changes, modifications and further applications that come within the spirit of the invention are desired to be protected.

What is claimed is:

Claim 1. A method of generating language models for speech recognition comprising:

- 5 identifying a plurality of utterances in training data corresponding to speech;
generating a frequency count of each utterance in the plurality of utterances;
generating a high-frequency plurality of utterances from the plurality of
utterances having a frequency that exceeds a predetermined frequency threshold;
generating a low-frequency plurality of utterances from the plurality of utterances
10 having a frequency that is below the predetermined frequency threshold;
generating a grammar-based language model using the high-frequency plurality of
utterances as training data; and
generating a statistical language model using the low-frequency plurality of
utterances as training data.

15

Claim 2. The method of claim 1, further comprising:

- performing a plurality of speech recognition experiments with the grammar-based
language model;
identifying a performance of the grammar-based language model in the speech
20 recognition experiments;
generating a second low-frequency plurality of utterances and second high-
frequency plurality of utterances from the high-frequency plurality of utterances when the

performance of the grammar-based language model is below a predetermined performance threshold;

generating a second grammar-based language model using the second high-frequency plurality of utterances as training data; and

5 generating a second statistical language model using the second low-frequency plurality of utterances as training data.

Claim 3. The method of claim 1, further comprising:

10 performing a plurality of speech recognition experiments with the statistical language model;

 identifying a performance of the statistical language model in the speech recognition experiments;

15 generating a second low-frequency plurality of utterances and second high-frequency plurality of utterances from the low-frequency plurality of utterances when the performance of the statistical language model is below a predetermined performance threshold;

 generating a second grammar-based language model using the second high-frequency plurality of utterances as training data; and

20 generating a second statistical language model using the second low-frequency plurality of utterances as training data.

Claim 4. The method of claim 1, further comprising:

performing a plurality of speech recognition experiments with the grammar-based language model;

identifying a performance of the grammar-based language model in the speech recognition experiments;

5 performing the plurality of speech recognition experiments with the statistical language model;

identifying a performance of the statistical language model in the speech recognition experiments; and

10 increasing the predetermined frequency threshold when the performance of the grammar-based language model is below a predetermined performance threshold and the performance of the statistical language model is below another predetermined performance threshold.

Claim 5. The method of claim 1, wherein a first processor core performs at least a portion of the generation of the grammar-based language model concurrently with a second processor core that performs at least a portion of generation of the statistical language model.

Claim 6. The method of claim 1 further comprising:

20 performing a speech recognition operation on speech data using the grammar-based language model to generate a first speech recognition result;

performing a speech recognition operation on the speech data using the statistical language model to generate a second speech recognition result; and

identifying a final speech recognition result as either of the first speech recognition result or the second speech recognition result when the first speech
5 recognition result is equivalent to the second speech recognition result.

Claim 7. The method of claim 6, the identification of the final speech recognition result further comprising:

identifying a first relative confidence score of the first speech recognition result;

10 identifying a second relative confidence score of the second speech recognition result;

identifying the first speech recognition result as the final speech recognition result when the first relative confidence score is greater than the second relative confidence score; and

15 identifying the second speech recognition result as the final speech recognition result when the second relative confidence is greater than the first relative confidence score.

Claim 8. The method of claim 6, wherein a first processor core performs at least a
20 portion of the speech recognition operation using the grammar-based language model concurrently with a second processor core that performs at least a portion of the speech recognition operation using the statistical language model.

Claim 9. A method of performing speech recognition comprising:

collecting a plurality of utterances;

generating a frequency count of each utterance in the plurality of utterances;

identifying a high-frequency segment of the plurality of utterances based upon a

5 predetermined frequency threshold;

identifying a low-frequency segment of the plurality of utterances based upon the
predetermined frequency threshold;

generating a family of grammar-based language models based on at least a portion
of the high-frequency segment;

10 generating a family of statistical language models based on at least a portion of
the low-frequency segment;

performing a first speech recognition using the family of grammar-based language
models;

15 performing a second speech recognition using the family of statistical language
models; and

determining a recognized speech based upon the first speech recognition and the
second speech recognition.

Claim 10. The method of claim 9, wherein at least one of the family of grammar-

20 based language models and the family of statistical language models includes more than
one language model.

Claim 11. An intelligent speech recognition system comprising:

at least one audio input;

at least one memory;

a family of grammar-based language models stored within the at least one

5 memory;

a family of statistical language models stored within the at least one memory; and

at least one processor operably connected to the at least one audio input and the at

least one memory and configured to (i) perform a first speech recognition using the

family of grammar-based language models, (ii) perform a second speech recognition

10 using the family of statistical language models, and (iii) determine a recognized speech

based upon the first speech recognition and the second speech recognition, wherein

the family of grammar-based language models is generated based upon a

generated frequency count of each utterance in a plurality of utterances identified as

being in a high-frequency segment of the plurality of utterances based upon a

15 predetermined frequency threshold; and

the family of statistical language models is generated based upon a generated

frequency count of each utterance in a plurality of utterances identified as being in a low-

frequency segment of the plurality of utterances based upon the predetermined frequency

threshold.

20

Claim 12. The system of claim 11, wherein:

the family of statistical language models includes a plurality of statistical

language models; and

the family of grammar-based language models includes a plurality of grammar-based language models.

Claim 13. The system of claim 11, wherein at least a portion of the intelligent speech
5 recognition system is located within or remotely from one or more client devices.

Claim 14. The system of claim 13, wherein the client device is selected from a group
consisting of an in-vehicle entertainment system, an end-user computer system, a mobile
computing device, an electronic device, a gaming device, a music player, a security
10 system, a network server or server system, and a telemedicine system.

Claim 15. The system of claim 11, wherein the at least one audio input comprises a
microphone.

15 Claim 16. The system of claim 11, further comprising:

a plurality of recognizers, each of the plurality of recognizers associated with a
respective one of the models in the family of grammar-based language models and the
family of statistical language models, and configured to generate a recognized output and
a confidence score associated with the recognized output based upon a model output from
20 the respective model in the family of grammar-based language models or the family of
statistical language models, the model output based upon utterance data provided to the
respective model.

Claim 17. The system of claim 16, wherein the at least one processor is configured to:

determine if the recognized output for a first utterance data is correct;

save the confidence score associated with each recognized output for the first

5 utterance data that is correct; and

define a respective range of confidence scores for each of the family of grammar-based language models and the family of statistical language models based upon the saved confidence scores.

10 Claim 18. The system of claim 17, wherein the at least one processor is further configured to:

obtain a second confidence score associated with each recognized output for a second utterance data;

compare the obtained second confidence score associated with each recognized
15 output to the respective range of confidence scores; and

determine the recognized speech based upon the comparison of the obtained second confidence scores.

Claim 19. The system of claim 11, wherein:

20 the family of grammar-based language models is stored within one of a client device and a cloud computing system operably connected to the client device through a network; and

the family of statistical language models is stored within the other of the client device and the cloud computing system.

Claim 20. The system of claim 19, wherein the family of grammar-based language
5 models is stored within the client device.

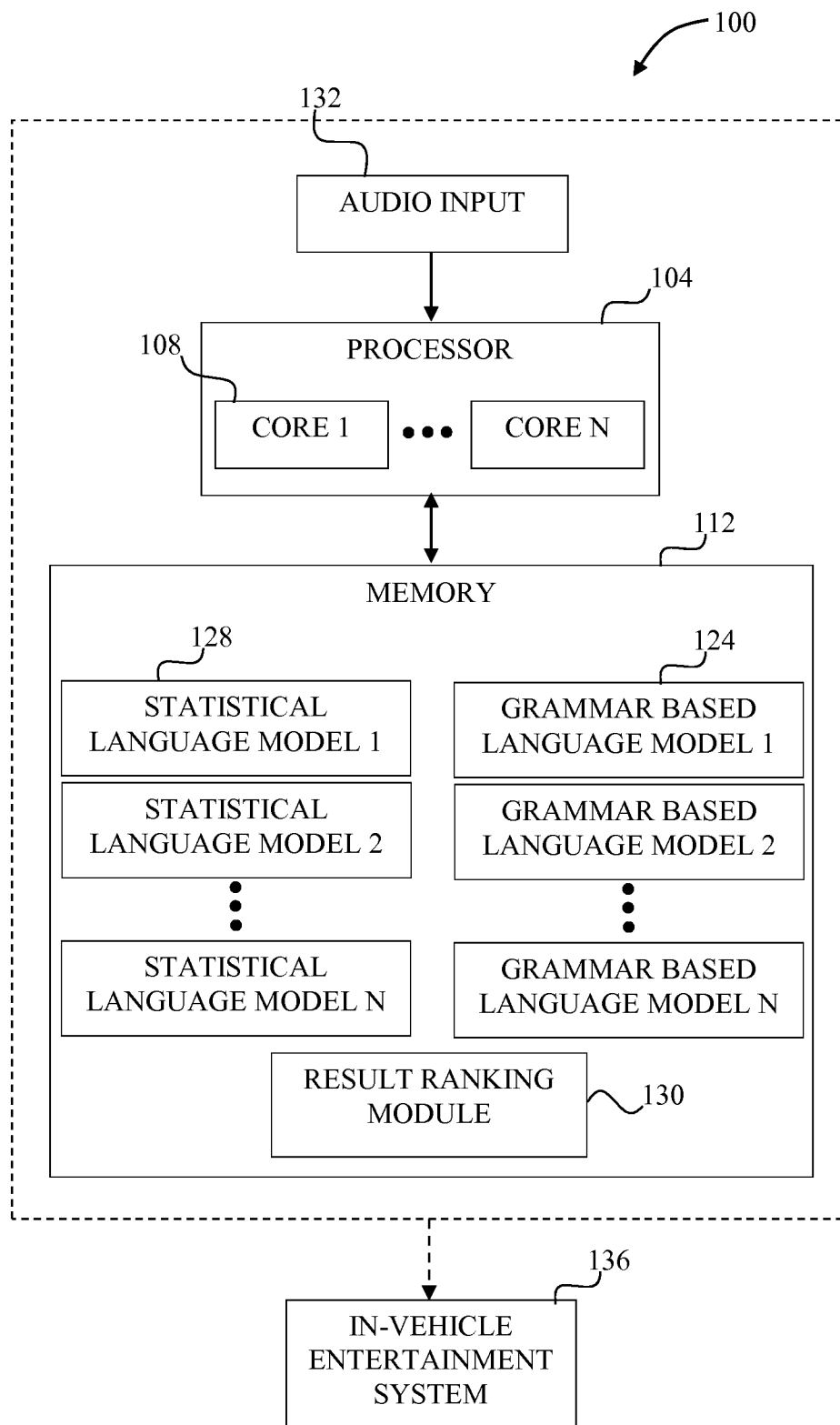


FIG. 1

Sheet 2/7

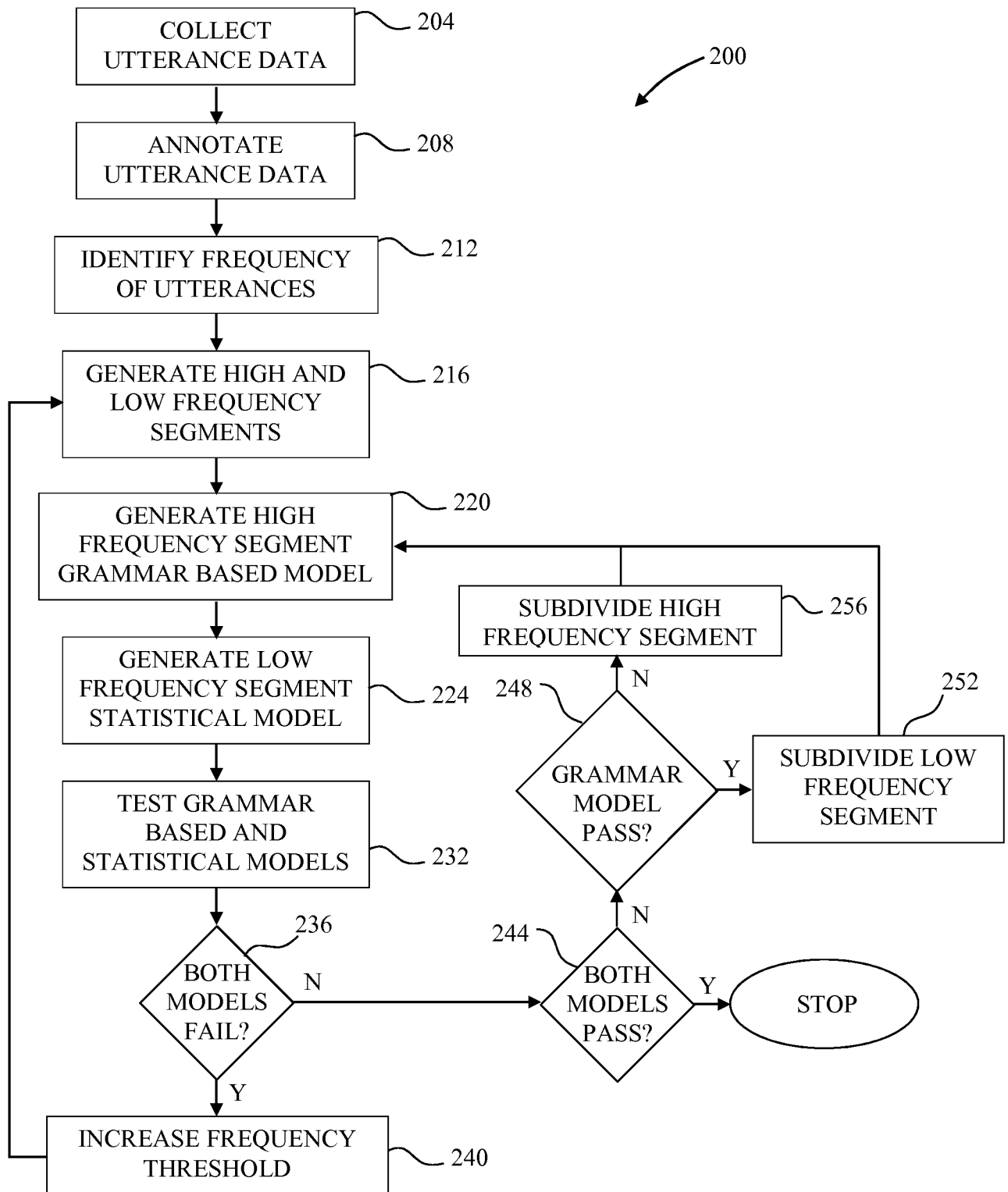


FIG. 2

Sheet 3/7

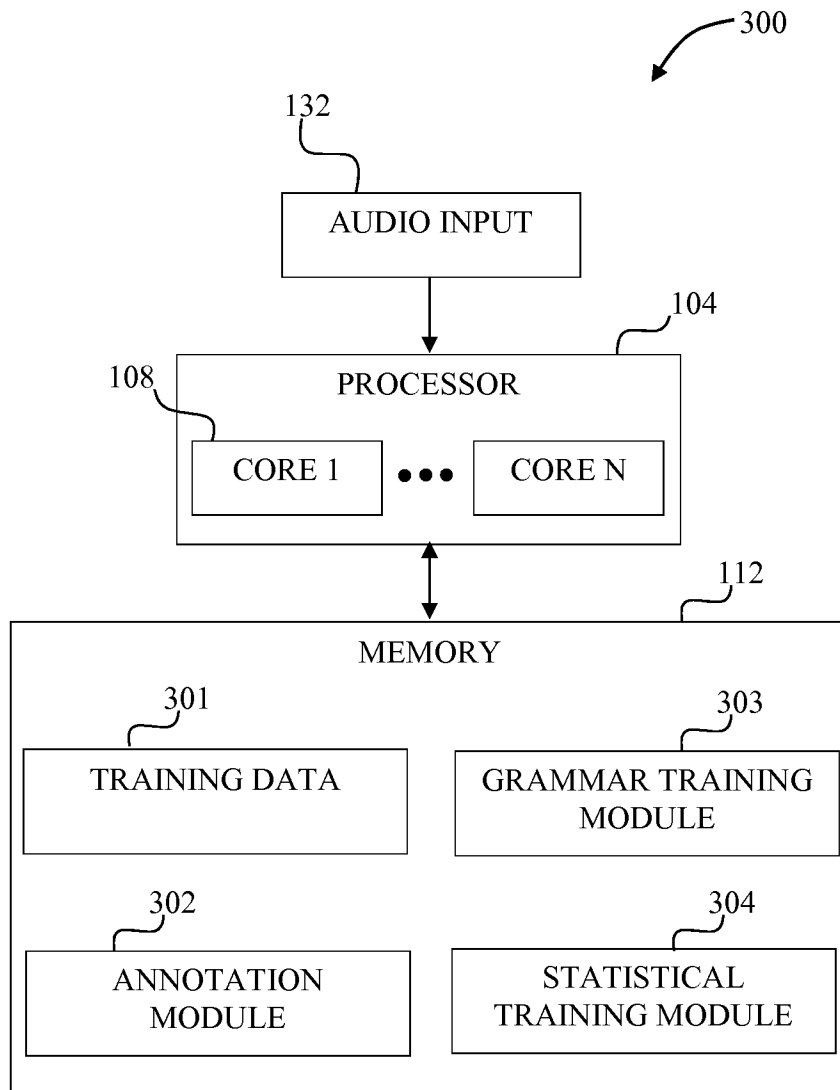


FIG. 3

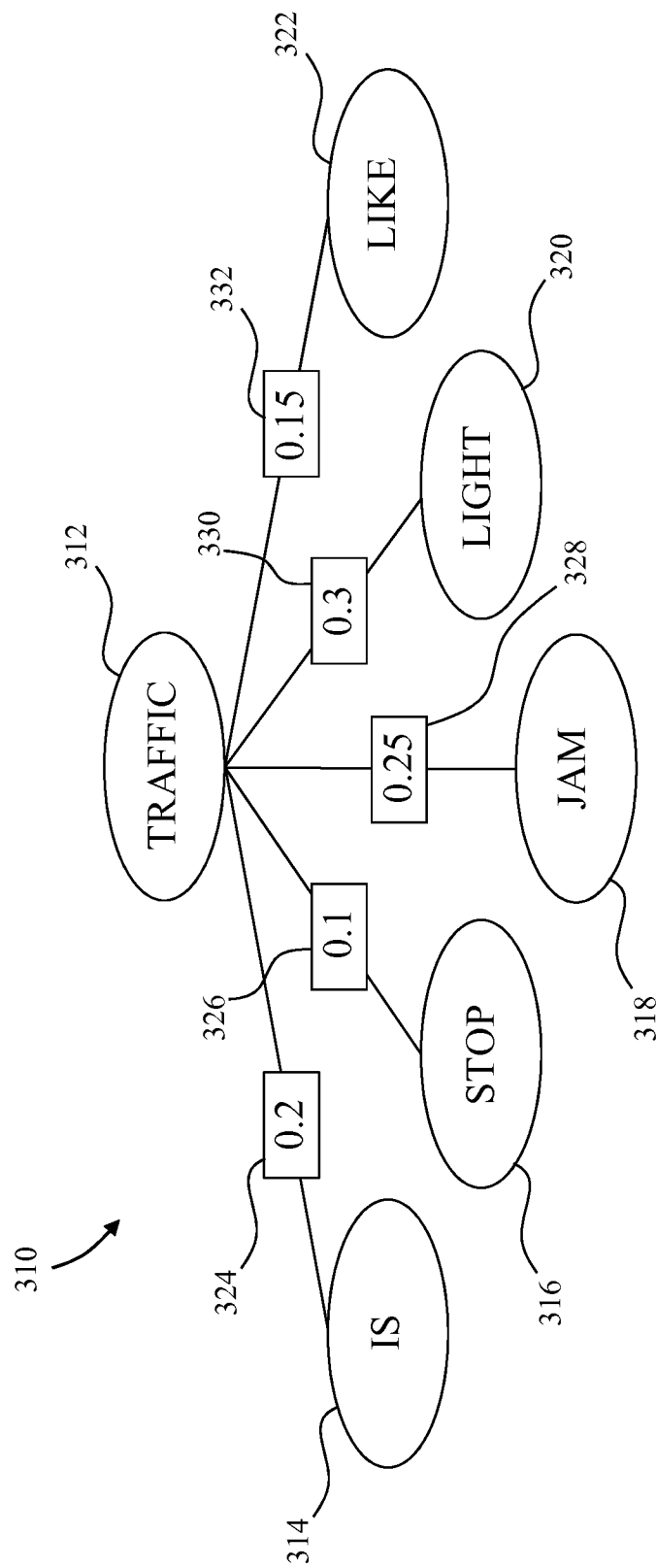


FIG. 4

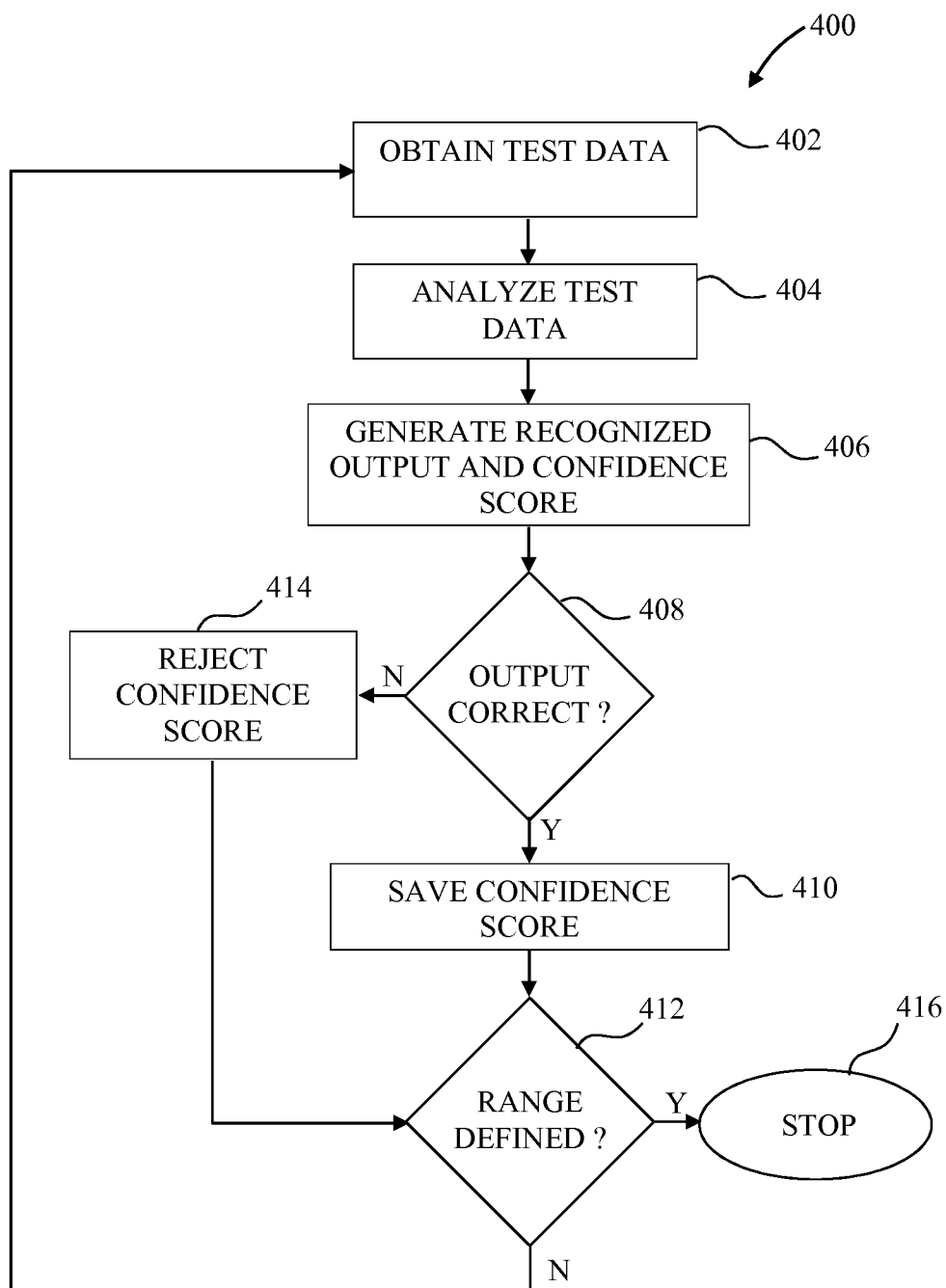


FIG. 5

Sheet 6/7

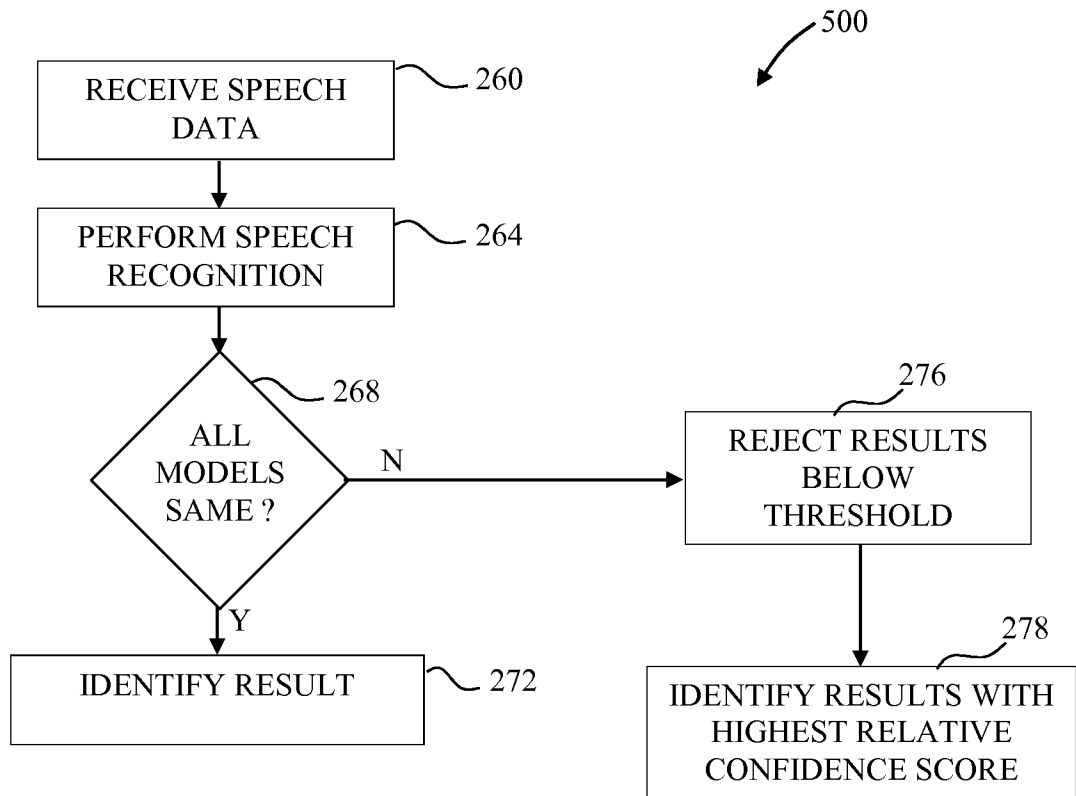


FIG. 6

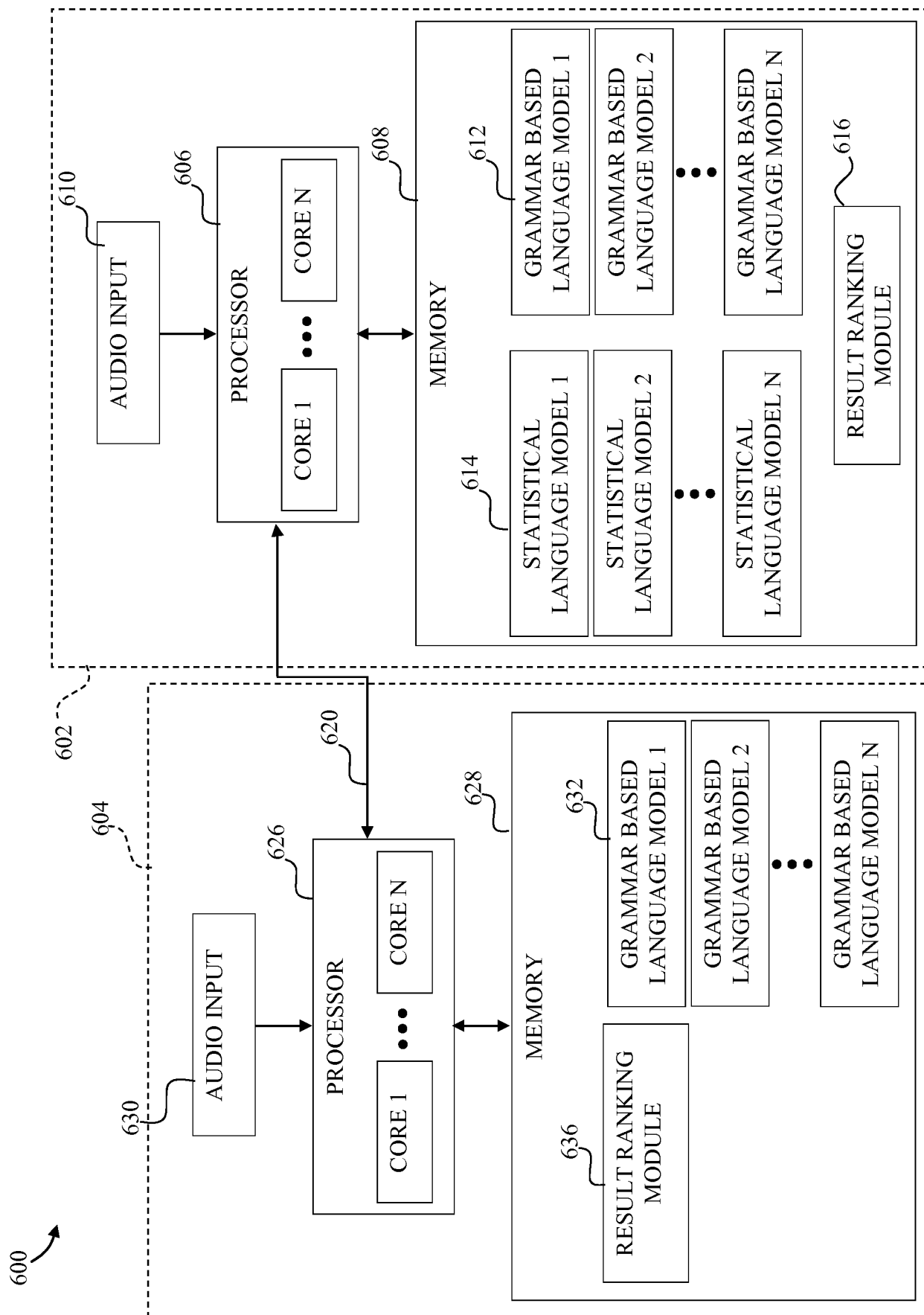


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2012/034261

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/06 G10L15/28 G10L15/18
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L G06F

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

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

EPO-Internal, IBM-TDB, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 2 034 472 A1 (SIEMENS AG [DE]) 11 March 2009 (2009-03-11)	11-20
A	page 6, lines 1-35; figures 3,4 -----	1,9
A	US 2004/148154 A1 (ACERO ALEJANDRO [US] ET AL) 29 July 2004 (2004-07-29) figure 5 paragraph [0007] paragraph [0063] paragraphs [0068], [0069] paragraphs [0073] - [0075] -----	1,9,11
A	US 2010/250240 A1 (SHU CHANG-QING [US]) 30 September 2010 (2010-09-30) figures 3,4 paragraphs [0008], [0022] paragraphs [0029] - [0047] -----	1,9,11



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

25 June 2012

Date of mailing of the international search report

06/07/2012

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Ramos Sánchez, U

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2012/034261

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
EP 2034472	A1	11-03-2009	DE 102007042971 A1	12-03-2009
			EP 2034472 A1	11-03-2009

US 2004148154	A1	29-07-2004	US 2004148154 A1	29-07-2004
			US 2004148170 A1	29-07-2004

US 2010250240	A1	30-09-2010	NONE	
