US008392188B1

(12) **United States Patent**
Riccardi

(10) **Patent No.: US 8,392,188 B1**
(45) **Date of Patent: Mar. 5, 2013**

(54) **METHOD AND SYSTEM FOR BUILDING A PHONOTACTIC MODEL FOR DOMAIN INDEPENDENT SPEECH RECOGNITION**

(75) Inventor: **Giuseppe Riccardi**, Hoboken, NJ (US)

(73) Assignee: **AT&T Intellectual Property II, L.P.**, Atlanta, GA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1321 days.

(21) Appl. No.: **09/956,907**

(22) Filed: **Sep. 21, 2001**

Related U.S. Application Data

(63) Continuation-in-part of application No. 09/690,721, filed on Oct. 18, 2000, now Pat. No. 7,085,720, and a continuation-in-part of application No. 09/690,903, filed on Oct. 18, 2000, now Pat. No. 6,681,206.

(60) Provisional application No. 60/235,861, filed on Sep. 27, 2000, provisional application No. 60/163,838, filed on Nov. 5, 1999.

(51) **Int. Cl.**
G10L 17/00 (2006.01)

(52) **U.S. Cl.** **704/249**; 704/244; 704/243

(58) **Field of Classification Search** 704/244, 704/231, 270, 250, 243, 245, 247, 249
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,777,600	A	10/1988	Saito et al.
4,827,521	A	5/1989	Bahl et al.
4,866,778	A	9/1989	Baker
4,882,759	A	11/1989	Bahl et al.
4,903,305	A	2/1990	Gillick et al.
5,005,127	A	4/1991	Kugimiya et al.
5,029,214	A	7/1991	Hollander
5,033,088	A	7/1991	Shipman
5,062,047	A	10/1991	Tanaka et al.
5,099,425	A	3/1992	Yuji et al.
5,210,689	A	5/1993	Baker
5,212,730	A	5/1993	Wheatley et al.
5,297,039	A	3/1994	Kanaegami et al.
5,323,316	A	6/1994	Kadashevich et al.
5,333,275	A	7/1994	Wheatley et al.
5,337,232	A	8/1994	Sakai et al.
5,357,596	A	10/1994	Takebayashi et al.
5,375,164	A	12/1994	Jennings
5,384,892	A	1/1995	Strong
5,390,272	A	2/1995	Repta et al.
5,434,906	A	7/1995	Robinson et al.
5,457,768	A	10/1995	Tsuboi et al.
5,500,920	A	3/1996	Kupiec
5,509,104	A	4/1996	Lee et al.
5,544,050	A	8/1996	Abe et al.
5,619,410	A	4/1997	Emori et al.
5,642,519	A	6/1997	Martin
5,651,095	A	7/1997	Ogden
5,659,731	A	8/1997	Gustafson
5,666,400	A	9/1997	McAllister et al.
5,675,707	A	10/1997	Gorin et al.

(Continued)

FOREIGN PATENT DOCUMENTS

WO	WO 00/62193	10/2000

OTHER PUBLICATIONS

U.S. Appl. No. 09/690,721, filed Oct. 18, 2000.*

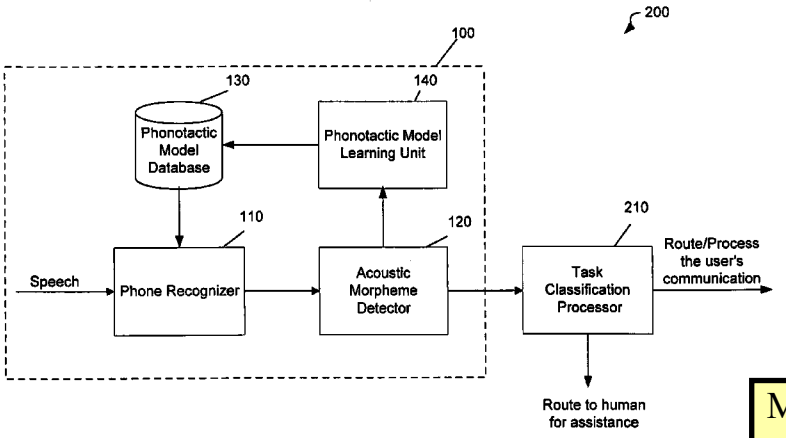
(Continued)

Primary Examiner — Huyen X. Vo

(57) **ABSTRACT**

The invention concerns a method and corresponding system for building a phonotactic model for domain independent speech recognition. The method may include recognizing phones from a user's input communication using a current phonotactic model, detecting morphemes (acoustic and/or non-acoustic) from the recognized phones, and outputting the detected morphemes for processing. The method also updates the phonotactic model with the detected morphemes and stores the new model in a database for use by the system during the next user interaction. The method may also include making task-type classification decisions based on the detected morphemes from the user's input communication.

46 Claims, 7 Drawing Sheets



MICROSOFT CORP.
EXHIBIT 1005

U.S. PATENT DOCUMENTS

5,719,921	A		2/1998	Vysotsky et al.	
5,724,481	A		3/1998	Garberg et al.	
5,724,594	A		3/1998	Pentheroudakis	
5,752,230	A	*	5/1998	Alonso-Cedo	704/270
5,794,193	A	*	8/1998	Gorin	704/250
5,832,480	A		11/1998	Byrd et al.	
5,839,106	A		11/1998	Bellegarda	
5,860,063	A		1/1999	Gorin et al.	
5,878,390	A	*	3/1999	Kawai et al.	704/231
5,905,774	A		5/1999	Tatchell et al.	
5,918,222	A		6/1999	Fukui et al.	
5,960,447	A		9/1999	Holt et al.	
5,983,180	A		11/1999	Robinson	
6,006,186	A		12/1999	Chen et al.	
6,021,384	A		2/2000	Gorin et al.	
6,023,673	A		2/2000	Bakis et al.	
6,044,337	A		3/2000	Gorin et al.	
6,064,957	A		5/2000	Brandow et al.	
6,138,085	A		10/2000	Richardson et al.	
6,173,261	B1		1/2001	Arai et al.	
6,192,110	B1		2/2001	Abella et al.	
6,208,964	B1	*	3/2001	Sabourin	704/244
6,233,553	B1		5/2001	Contolini et al.	
6,243,683	B1		6/2001	Peters	
6,308,156	B1		10/2001	Barry et al.	
6,311,152	B1		10/2001	Bai et al.	
6,317,707	B1		11/2001	Bangalore et al.	
6,397,179	B2		5/2002	Crespo et al.	
6,601,026	B2		7/2003	Appelt et al.	
6,681,206	B1		1/2004	Gorin et al.	
6,745,161	B1		6/2004	Arnold et al.	
6,879,956	B1		4/2005	Honda et al.	
6,895,377	B2		5/2005	Kroeker et al.	
6,941,266	B1		9/2005	Gorin et al.	
6,961,954	B1		11/2005	Maybury et al.	
7,082,578	B1		7/2006	Fishkin et al.	
7,085,720	B1		8/2006	Gorin et al.	
7,286,984	B1		10/2007	Gorin et al.	
7,440,897	B1		10/2008	Gorin et al.	
7,620,548	B2		11/2009	Gorin et al.	
8,010,361	B2		8/2011	Gorin et al.	
2004/0199375	A1		10/2004	Ehsani et al.	

OTHER PUBLICATIONS

A.L. Gorin, "Processing of Semantic Information in Fluently Spoken Language", to appear in the Proceedings of the Int;\l. Conf. on Spoken Language Processing (CSLP), Oct. 1996, AT&T Research, pp. 1-4.

A.L. Gorin et al., "Spoken Language Acquisition for Automated Call Routing in a Telephone Network", Proceedings 1993 IEEE Workshop on Automatic Speech Recognition, Dec. 12-15, 1993, p. 99-100.

A.L. Gorin et al.. "Spoken Dialog as a Feedback control System", ESCA Workshop on Spoken Dialogue Systems—Theories and Applications, Vigso, Denmark, May 30-Jun. 2, 1995, AT&T Bell Laboratories, pp. 173-175.

A.L. Gorin et al., "Spoken Language Acquisition for Automated Call Routing", ICSLP 94, 1994 International Conference of Spoken Language Processing, Sep. 18-22, 1994, Pacific Convention Plaza Yokohama, pp. 1483-1486.

A.L. Gorin et al., "On Automated Language Acquisition" J. Accoust. Soc. Am. 97(6) Jun. 1995, pp. 3441-3461.

A. L. Gorin et al., "An Experiment in Spoken Language Acquisition", IEEE Transactions on Speech and Audio Processing, vol. 2, No. 1, Par II, Jan. 1994, pp. 224-240.

A.L. Gorin et al., "Automated Call Routing in a Telecommunications Network", Proceedings, Second IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Sep. 26-27, 1994, pp. 137-140.

R. Cole et al., "The Challenge of Spoken Language Systems: I Research Direction for the Nineties", IEEE Transactions on Speech and Audio Processing; Jan. 1995, vol. 3, Issue 1, pp. 1-21.

L.G. Miller et al., "A Structured Network Architecture for Adaptive Language Acquisition", ICASSP '92: Acoustics, Speech & Signal Processing Conference; 1992 vol. 1, I-201-I-204.

A.L. Gorin et al., "On Adaptive Acquisition of Spoken Language", Neural Networks for Signal Processing, 1991; 1991, pp. 422-431.

Ponceleon et al., "Automatic Discovery of Salient Segments in Imperfect Speech Transcripts", CIKM '01, Atlanta, Georgia, Nov. 5-10, 2001.

A.L. Gorin et al. "How May I Help You?", AT&T Las Research, p. 1-2 and 4, as it was to appear in the Proc. of IVITA, Philadelphia, Oct. 1996.

Irene Langkilde et al., "Automatic Prediction of Problematic Human-Computer Dialogues in How May I Help You?", AT&T Labs Research, pp. 1-4.

U.S. Appl. No. 09/699,494, "Method for Automated Task Classification and Routing", filed Oct. 31, 2000, by A. Gorin et al.

U.S. Appl. No. 09/699,495, "Automated Task Classification System", filed Oct. 31, 2000, by A. Gorin et al.

U.S. Appl. No. 09/699,496, "System and Method for Conducting Dialog with a User in an automatic Task Classification System", filed Oct. 31, 2000, by A. Gorin et al.

U.S. Appl. No. 09/712,192, "Method and System for Predicting Problematic Situations in Automated Dialog", filed Nov. 15, 2000, by A. Gorin et al.

A.L. Gorin et al. U.S. Appl. No. 10/158,082, "Method and System for Automatically Detecting Morphemes in a Task Classification System Using Lattices", filed May 31, 2002.

Aberdeen et al., "Mitre: Description of the Alembic system as used in MET", Proceedings of the TIPSTER 24-Month Workshop, pp. 461-462, 1996.

Abney, Steven, "Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax" University to Tubingen, Oct. 17, 1991.

Humphereys et al. University of Sheffield: Description of the LaSIE-II System as Used for MUC-7 In Proceedings of the Seventh Message Understanding Conferences, 1998.

Rahim et al. "Robust numeric recognition in spoken language dialogue", Speech Communication, vol. 34, 2001, pp. 195-212.

Huang et al. "Information extraction from voicemail", Association for computational Linguistics, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 2001.

Bikel et al. "An Algorithm that learns what's in a Name", Machine Learning, vol. 34, pp. 211-231, Kluwer Academic Publishers, 1999.

Gotoh et al., "Named Entity Tagged Language Models," IEEE International Conference in Acoustics, Speech and Signal Processing, ICASSP '99, Mar. 15-19, 1999, pp. 513-516.

Gaizaaukas et al., "Description of the LaSIE system as used for MUC-6," In Processing of the Sixth Message Understanding Conference (MUC-6), pp. 207-220, Nov. 1995.

* cited by examiner

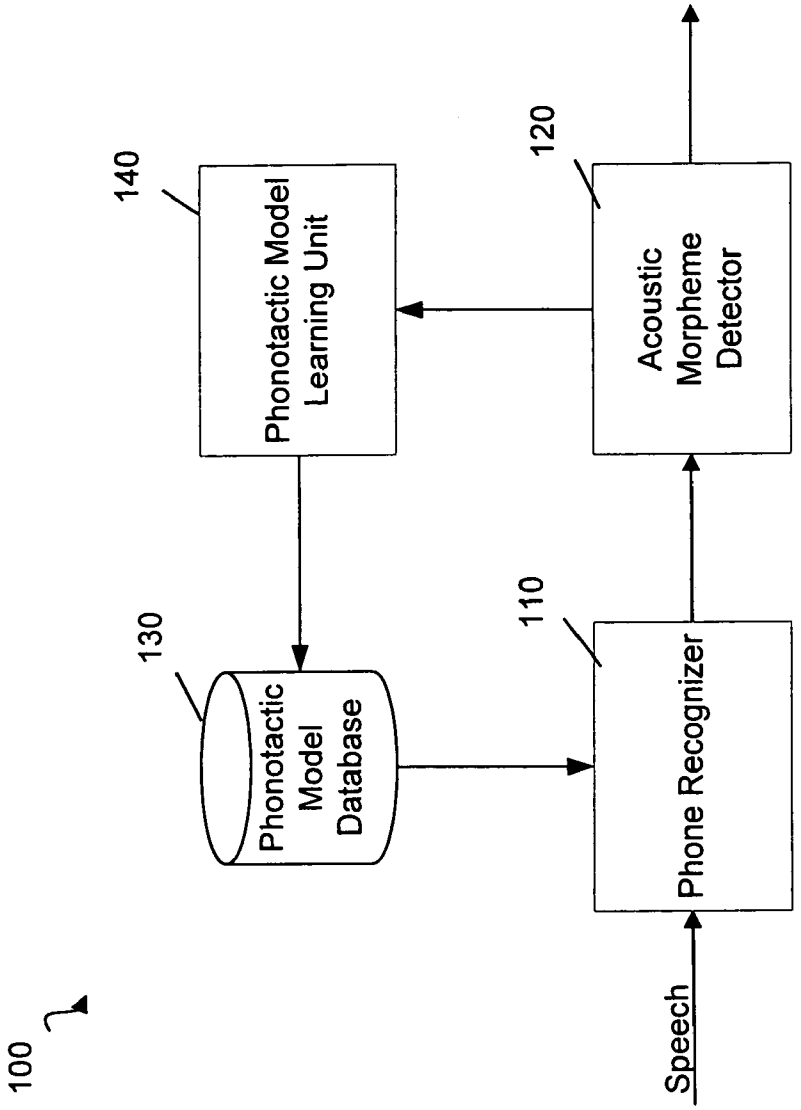

100
Speech
Phone Recognizer
110
Phonotactic Model Database
130
Acoustic Morpheme Detector
120
Phonotactic Model Learning Unit
140

Fig. 1

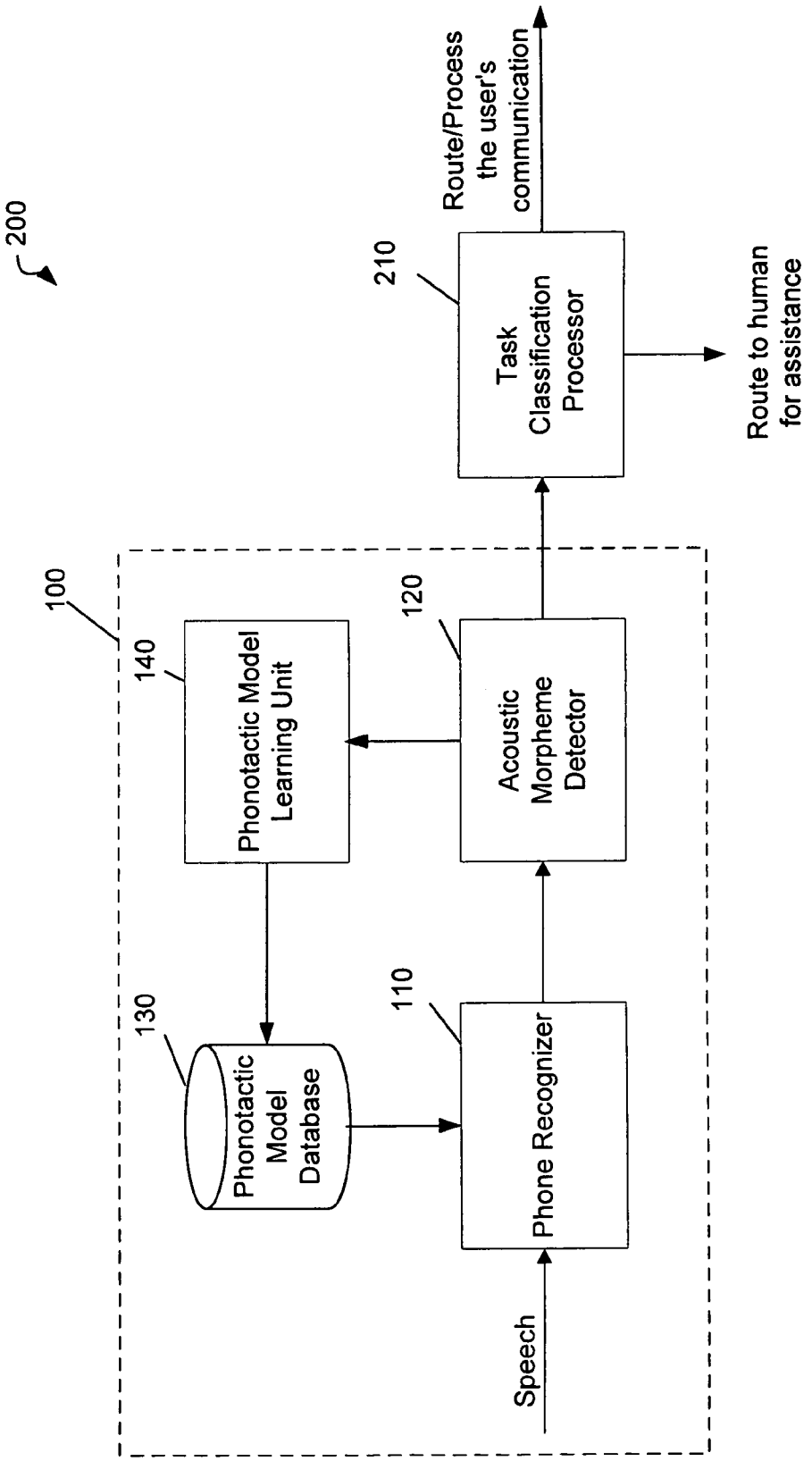

200
100
Speech
Phone Recognizer
110
Phonotactic Model Database
130
Phonotactic Model Learning Unit
140
Acoustic Morpheme Detector
120
Task Classification Processor
210
Route/Process the user's communication
Route to human for assistance

Fig. 2

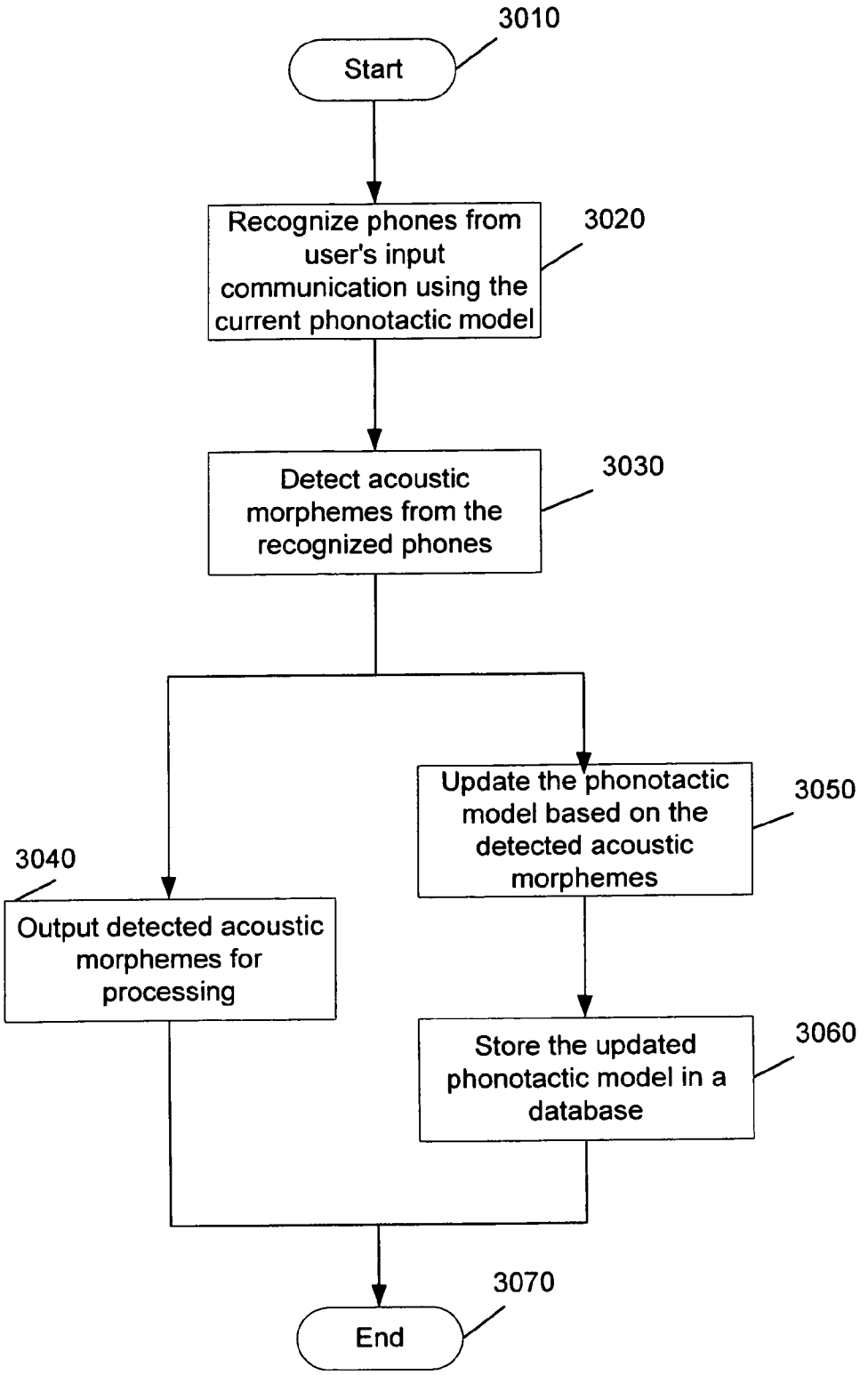

3010
Start
Recognize phones from user's input communication using the current phonotactic model
3020
Detect acoustic morphemes from the recognized phones
3030
Update the phonotactic model based on the detected acoustic morphemes
3050
3040
Output detected acoustic morphemes for processing
Store the updated phonotactic model in a database
3060
3070
End

Fig. 3

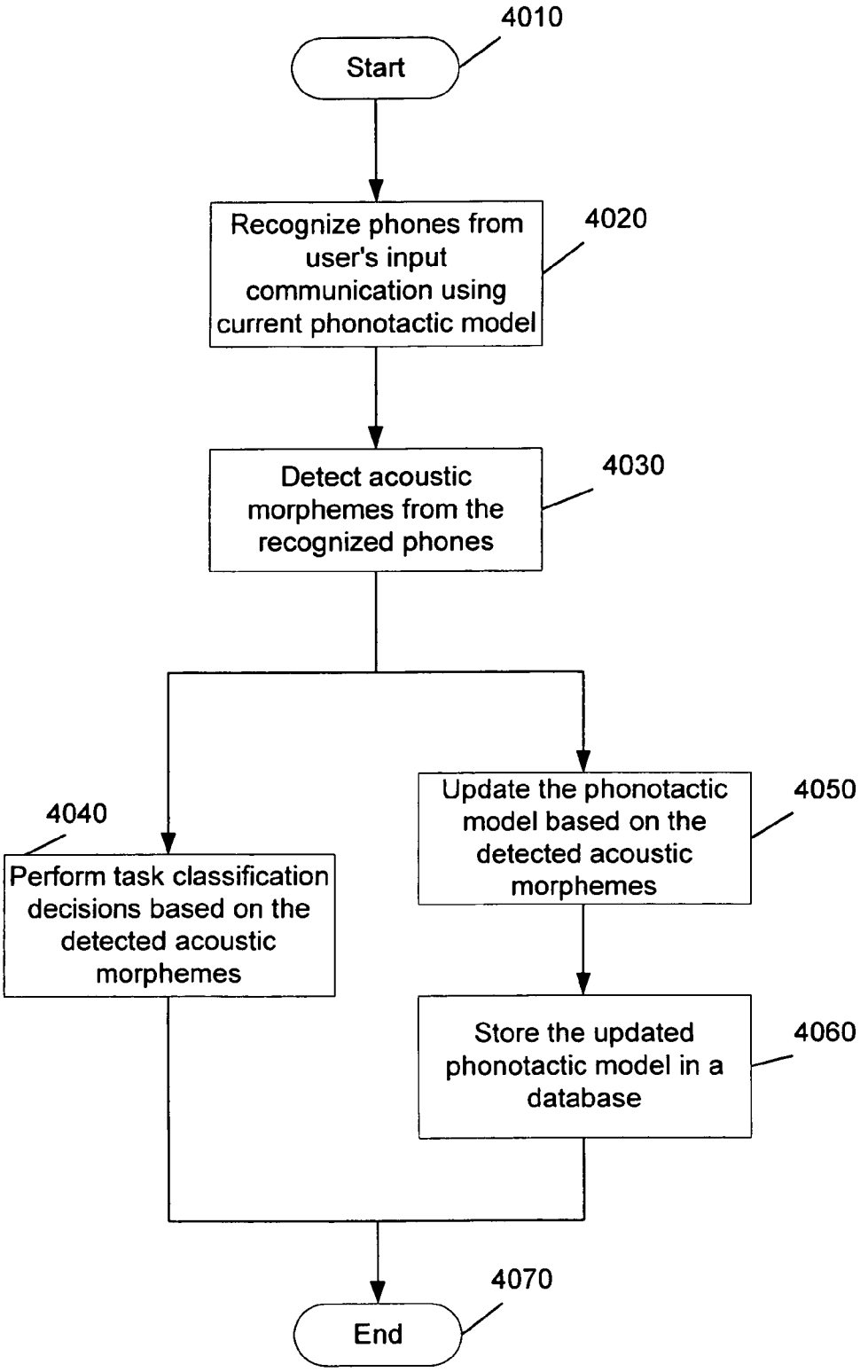

4010
Start
Recognize phones from user's input communication using current phonotactic model
4020
Detect acoustic morphemes from the recognized phones
4030
4040
Perform task classification decisions based on the detected acoustic morphemes
Update the phonotactic model based on the detected acoustic morphemes
4050
Store the updated phonotactic model in a database
4060
4070
End

Fig. 4

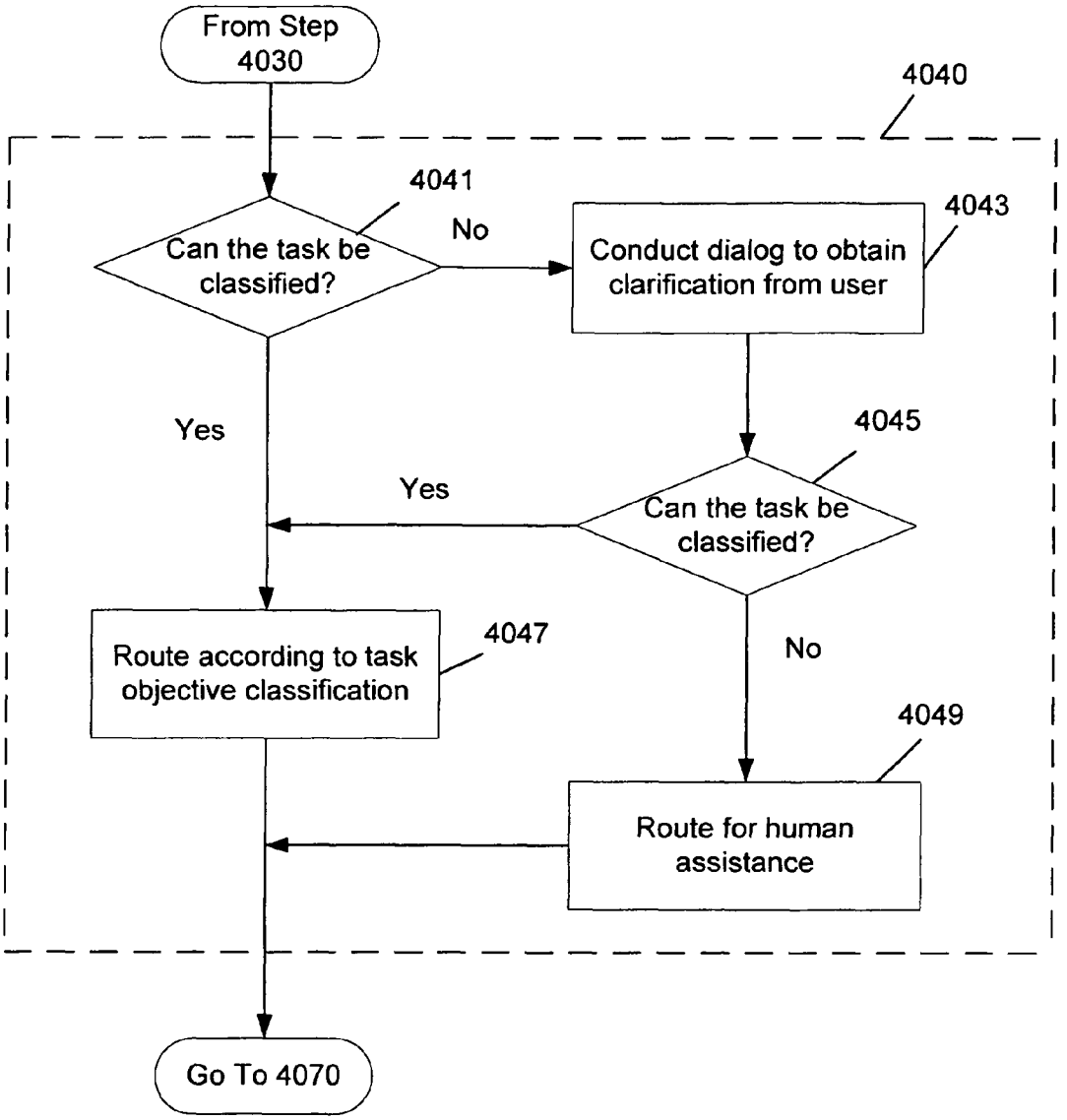

From Step 4030
4040
4041
Can the task be classified?
No
Conduct dialog to obtain clarification from user
4043
Yes
4045
Yes
Can the task be classified?
Route according to task objective classification
4047
No
4049
Route for human assistance
Go To 4070

Fig. 5

MI	WMI		
7.75	0.0611	K ao l	Call
18.08	0.0596	n ah m B er	Number
9.15	0.0621	f ao r	Four
14.57	0.0280	y eh s	Yes
8.31	0.0374	w ah n	One
16.66	0.0355	s eh v ax n	Seven
36.00	0.0277	K ao l ih ng K aa r D	Calling Card
3.63	0.0276	T uw	Two/To
6.12	0.0227	l ay K	Like

Fig. 6

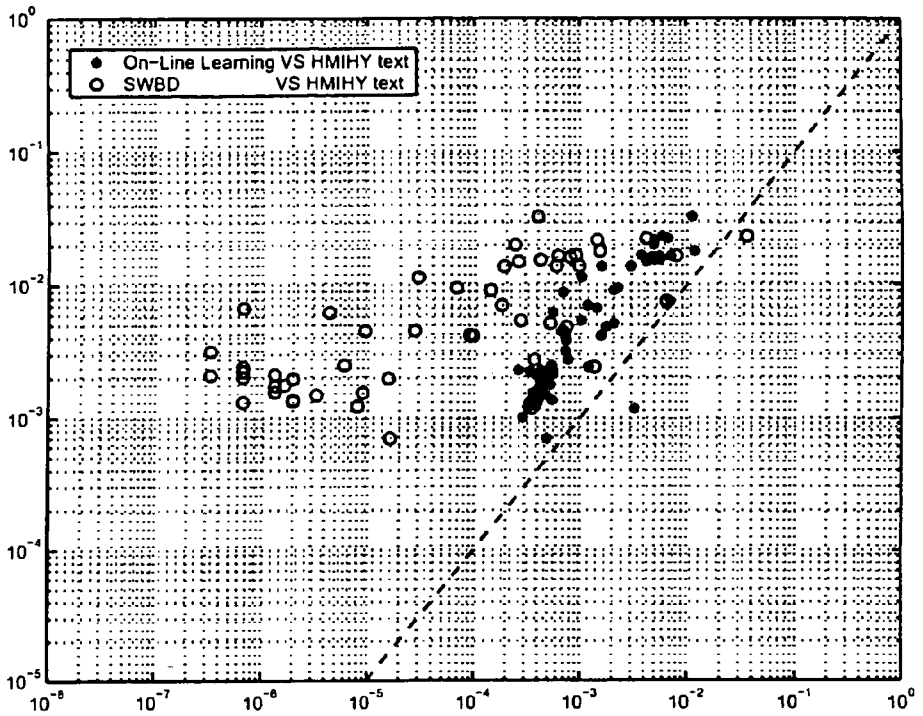

On-Line Learning VS HMIHY text
SWBD VS HMIHY text
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}

Fig. 7

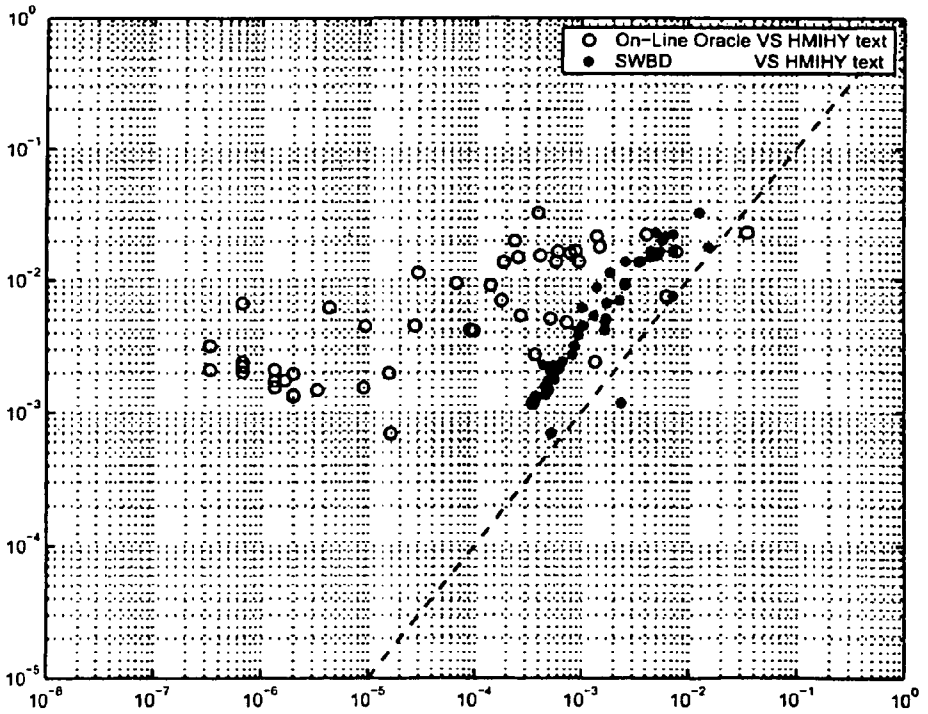

On-Line Oracle VS HMIHY text
SWBD VS HMIHY text
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}

Fig. 8

METHOD AND SYSTEM FOR BUILDING A PHONOTACTIC MODEL FOR DOMAIN INDEPENDENT SPEECH RECOGNITION

CLAIM FOR PRIORITY/CROSS REFERENCE TO RELATED APPLICATIONS

This non-provisional application claims the benefit of U.S. Provisional Patent Application No. 60/235,861, filed Sep. 27, 2000, which is incorporated by reference in its entirety. This non-provisional application is also a continuation-in-part of U.S. patent application Ser. Nos. 09/690,721 now U.S. Pat. No. 7,085,720 and 09/690,903 now U.S. Pat. No. 6,681,206, both filed Oct. 18, 2000, which claim the benefit of U.S. Provisional Patent Application No. 60/163,838, filed Nov. 5, 1999, which are also incorporated herein by reference in their entireties.

TECHNICAL FIELD

The invention relates to automated systems for communication recognition and understanding.

BACKGROUND OF THE INVENTION

Conventional methods for constructing spoken language systems involve training speech and language models prior to operation by transcribing speech input and finding existing patterns. This speech is manually transcribed and each utterance is then semantically labeled. The resultant database is exploited to train stochastic language models for recognition and understanding. These models are further adapted for different dialog states. Examples of such methods are shown in U.S. Pat. Nos. 5,675,707, 5,860,063, 6,044,337, 6,173,261, 6,021,384 and 6,192,110, each of which is incorporated by reference herein in its entirety.

This transcription and labeling process is a major bottleneck in new application development and refinement of existing ones. For incremental training of a deployed natural spoken dialog system, current technology would potentially require transcribing millions of transactions. This process is both time-consuming and prohibitively expensive.

SUMMARY OF THE INVENTION

The invention concerns a method and corresponding system for building a phonotactic model for domain independent speech recognition. The method may include recognizing phones from a user's input communication using a current phonotactic model, detecting morphemes (acoustic and/or non-acoustic) from the recognized phones, and outputting the detected morphemes for processing. The method also updates the phonotactic model with the detected morphemes and stores the new model in a database for use by the system during the next user interaction. The method may also include making task-type classification decisions based on the detected morphemes from the user's input communication.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is described in detail with reference to the following drawings wherein like numerals reference like elements, and wherein:

FIG. **1** is a block diagram of an exemplary speech recognition system of the invention;

FIG. **2** is a block diagram of an exemplary speech recognition system of the invention used in conjunction with a task classification system;

FIG. **3** is a flowchart illustrating an exemplary phonotactic model updating process;

FIG. **4** is a flowchart illustrating an exemplary phonotactic model updating process in a task classification system;

FIG. **5** is a detailed flowchart of the exemplary task classification process shown in FIG. **4**;

FIG. **6** is an exemplary excerpt from the set of acquired acoustic morphemes;

FIG. **7** is a graph illustrating the on-line learning experimental results as compared to the estimates from text transcriptions; and

FIG. **8** is a graph illustrating the on-line oracle experimental results as compared to the estimates from text transcriptions.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Traditional ASR techniques require the domain-specific knowledge of acoustic, lexicon data and more importantly the word probability distributions. However, this invention concerns a method for unsupervised learning of acoustic and lexical units from out-of-domain speech data. The new lexical units are used for fast adaptation of language model probabilities to a new domain. Starting from a lexicon and language model, the relevant language statistics of task classification systems may be learned.

The invention concerns spoken language systems that adapts on-line to new domains. State-of-the art research or deployed spoken dialog systems perform constrained tasks (e.g., travel information, stock quotes, etc.) and they achieve high task completion rates. These systems are programmed to answer users' queries as long as they belong to the apriori defined domain. However, conventional spoken dialog systems do not react to changes in the task (e.g., a new category for a call-routing task) and adapt to unseen speech and language events. The crucial features for such an adaptive system are the acoustic and lexical units.

The conventional large vocabulary speech recognition framework requires acoustic and language model to be trained on domain-specific data. This data is usually collected through human-to-machine interaction protocols and speech utterances are transcribed for the purpose of acoustic and language model training. These models perform poorly in out-of-domain conditions and are not suitable for on-line learning of language. This is true despite the fact that large vocabulary lexicons can have millions of words and reduce the out-of vocabulary rate to zero. However, the main reason for such poor performance is the mismatch of the language model probabilities.

An alternative approach to large vocabulary recognition is to model phone sequences. There has been active research on this topic and high phone recognition rates have been achieved for restricted domains and language. While it is widely accepted that phone recognition performance is inferior to word-based large vocabulary speech recognition tasks (where large databases are available to train language models), phone recognition allows a system to perform task-independent speech recognition. For these reasons, task-independent speech recognizers should combine the accuracy of a word-based system and the acoustic event granularity of a phone-based system.

Baseline approaches to the teaching of speech recognition systems are found in U.S. Pat. Nos. 5,675,707 and 5,860,063,

6,173,261, and 6,192,110, which have been incorporated above herein by reference in their entireties.

While the morphemes may be non-acoustic (i.e., made up of non-verbal sub-morphemes such as tablet strokes, gestures, body movements, etc.), for ease of discussion, the systems and methods illustrated in the drawings and discussed in the below concern only acoustic morphemes. Consequently, the invention should not be limited to just acoustic morphemes and should encompass the utilization of any sub-units of any known or future method of communication for the purposes of recognition and understanding.

Furthermore, while the terms "speech", "phrase" and "utterance", used throughout the description below, may connote only spoken language, it is important to note in the context of this invention, "speech", "phrase" and "utterance" may include verbal and/or non-verbal sub-units (or sub-morphemes). Therefore, "speech", "phrase" and "utterance" may comprise non-verbal sub-units, verbal sub-units or a combination of verbal and non-verbal sub-units within the sprit and scope of this invention.

Non-verbal speech may include but are not limited to gestures, body movements, head movements, non-responses, text, keyboard entries, keypad entries, mouse clicks, DTMF codes, pointers, stylus, cable set-top box entries, graphical user interface entries and touchscreen entries, or a combination thereof. Multimodal information is received using multiple channels (i.e., aural, visual, etc.). The user's input communication may also be derived from the verbal and non-verbal speech and the user's or the machine's environment. Basically, any manner of communication falls within the intended scope of the invention. However, for ease of discussion, we will focus on verbal speech in the examples and embodiments set forth below.

FIG. **1** is an exemplary block diagram of an exemplary speech recognition and understanding system **100** that utilizes acoustic morphemes. The speech recognition and understanding system **100** may include a phone recognizer **110**, an acoustic morpheme detector **120**, a phonotactic model learning unit **140**, and a phonotactic database **130**.

In the speech recognition and understanding system **100**, the phone recognizer **110** may receive an input communications, such as speech, from a user, customer, etc., in the form of verbal and/or non-verbal speech. As discussed above, the input speech may, of course, be expressed in verbal speech, non-verbal speech, multimodal forms, or using a mix of verbal and non-verbal speech.

The phone recognizer **110** is the task-independent component of the speech recognition and understanding system **100** which transcribes speech utterances from one domain and is trained on another domain. Prior to being integrated in the system **100**, the phone recognizer **110** may be initially trained from an off-the-shelf database, for example. The database may be generated from recordings of users talking with human agents, responding to the prompt "AT&T, How may I help you?" (HMIHY). The characteristics of this data and early experiments are detailed in U.S. Pat. No. 5,675,707, for example, which has been incorporated herein by reference in its entirety. In any event, the phone recognizer **110** may be of any known design and performs the function of recognizing, or spotting, the existence of one or more phone in the user's input speech.

In an embodiment for recognizing non-acoustic morphemes, the phone recognizer **110** may be replaced in the figure by a sub-morpheme recognizer. The sub-morpheme recognizer would operate similar to the phone recognizer **110**, but it would receive raw non-acoustic or a mixture of acoustic and non-acoustic speech from the user. Again, although the invention encompasses non-acoustic speech, for ease of discussion, only acoustic morphemes will be discussed in reference to the figures.

The acoustic morpheme detector **120** detects the acoustic morphemes present in the recognized phones from user's input request. The acoustic morphemes detected by acoustic morpheme detector **120** may be used for processing by any number of automated systems known to those of ordinary skill in the art. For example, the acoustic morphemes detected by the acoustic morpheme detector **120** may include a large number of verbal and non-verbal speech fragments or sub-morphemes (illustrated as phone-phrases for ease of discussion), each of which may be related to one of a predetermined set of task objectives. In this respect, each of the acoustic morphemes may be labeled with its associated task objective. FIG. **2** shows an example of one such task classification system **200**, as discussed below.

The acoustic morphemes detected by the acoustic morpheme detector **120** are also provided to the phonotactic model learning unit **140** for building a new phonotactic model. The phonotactic model learning unit **140** generates a new phonotactic model based on the detected acoustic morphemes found in the recognized phones from the user's input speech. The newly generated phonotactic model is then stored in the phonotactic model database **130** for use by the phone recognizer **110**.

FIG. **2** shows the use of detected acoustic morphemes resulting from the automatically generated phontotactic model in a task classification system **200**. The task classification system **200** may include a task classification processor **210** along with the same elements as the speech recognition and understanding system **100**. The task classification processor **210** may apply a confidence function, based on the probabilistic relation between the recognized acoustic morphemes and selected task objectives, and makes a decision either to implement a particular task objective, or makes a determination that no decision is likely in which case the user may be defaulted to a human or automated system for assistance. Within this framework, on-line estimates of word or morpheme probabilities are learned for task-independent speech recognition. In fact, a classifier that operates on words, morphemes (acoustic and non-acoustic), and/or sub-morphemes, etc., may be used within the spirit and scope of the invention.

An exemplary process of the invention will now be described in FIGS. **3-5** with reference to the exemplary structure shown in FIGS. **1** and **2**. FIG. **3** illustrates a possible process of automatically building a phonotactic model using acoustic morphemes for use in the speech recognition and understanding system **100**. The process begins at step **3010** and proceeds to step **3020** where the phone recognizer **110** receives raw speech from a user and recognizes phones in the speech using the current phonotactic model stored in the phonotactic model database **130**. In the phonotactic model, each word in the speech corpus is replaced with its most likely pronunciation and trained n-gram model based on the Variable N-gram Stochastic Automaton (VNSA).

The acoustic morpheme detector **120** may detect phone sequences by computing information theoretic measures, such as the weighted mutual information $I_w(x,y)=P(x,y)MI(x,y)=P(x,y)\log P(x,y)/P(x)P(y)$, where x and y are two phone sequences. Weighted mutual information compensates for probability estimates biases introduced by rare events in the mutual information, MI(x,y). Alternative and computationally expensive methods for computing acoustic morphemes are entropy-minimizing units and multi-gram approaches. It can be shown both theoretically and experimentally that the

n-best I_w is a close approximation of the exact entropy minimization problem. Such methods are shown in U.S. Pat. No. 6,021,384, which has been incorporated by reference above.

Excerpts from the detected acoustic morphemes are shown in FIG. **6**. The first and second column contain the I_w and MI values, the third and fourth are the corresponding acoustic morpheme sequence and its most likely word sequence. The phonotactic model learning unit **140** may for example, incorporate the top N (N=500) acoustic morphemes in the stochastic phonotactic models which takes advantage of the long-spanning acoustic morphemes.

After the acoustic morphemes have been detected, the process proceeds on a dual track of processing and training. In the processing track, in step **3040**, the acoustic morphemes are output for processing. One of the possible processes for detected acoustic morphemes may be to the map them into word sequences using a lexical access unit. In general, this process is not needed if the semantic units are not based on words. However, word-based language models are the natural candidates for cross-domain modeling, given the large availability of transcribed speech databases and the effectiveness word-based large vocabulary speech recognition.

Acoustic morphemes can be mapped to zero, one or many word sequences, for a given baseform lexicon. A conservative strategy to lexical access may be taken which considers only the exact match between baseforms in the lexicon and acoustic morphemes. This conservative strategy is dictated by the need to learn the most relevant lexical features of the domain with high precision and high rejection. These lexical features may be used to adapt on-line an off-the-shelf large vocabulary language model. The acquisition algorithm discussed above has selected 500 units that were matched against their most likely word mapping drawn from the lexicon. For example, in FIG. **6**, a list of words and phrases was compiled. While the list of lexical items is short, due to the conservative strategy, they cover 40% of the probability mass of the speech corpus.

Alternatively, the detected acoustic morphemes may be used by a task classification system **200** from FIG. **2**, as depicted in the exemplary flowchart shown in FIG. **4**. FIG. **4** includes steps **4010**, **4020**, **4030**, **4050**, **4060** and **4070** that correspond to steps **3010**, **3020**, **3030**, **3050**, **3060** and **3070** of FIG. **3**. However, FIG. **4** also includes step **4040** where the task classification processor **210** makes task classification decisions based on the detected acoustic morphemes.

As an example of processes that may be performed as part of a possible task classification process using acoustic morphemes, FIG. **5** is provided. In FIG. **5**, from step **4030** where acoustic morphemes are detected from recognized phones from a user's input communication, the process proceeds to step **4040** and sub-steps **4041-4049** in which the task classification processor **210** performs task classifications based on the detected acoustic morphemes. The task classification processor **210** may apply a confidence function based on the probabilistic relation between the recognized acoustic morphemes and selected task objectives, for example.

In step **4041**, the task classification processor **210** determines whether a task can be classified based on the detected acoustic morpheme. If the task can be classified, in step **4047**, the task classification processor **210** routes the user/customer according to the classified task objective. The process then goes to step **4070** and ends.

If the task cannot be classified in step **4041** (i.e. a low confidence level has been generated), in step **4043**, a dialog module (located internally or externally) the task classification processor **210** conducts dialog with the user/customer to obtain clarification of the task objective. After dialog has been conducted with the user/customer, in step **4045**, the task classification processor **210** determines whether the task can now be classified based on the additional dialog. If the task can be classified, the process proceeds to step **4047** and the user/customer is routed in accordance with the classified task objective and the process ends at step **4070**. However, if task can still not be classified, in step **4049**, the user/customer is routed to a human for assistance and then the process goes to step **4070** and ends.

Referring back to FIG. **3**, after the acoustic morphemes have been detected in step **3030**, they may be used in a processed for building a phonotactic model using the phonotactic model learning unit **140** for ultimate use by the phone recognizer **110** in recognizing phone sequences in the user's next utterance. In building the phonotactic model, the phonotactic learning unit **140** transforms a prior probability distribution from one domain to another without speech transcriptions (on-line adaptation). This is the primary constraint of an adaptive spoken dialog system, which can react and recover to novel events is the syntactic and semantic input channel. Most of the literature in language model adaptation is for batch processing (with speech transcriptions) and assumes that the probability space is the same both for the source and target domain, namely the word sequence space.

The building of the phonotactic model by the phohtactic model learning unit **140** according to the invention involves at least two stochastic modeling processes. The first is to map phone sequence statistics into word statistics. If the most likely pronunciation of the word collect is K ae l eh K t, then P(collect)≠P(K ae l eh K t). In general terms:

$$P_{on-line}(\omega) = \sum_{f_i \in B_\omega} P(f_i)$$

where f_i is a baseform of the word ω drawn from the set B_ω. It is assumed from here forward that $P_{on-line}(\omega) \approx P(f_I)$, where f, is the most likely pronunciation.

FIG. **7** is a scatter plot (log $P_{HMIHY}(\omega)$, log $P_{on-line}(\omega)$) (asterisk) and (log $P_{HMIHY}(\omega)$, log $P_{switchboard}(\omega)$ (circle), where both $P_{HMIHY}(\omega)$ and $P_{switchboard}(\omega)$ have been estimated from the hand-labeled speech transcriptions. Most word probability estimates lie along the diagonal (dashed line) and closely approximate the true empirical distribution $P_{HMIHY}(\omega)$. In order to get an upper bound on the on-line probability estimates the phone sequence with lowest string edit distance from the reference phone transcription was computed. Thus, we have compiled the oracle phone transcription training set and estimated $P_{oracle\ online}(\omega)$.

FIG. **8** is a scatter plot of the log-probabilities for the on-line oracle estimates. From FIG. **8** we see that the probabilities of the selected words and phrases have been estimated with high precision with respect to the best possible guess (oracle).

The second problem is to transform the word probability vector of the prior distribution using the word statistics learned on-line. Given the small lexical coverage and large probability mass of the acquired features the two distributions have been merged with following scheme:

$$P_{target}(\omega) = \begin{cases} P_{on-line}(\omega) & w \in S \\ 1 - \alpha P_{switchboard}(\omega) & w \in S \\ \beta & \end{cases}$$

where:

$$\alpha=\Sigma_{w\in S}P_{on\text{-}line}(\omega),\beta=\Sigma_{w\in S},\Sigma_{w\in V}P(\omega)$$

and S is the sec of selected words. This model has been tested to measure the perplexity of the HMIHY test set perplexity for a unigram model. The unigram perplexity on the HMIHY test set, in matched language model condition ($P_{HMIHY}(\omega)$ is 128.3. The test perplexity using the switchboard language model ($P_{switchboard}(\omega)$) is 715.9. The test set perplexity was tested with the new word probability $P_{target}(\omega)$ and achieved a 39.4% relative improvement (433.9) with respect to the switchboard baseline.

Once the new phonotactic model is built, in step **3060** it is stored in the phonotactic model database **130** for use by the phone recognizer **110** during the user's next utterance or interaction. The database **130** may be any mechanism, memory, file etc. capable of storing information and data known to those of skill in the art.

As shown in FIGS. **1** and **2**, the method of this invention may be implemented using a programmed processor. However, method can also be implemented on a general-purpose or a special purpose computer, a programmed microprocessor or microcontroller, peripheral integrated circuit elements, an application-specific integrated circuit (ASIC) or other integrated circuits, hardware/electronic logic circuits, such as a discrete element circuit, a programmable logic device, such as a PLD, PLA, FPGA, or PAL, or the like. In general, any device on which the finite state machine capable of implementing the flowcharts shown in FIGS. **3-5** can be used to implement the speech recognition and understanding system functions of this invention.

While the invention has been described with reference to the above embodiments, it is to be understood that these embodiments are purely exemplary in nature. Thus, the invention is not restricted to the particular forms shown in the foregoing embodiments. Various modifications and alterations can be made thereto without departing from the spirit and scope of the invention.

What is claimed is:

1. A method for building a phonotactic model for domain independent speech recognition, comprising:

recognizing phones from a user's input communication using a current phonotactic model stored in a database;

detecting morphemes from the recognized phones;

creating, via a new processor, a new phonotactic model using the detected morphemes, the creating the new phonotactic model comprising transforming a prior probability distribution associated with a first domain to a prior probability distribution associated with a second domain;

replacing the current phonotactic model with the new phonotactic model in the database; and

outputting the detected morphemes for processing.

2. The method of claim **1**, wherein the morphemes include at least one of verbal speech or non-verbal speech.

3. The method of claim **2**, wherein the non-verbal speech includes use of at least one of gestures, body movements, head movements, non-responses, text, keyboard entries, keypad entries, mouse clicks, DTMF codes, pointers, stylus, cable set-top box entries, graphical user interface entries or touchscreen entries.

4. The method of claim **2**, wherein the user's input communication is derived from the verbal and non-verbal speech and an environment of the user.

5. The method of claim **1**, wherein the morphemes are expressed in multimodal form.

6. The method of claim **1**, wherein the morphemes in the user's input communication are derived from actions of the user, including a focus of attention of the user.

7. The method of claim **1**, further comprising:

mapping the detected acoustic morphemes into words.

8. The method of claim **1**, wherein the method operates in conjunction with one or more communication networks, the communication networks including a telephone network, the Internet, an intranet, cable TV network, a local area network (LAN), and a wireless communication network.

9. The method of claim **1**, wherein the method is used for customer care purposes.

10. The method of claim **1**, wherein the morphemes are detected using a measure of commonality within a language of the morphemes.

11. The method of claim **10**, wherein the commonality measure is a mutual information measure.

12. The method of claim **11**, wherein each of the plurality of detected morphemes has a mutual information measure exceeding a predetermined threshold.

13. The method of claim **1**, wherein detecting acoustic morphemes includes using a confidence function.

14. The method of claim **1**, wherein the new phonotactic model is used when recognizing phones for a future input communication from a user.

15. The method of claim **1**, wherein the method is iterative.

16. The method of claim **1**, wherein the morphemes are at least one of acoustic morphemes or non-acoustic morphemes.

17. The method of claim **1**, wherein creating a new phonotactic model using the detected morphemes further comprises:

mapping phone sequence statistics into word statistics; and

transforming a word probability vector of the prior probability distribution of the first domain using the mapped word statistics.

18. A method of task classification using a phonotactic model built for domain independent speech recognition, comprising:

recognizing phones from a user's input communication using a current phonotactic model stored in a database;

detecting morphemes from the recognized phones;

creating, via a processor, a new phonotactic model using the detected morpheme, the creating the new phonotactic model comprising transforming a prior probability distribution associated with a first domain to a prior probability distribution associated with a second domain;

replacing the current phonotactic model with the new phonotactic model in the database; and

making a task-type classification decision based on the detected morphemes from the user's input communication.

19. The method of claim **18**, wherein the morphemes include at least one of verbal speech or non-verbal speech.

20. The method of claim **19**, wherein the non-verbal speech includes the use of at least one of gestures, body movements, head movements, non-responses, text, keyboard entries, keypad entries, mouse clicks, DTMF codes, pointers, stylus, cable set-top box entries, graphical user interface entries or touchscreen entries.

21. The method of claim **19**, wherein the user's input communication is derived from the verbal and non-verbal speech and an environment of the user.

22. The method of claim **19**, wherein each of the verbal and non-verbal speech are directed to one of a set of predeter-

mined task objectives and each of the verbal and non-verbal speech is labeled with one of the predetermined task objectives to which it is directed.

23. The method of claim **18**, wherein the morphemes are expressed in multimodal form.

24. The method of claim **18**, wherein the morphemes in the user's input communication are derived from actions of the user, including a focus of attention of the user.

25. The method of claim **18**, further comprising entering into a dialog with the user to obtain a feedback response from the user.

26. The method of claim **25**, wherein entering into a dialog with the user to obtain a feedback response from the user, further prompting the user to provide a feedback response including additional information with respect to an initial input communication of the user.

27. The method of claim **25**, wherein the user is prompted to provide a feedback response that includes confirmation with respect to at least one set of task objectives determined in the task-type classification decision.

28. The method of claim **18**, wherein the input communication is routed based on the task-type classification decision.

29. The method of claim **28**, wherein a task objective is performed after the input communication is routed.

30. The method of claim **18**, wherein the method operates in conjunction with one or more communication networks, the communication networks including a telephone network, the Internet, an intranet, cable TV network, a local area network (LAN), and a wireless communication network.

31. The method of claim **18**, wherein the method is used for customer care purposes.

32. The method of claim **18**, wherein the classification decisions decision and corresponding user input communications communication are collected for automated learning purposes.

33. The method of claim **18**, wherein a relationship between the detected morphemes and a predetermined set of task objectives includes a measure of usefulness of one of the morphemes to a specified one of the predetermined set of task objectives.

34. The method of claim **18**, wherein a relationship between the detected morphemes and a predetermined set of task objectives includes a measure of commonality within a language of the morphemes.

35. The method of claim **34**, wherein the commonality measure is a mutual information measure.

36. The method of claim **35**, wherein each of the detected morphemes has a mutual information measure exceeding a predetermined threshold.

37. The method of claim **18**, wherein making a task-type classification decision includes using a confidence function.

38. The method of claim **18**, wherein the input communication from the user represents a request for at least one of a set of predetermined task objectives.

39. The method of claim **18**, wherein the input communication is responsive to a query of a of form "How may I help you?".

40. The method of claim **18**, wherein the new phonotactic model is used when recognizing phones for a future input communication from a user.

41. The method of claim **18**, wherein the method is iterative.

42. The method of claim **18**, wherein the morphemes are at least one of acoustic morphemes or non-acoustic morphemes.

43. The method of claim **18**, wherein creating a new phonotactic model using the detected morphemes further comprises:

mapping phone sequence statistics into word statistics; and

transforming a word probability vector of the prior probability distribution of the first domain using the mapped word statistics.

44. A system that builds a phonotactic model for domain independent speech recognition, comprising:

a database that stores phonotactic models;

a recognizer that recognizes phones from a user's input communication using a current phonotactic model stored in the database;

a morpheme detector that detects morphemes from the recognized phones and outputs the detected morphemes for processing; and

a learning unit that creates a new phonotactic model using the detected morphemes and transforms a prior probability distribution associated with a first domain to a prior probability distribution associated with a second domain, wherein the learning unit replaces the current phonotactic model with the new phonotactic model in the database.

45. The system of claim **44**, further comprising:

a task classification processor that makes task classification decisions based on the detected morphemes from the user's input communication.

46. The system of claim **44**, wherein the learning unit is configured to perform:

mapping phone sequence statistics into word statistics; and

transforming a word probability vector of the prior probability distribution of the first domain using the mapped word statistics.

* * * * *

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 8,392,188 B1
APPLICATION NO. : 09/956907
DATED : March 5, 2013
INVENTOR(S) : Giuseppe Riccardi

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In the Claims

Col. 7, Claim 1, line 45, change "the recognized phones" to --the phones--

Col. 7, Claim 1, line 47, change "the detected morphemes" to --the morphemes--

Col. 7, Claim 1, line 54, change "the detected morphemes" to --the morphemes--

Col. 7, Claim 4, line 64, change "the verbal and non-verbal" to --the at least one of verbal or non-verbal--

Col. 8, Claim 7, line 5, change "the detected acoustic morphemes" to --the morphemes--

Col. 8, Claim 12, lines 19-20, change "the plurality of detected morphemes" to --the morphemes--

Col. 8, Claim 13, line 22, change "detecting acoustic" to --detecting--

Col. 8, Claim 17, line 31, change "creating a new" to --creating the new--

Col. 8, Claim 17, line 32, change "using the detected morphemes" to --using the morphemes--

Col. 8, Claim 17, line 36, change "the mapped" to --the--

Col. 8, Claim 18, line 45, change "the detected morphemes" to --the morphemes--

Col. 8, Claim 18, line 53, change "detected morphemes" to --morphemes--

Signed and Sealed this
Seventeenth Day of June, 2014

Michelle K. Lee

Michelle K. Lee
Deputy Director of the United States Patent and Trademark Office

Col. 8, Claim 21, line 64, change “from the verbal and non-verbal” to --from the at least one of the verbal or non-verbal--

Col. 8, Claim 22, line 66, change “of the verbal and non-verbal” to --of the at least one of the verbal or non-verbal--

Col. 9, Claim 22, line 1, change “of the verbal and non-verbal” to --of the at least one of the verbal or non-verbal--

Col. 9, Claim 26, lines 14-15, change “from the user, further prompting the user” to --from the user prompts the user--

Col. 9, Claim 32, lines 34-35, change “decisions decision and corresponding user input communications” to --decision and corresponding user input--

Col. 9, Claim 33, line 38, change “the detected morphemes” to --the morphemes--

Col. 9, Claim 34, line 43, change “the detected morphemes” to --the morphemes--

Col. 9, Claim 36, line 48, change “each of the detected” to --each of the--

Col. 10, Claim 39, line 7, change “of a of form” to --of a form--

Col. 10, Claim 40, line 11, change “from a user” to --from the user--

Col. 10, Claim 43, line 17, change “the detected morphemes” to --the morphemes--

Col. 10, Claim 43, line 21, change “the mapped” to --the--

Col. 10, Claim 44, line 30, change “recognized phones and outputs the detected morphemes” to --phones and outputs the morphemes--

Col. 10, Claim 44, line 33, change “the detected morphemes” to --the morphemes--

Col. 10, Claim 45, line 41, change “the detected morphemes” to --the morphemes--

Col. 10, Claim 46, line 47, change “the mapped” to --the--