US 20090319270A1

(19) United States

(12) Patent Application Publication

GROSS

(10) Pub. No.: US 2009/0319270 A1

(43) Pub. Date: Dec. 24, 2009

(54) CAPTCHA USING CHALLENGES OPTIMIZED FOR DISTINGUISHING BETWEEN HUMANS AND MACHINES

(76) Inventor: JOHN NICHOLAS GROSS, BERKELEY, CA (US)

Correspondence Address:
J. NICHOLAS GROSS, ATTORNEY
2030 ADDISON ST., SUITE 610
BERKELEY, CA 94704 (US)

(21) Appl. No.: 12/484,800

(22) Filed: Jun. 15, 2009

Related U.S. Application Data

(60) Provisional application No. 61/074,979, filed on Jun. 23, 2008.

Publication Classification

(51) Int. Cl.
G10L 15/02 (2006.01)

(52) U.S. Cl. 704/246; 704/E15.001

(57) ABSTRACT

An audible based electronic challenge system is used to control access to a computing resource by using a test to identify an origin of a voice. The test is based on analyzing a spoken utterance using optimized challenge items selected for their discrimination capability to determine if it was articulated by an unauthorized human or a text to speech (TTS) system.

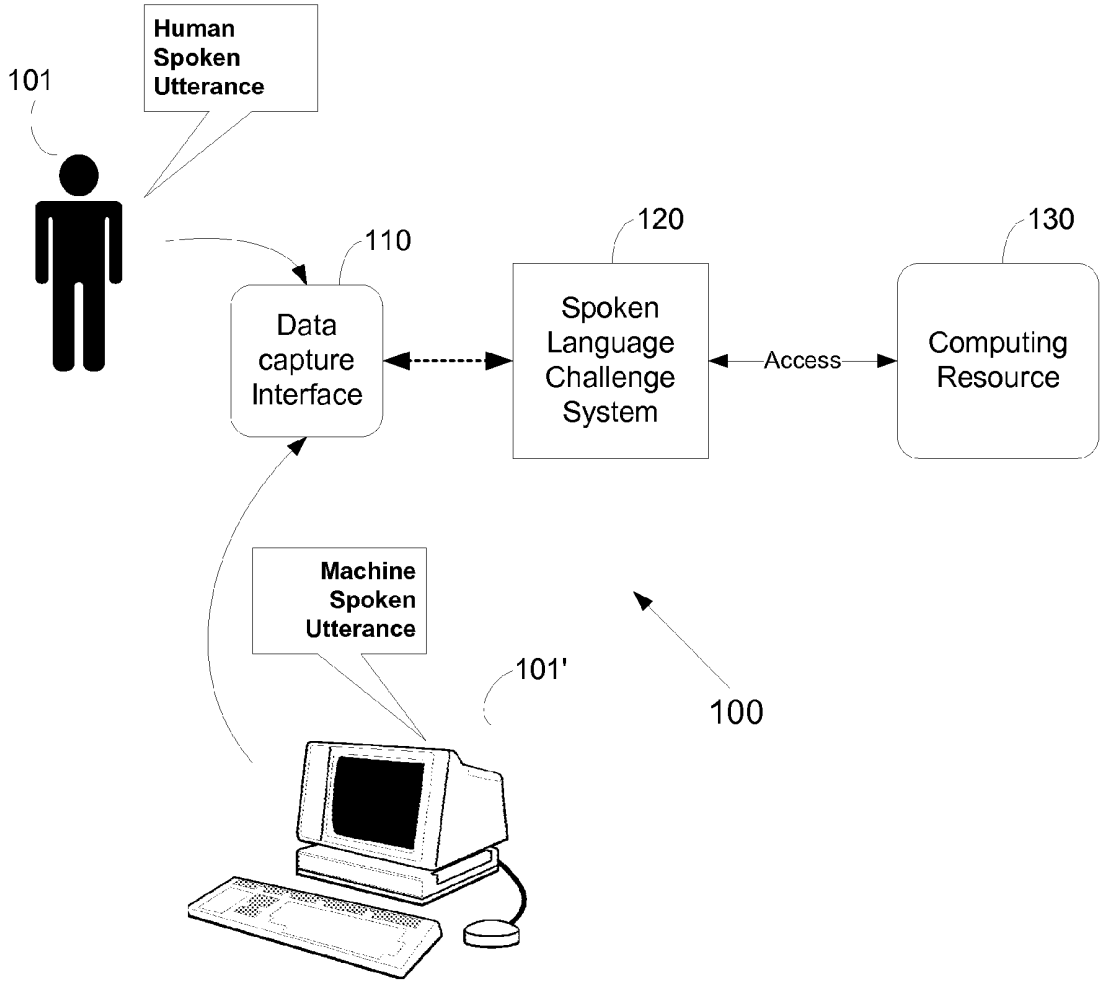


FIGURE 1

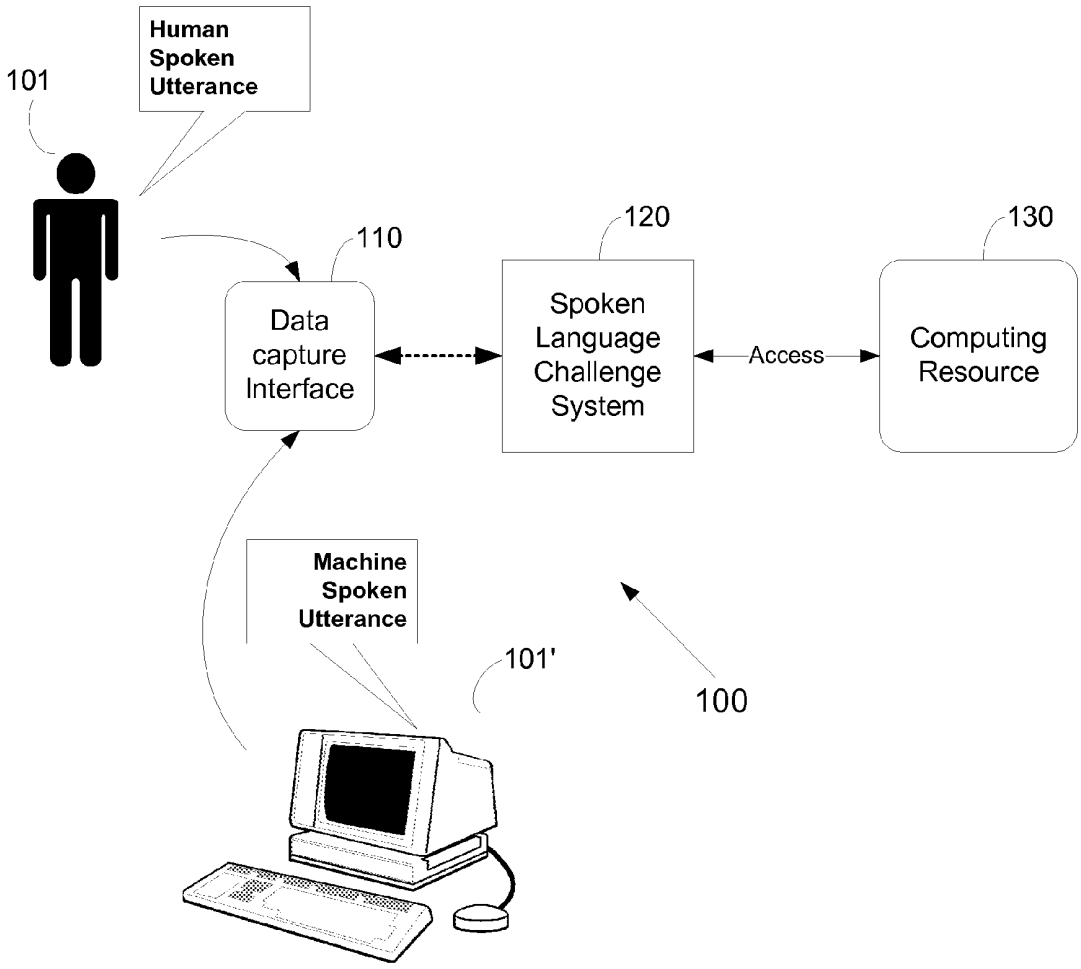

Human
Spoken
Utterance
101
110
120
130
Data
capture
Interface
Spoken
Language
Challenge
System
Access
Computing
Resource
Machine
Spoken
Utterance
101'
100

FIG. 2

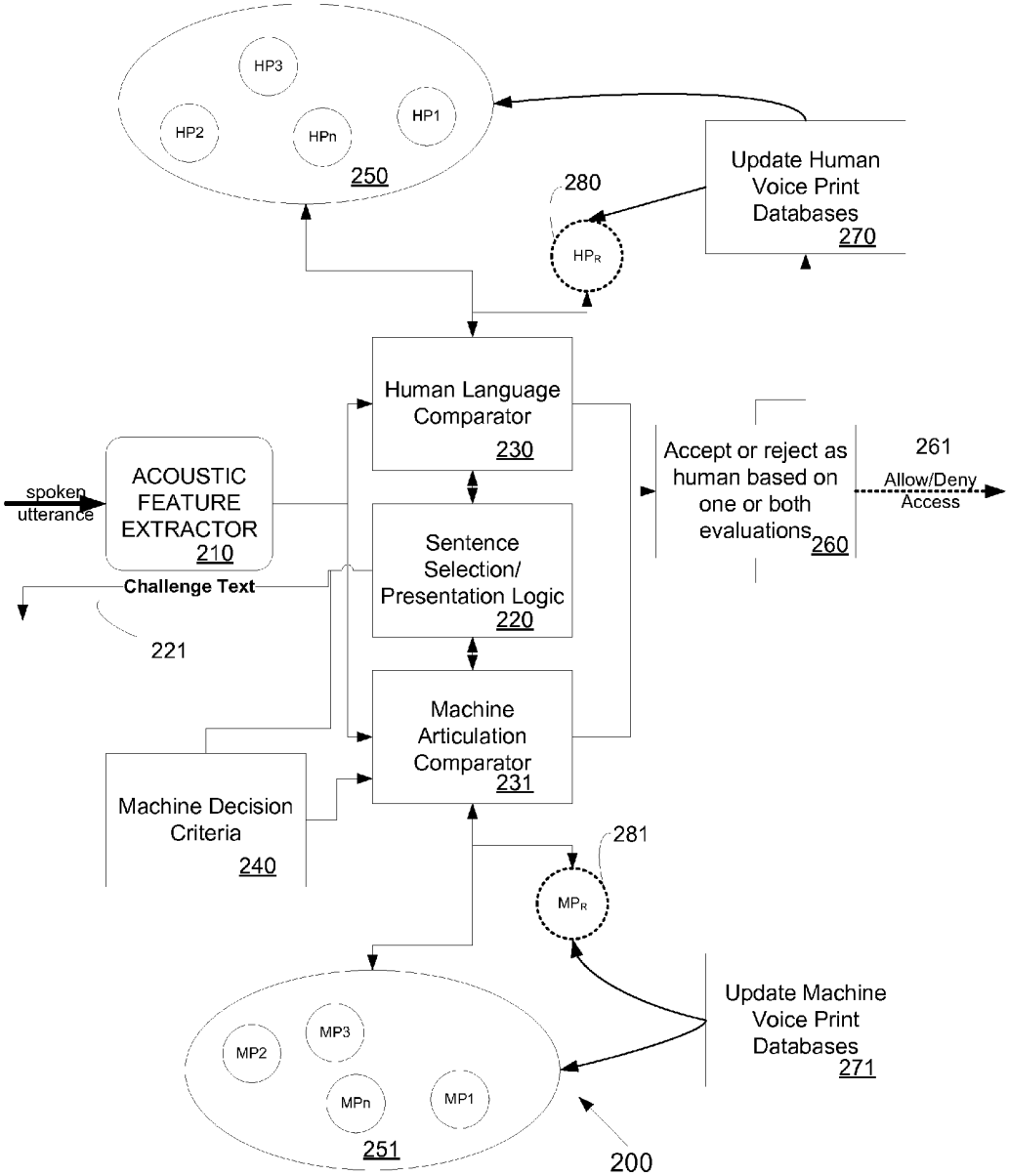

HP3
HP2
HPn
HP1
250
Update Human Voice Print Databases
270
280
HPR
Human Language Comparator
230
spoken utterance
ACOUSTIC FEATURE EXTRACTOR
210
Challenge Text
221
Sentence Selection/ Presentation Logic
220
Machine Articulation Comparator
231
Accept or reject as human based on one or both evaluations
260
261
Allow/Deny Access
Machine Decision Criteria
240
281
MPR
Update Machine Voice Print Databases
271
MP2
MP3
MPn
MP1
251
200

FIGURE 3

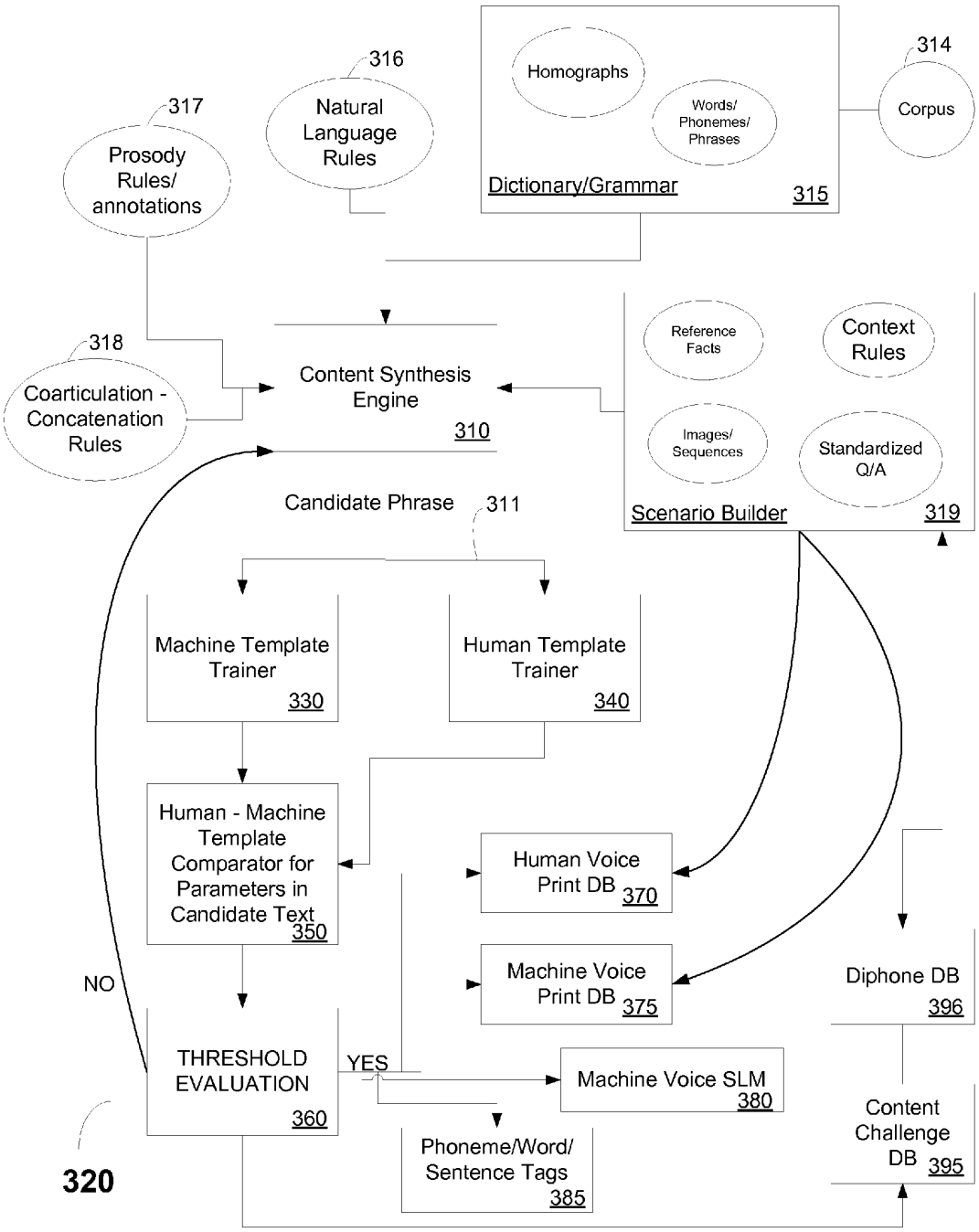

316
Natural Language Rules
317
Prosody Rules/ annotations
Homographs
Words/ Phonemes/ Phrases
Dictionary/Grammar
315
314
Corpus
318
Coarticulation - Concatenation Rules
Content Synthesis Engine
310
Reference Facts
Context Rules
Images/ Sequences
Standardized Q/A
Scenario Builder
319
Candidate Phrase
311
Machine Template Trainer
330
Human Template Trainer
340
Human - Machine Template Comparator for Parameters in Candidate Text
350
Human Voice Print DB
370
Machine Voice Print DB
375
Diphone DB
396
NO
THRESHOLD EVALUATION
360
YES
Machine Voice SLM
380
Phoneme/Word/ Sentence Tags
385
Content Challenge DB
395
320

FIG. 4A

Please speak the text in the box below into the microphone using your normal voice

The quick brown fox jumps over a lazy dog

FIG. 4B

Please look at the picture below and using your normal voice speak into the microphone and tell me what the person is doing:

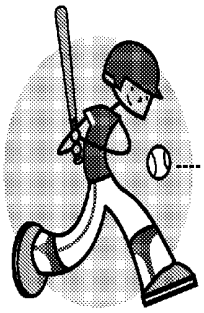
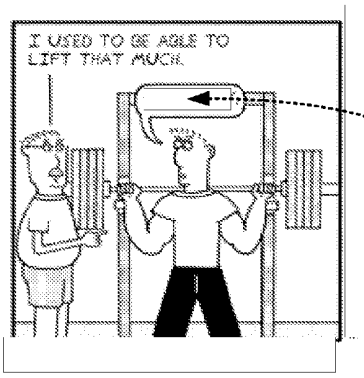


FIG. 4C

Please look at the picture below and using your normal voice speak only one of the responses below that you would give for the other person into the microphone:



Yeah right old man
You don't look like a spinach eater now
I'm just warming up ... I can really do twice this much
My cellphone is ringing

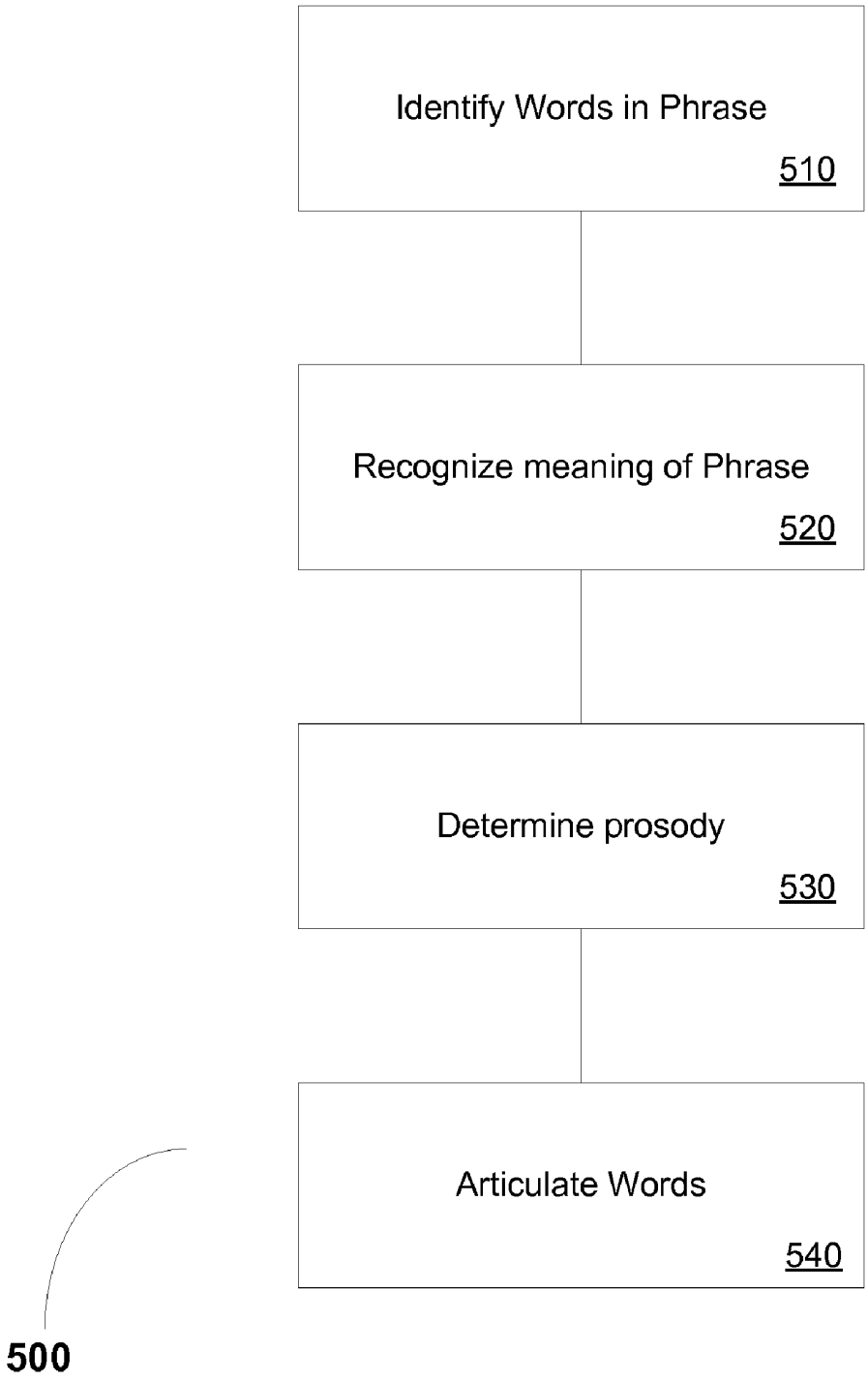
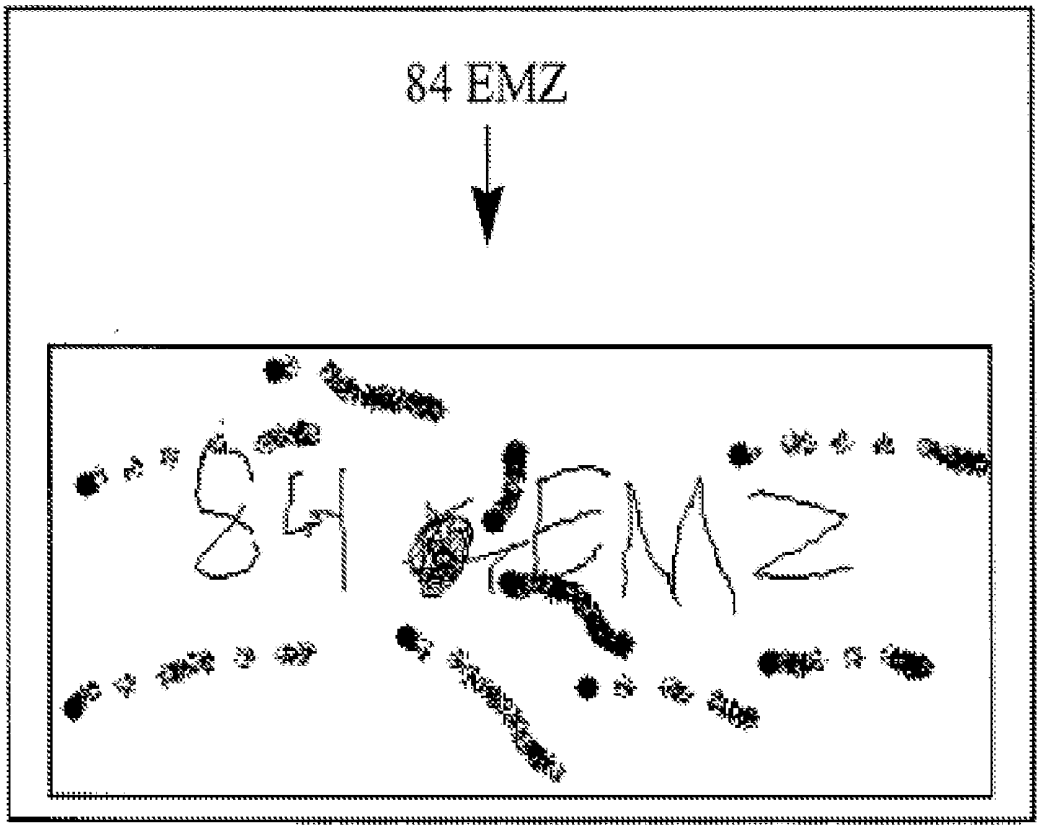


FIGURE 5: PRIOR ART

FIGURE 6: PRIOR ART



CAPTCHA USING CHALLENGES OPTIMIZED FOR DISTINGUISHING BETWEEN HUMANS AND MACHINES

RELATED APPLICATION DATA

[0001] The present application claims the benefit under 35 U.S.C. 119(e) of the priority date of Provisional Application Ser. No. 61/074,979 filed Jun. 23, 2008 which is hereby incorporated by reference. The application is further related to the following applications, all of which are filed on this same date and incorporated by reference herein:

[0002] System & Method for Verifying Origin of Input Through Spoken Language Analysis Ser. No. _____ (attorney docket number 2009-2)

[0003] System & Method for Generating Challenge Items for CAPTHCAs; Ser. No. _____ (attorney docket number 2009-3)

FIELD OF THE INVENTION

[0004] The present invention relates to electronic systems for detecting and differentiating input provided by humans and machines. These systems are used primarily in Internet applications for verifying that data originating from a source is from a human, and not from an unauthorized computer program/software agent/robot. In other applications entities can be fingerprinted to detect unauthorized accesses regardless of their origin.

BACKGROUND

[0005] CATPCHA (Completely Automatic Public Turing Tests To Tell Humans And Computers Apart) systems are well known in the art. Examples of such are used by Yahoo! (Gimpy type), Xerox PARC (Baffle type); so-called Bongo, Pix and Pessimal types are also known in the art. One of the first such visual based systems is described in U.S. patent application Ser. No. 10/790,611 to Reshef, which is hereby incorporated by reference herein.

[0006] Generally speaking, the goal of visual based CAPTCHAs is to present an optical image which is only decipherable/comprehensible by a human. To this end, the bulk of these systems rely primarily on some combination of pseudorandom letters and numbers which are placed in front of an obfuscating background, or subjected to visual degradation to make them machine-unrecognizable. A good background on such technologies can be found in the article "Is it Human or Computer? Defending E-Commerce with Captchas," by Clark Pope and Khushpreet Kaur in IT PRO, March-April 2005, p. 43-49, which is hereby incorporated by reference herein. An example of a typical CAPTCHA of the prior art is shown in FIG. **6**. The person looking at the image presented would have to determine that the text shown corresponds to the characters "84EMZ."

[0007] An article entitled What's Up CAPTCHA?—A CAPTCHA Based On Image Orientation by Gossweiler et al. incorporated by reference herein makes use of social feedback mechanisms to select appropriate challenge materials for visual CAPTCHs. The integration of aggregated human feedback allows for better selection of CAPTCHAs that are best optimized for discriminating against machines.

[0008] Recently, however, several sophisticated machine vision systems have achieved significant success in "breaking" the conventional optical CAPTCHA systems. For an example of such system, see "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA" by Mori and Malik, also incorporated by reference herein and which is available at the University of California Berkeley Computer Science Department website. Thus, traditional forms of CAPTCHA appear to be at risk of becoming obsolete before they gain widespread adoption.

[0009] Audio CAPTCHas are also known in the art. For an example of such system please see the above article to Pope and Kaur, page 45. Generally speaking, these types of systems take a random sequence of recordings of words, numbers, etc., combine them, and then ask the user to input—via keyboard or mouse—whatever is "heard" by the user into the system to determine if the message is comprehended. A drawback of this approach, of course, is that speech recognizers are improving rapidly; an article by Reynolds and Heck entitled "Automatic Speaker Recognition Recent Progress, Current Applications, and Future Trends" presented at AAAS 2000 Meeting Humans, Computers and Speech Symposium 19 Feb. 2000—incorporated by reference herein—makes it clear that machines are in fact better analyzers and recognizers of speech than are humans at this point. Consequently, audio CAPTCHAs of this type are similarly doomed to failure at this point.

[0010] The Reynolds et al article also notes that speech verification systems are well-known in the art. These systems are basically used as a form of human biometric analyzer, so that a person can access sensitive information over a communications link using his/her voice. A voice print for the particular user is created using a conventional Hidden Markov Model (HMM) during an enrollment/training session. Later when the user attempts to access the system—for example, in a banking application the user may wish to transfer funds from an account—the system compares certain captured audio data from the user against the prior recording to see if there is a sufficiently close biometric match. Identities are typically confirmed by measuring such intrinsic personal traits as lung capacity, nasal passages and larynx size. Again, since speech recognizers are extremely accurate in evaluating speech data, a very reliable verification can be made to determine if the identity of the person matches the prior recorded voice print. Speaker verification systems are well-known and are disclosed, for example in such references as U.S. Pat. No. 5,897,616; 6,681,205 and Publication No. 20030125944 which are incorporated by reference herein.

[0011] Another article by Shucker—Spoofing and Anti-Spoofing Measures, Information Security Technical Report, Vol. 7, No. 4, pages 56-62, 2002 explains that these verification systems are very hard to fool with tape recording equipment and the like, because such systems cannot duplicate the physical characteristics noted above. Thus, some speaker-verification technology has ways of testing for "liveness." They specifically analyze for acoustic patterns suggesting that the voice has been recorded using a process called anti-spoofing. Another application of this technique for fingerprinting is also described generally in U.S. Pat. No. 6,851,051 to Bolle et al. which is incorporated by reference herein. Other biometric techniques for uniquely differentiating humans apart are disclosed in US Publication No. 20050185847A1 to Rowe which is also incorporated by reference herein.

[0012] To date, therefore, while verification systems have been used for distinguishing between humans, they have been designed or employed on a limited basis for the purpose of distinguishing between a computer speaking and a human

speaking as part of a CAPTCHA type tester/analyzer. This is despite the fact that a recent article entitled "The Artificial of Conversation" published at: http://htmltimes(dot)com/turing-test-machine-intelligence(dot)php implies that conventional Turing tests do not even bother examining computer system vocalizations since they are too difficult.

[0013] A recent article entitled "Accessible Voice CAPTCHAs for Internet Telephony" by Markkola et al. incorporated by reference herein describes a Skype challenge system that requires the user to speak a number of random digits. This illustrates that there is known value in using spoken CAPTCHAs.

[0014] Some recent filings by Rajakumar (US Publication No. 20070280436, 20070282605 and 20060248019) also incorporated by reference herein also discuss the use of a voice database for registering the names of known fraudsters. Thereafter when a person attempts access the system can detect whether the person calling is already registered and is therefore blocked based on his/her voiceprint.

[0015] A further filing by Maislos et al. (US Publication No. 20090055193) (Ser. No. 12/034,736) is also incorporated by reference herein. The Maislos system—while purportedly using voice to differentiate between humans and computing systems, and even different demographic groups—is only recently filed and does not contain many details on how to optimize such discrimination, or how to formulate appropriate challenges. Another company identified as Persay is also believed to be researching voice based CAPTHCA systems; see e.g. www(dot)persay(dot)com and accompanying literature for their SPID system.

SUMMARY OF THE INVENTION

[0016] An object of the present invention, therefore, is to overcome the aforementioned limitations of the prior art. It will be understood from the Detailed Description that the inventions can be implemented in a multitude of different embodiments. Furthermore, it will be readily appreciated by skilled artisans that such different embodiments will likely include only one or more of the aforementioned objects of the present inventions. Thus, the absence of one or more of such characteristics in any particular embodiment should not be construed as limiting the scope of the present inventions.

[0017] A first aspect of the invention concerns a method of identifying a source of data input to a computing system comprising: receiving speech utterance from an entity related to randomly selected challenge text; wherein the challenge text represents a selected set of one more contiguous words which when articulated have a measurable difference in acoustical characteristics between a reference human voice and a reference computer synthesized voice that exceeds a target threshold; processing the speech utterance with the computing system to compute first acoustical characteristics of the entity; and generating a determination of whether the speech utterance originated from a machine or a human.

[0018] In preferred embodiments additional steps may be performed including: identifying a first computer synthesized voice that best correlates to the entity; and selecting the randomly selected challenge text based on an identity of the first entity so as to maximize a difference in acoustical characteristics. In addition in preferred embodiments the challenge text can be selected in part based on a confirmation from a human listener that an articulation of such challenge text originated from a computer synthesized voice. Also preferred embodiments may have the steps: soliciting utterances from a plurality of separate computing machines to determine their respective acoustical characteristics; and storing the plurality of associated acoustical characteristics in a database of known computing entities. Multiple samples of individual challenge sentences are preferably collected. The challenge text is preferably selected in part based on a difference in time for rendering such text into audible form by a human and a computing machine.

[0019] Some preferred embodiments include a step: granting or denying access to data and/or a data processing device based on the results of the CAPTCHA, including a signup for an email account or a blog posting. For others the step of granting or denying access to an advertisement based on the determination is performed. Other preferred embodiments perform a separate automated visual challenge test so that both visual processing and articulation processing is considered in one or more of the determinations.

[0020] For some applications a prosody score associated with the speech utterance is also preferably considered during step (c). The first test text preferably consists of a sentence presented in visual form for the entity to articulate, and/or a sentence presented in audible form for the entity to repeat. The first text data can also consist of a question presented in visual form and further includes an image cue associated with the question as well as a separate set of acceptable responses presented in visual form.

[0021] In some preferred embodiments an additional step is performed: selecting one or more second computing systems for separately performing the process based on a performance and/or cost requirement. These one or more second computing systems can be selected based on a language spoken by the entity, an IP address or geographic region associated with the entity, an auction process in which such one more second computing systems bid for the right to process the speech utterance, etc. The challenge text can also be selected based on a detected accent and/or geographic region associated with the entity.

[0022] A set of sentences for inclusion as challenge text preferably is based on an articulation difficulty score for a computer synthesis engine. In other cases the sentences are automatically generated based on a measured concatenation difficulty for a set of diphones. In still other applications the set of sentences are extracted automatically from a corpus that includes web logs, newspapers, books and/or the Internet.

[0023] Another aspect of the invention concerns a method of implementing a CATPCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) to identify a source of data input to a computing system comprising: presenting an image CATPCHA to an entity, which image CAPTCHA includes one or more visually distorted words, phrases or images as a challenge item; receiving a speech utterance from an entity related to the challenge text; processing the speech utterance to generate a determination of whether the speech utterance originated from a machine or a human.

[0024] The image CAPTCHA can take various forms, and preferably includes at least two distinct words. The determination preferably includes a first score based on computing acoustical characteristics of the speech utterance, and a second score based on recognizing the speech utterance to determine if the one or more visually distorted words, phrases or images are correctly identified. The scoring can also take into account a time required for the entity to determine the image CAPTCHA. In yet other applications the image CAPTCHA

is revealed in distinct stages which span a predetermined time period, and with each stage presenting additional visual information.

[0025] Still another aspect concerns a method of identifying a source of data input to a computing system comprising: associating a first challenge item with a first set of individual text descriptors; wherein the first set of text descriptors are based on feedback provided by a group of human reviewers; associating a second challenge item with a second set of individual text descriptors; wherein the second set of individual text descriptors are also based on feedback provided by a group of human reviewers; identifying at least a first reference correlation between at least a first reference text descriptor for the first challenge item and a second reference text descriptor for the second challenge item, including a probability that a human reviewer identifying the first reference text descriptor when presented with the first challenge item also provides the second reference text descriptor when presented with the second challenge item, or vice-versa; presenting the first challenge item to an entity as part of an automated access challenge adapted to distinguish humans from computing machines; receiving speech utterance from the entity related to the first challenge item to determine a first input text descriptor; presenting the second challenge item to the entity as part of the automated access challenge adapted to distinguish humans from computing machines; receiving speech utterance from the entity related to the second challenge item to determine a second input text descriptor; comparing the first and second input text descriptors to identify the reference correlation between them as measured; and generating a determination of whether the speech utterance originated from a machine or a human based on a value of the reference correlation.

[0026] The challenge items preferably include an image, a question, or different types, including a first type which includes an image, and a second type which includes a question. The first and second challenge items can also be selected based on a value of the reference correlation.

[0027] A further aspect concerns a method of identifying a source of data input to a computing system comprising: receiving speech utterance from an entity related to randomly selected challenge text; processing the speech utterance with the computing system to compute first acoustical characteristics of the entity; comparing the first acoustical characteristics with at least one reference set of acoustical characteristics for a human voice to identify a first score for the speech utterance; comparing the first acoustical characteristics with at least one reference set of acoustical characteristics for a computer synthesized voice in parallel to generate a second score for the speech utterance; generating a determination of whether the speech utterance originated from a machine or a human based on the first score and the second score.

[0028] Some of the steps are preferably done by separate entities using separate computing machines, which can participate in an auction to identify the first and/or second scores.

[0029] Another aspect concerns a method of identifying a source of data input to a computing system comprising: selecting first test text data to be articulated as a speech utterance by an entity providing input to the computing system; receiving the speech utterance from the entity; generating first recognized speech data from the speech utterance corresponding to the first test text data; processing the first recognized speech data with the computing system to generate an initial determination of whether the speech utterance originated from a machine or a human; optionally repeating steps above based on a confidence score for the initial determination using second test text data, which second test text data is derived dynamically from content presented in the first recognized speech data; and processing the second recognized speech data with the computing system to generate a final determination of whether the speech utterance originated from a machine or a human.

[0030] Another aspect is directed to a method of controlling access to a computing system comprising: selecting first test text data to be articulated as a first speech utterance by a first entity providing input to the computing system; storing a voice print for the first entity at the computing system based on the first speech utterance being converted into recognized speech data; wherein the first entity can include either a human or a computer using a synthesized voice; receiving a second speech utterance by a second entity; processing the second recognized speech data with the computing system to determine whether the second speech utterance also originated from the first entity; controlling whether the second entity is allowed to access an account and/or data based on comparing the voice print to the second recognized speech data.

[0031] The access is preferably used for one or more of the following:

[0032] a) establishing an online account; and/or

[0033] b) accessing an online account; and/or

[0034] c) establishing a universal online ID; and/or

[0035] d) accessing a universal online ID; and/or

[0036] e) sending email; and/or

[0037] f) accessing email; and/or

[0038] g) posting on a message board; and/or

[0039] h) posting on a web log; and/or

[0040] i) posting on a social network site page;

[0041] j) buying or selling on an auction site; and/or

[0042] k) posting a recommendation for an item/service; and/or

[0043] l) selecting an electronic ad.

[0044] Yet another aspect concerns a method of identifying a source of data input to a computing system using prosodic elements of speech comprising: presenting a challenge item to an entity, which challenge item is associated with a reference set of words and associated reference prosodic scores; receiving speech utterance from an entity related to the challenge item including an input set of words; processing the speech utterance with the computing system to compute input prosodic scores of the input set of words; comparing the input prosodic scores and the reference prosodic scores; generating a determination of whether the speech utterance originated from a machine or a human based on the comparing.

[0045] Some preferred embodiments include the step: recognizing the input set of words to compute an additional prosodic score based on an identity of the input set of words, and comparing the additional prosodic words to a second reference prosodic score related to a content of the reference set of words. The challenge item is preferably supplemented with visual cues, the visual cues being adapted to induce the reference prosodic scores. The visual cues are preferably selected from a database of visual cues determined by reference to a database of human vocalizations to most likely result in the reference prosodic scores.

[0046] Still another aspect involves a method of identifying a source of data input to a computing system using prosodic elements of speech comprising: presenting a challenge item

to an entity, which challenge item is associated with a reference set of words and associated prosodic characteristics; receiving speech utterance from an entity related to the challenge item; wherein the reference set of words represents a selected set of one more contiguous words which when vocalized have a measurable difference in prosodic characteristics between a reference human voice and a reference computer synthesized voice that exceeds a target threshold; processing the speech utterance with the computing system to compute first prosodic characteristics of the entity; generating a determination of whether the speech utterance originated from a machine or a human based on the processing.

[0047] Some preferred embodiments include the steps: estimating a first computer synthesized voice that best correlates to the entity; and selecting the challenge item based on an identity of the first entity so as to maximize a difference in prosodic characteristics. Other embodiments include further steps: soliciting utterances from a plurality of separate computing machines to determine their respective prosodic characteristics; and storing the plurality of associated prosodic characteristics in a database of known computing entities. Multiple samples of individual challenge sentences are preferably collected. The visual cues are preferably added to induce the entity to vocalize the reference set of words using the reference human voice.

[0048] Other aspects concern a method of implementing a CATPCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) to identify a source of data input to a computing system comprising: training the computing system with samples of human voices and computer synthesized voices articulating a set of reference challenge items; receiving a speech utterance from an entity related to one of the set of reference challenge items; determining with the trained computer system whether the speech utterance was vocalized by a machine or a human.

[0049] A set of human test subjects preferably are used to identify whether a reference challenge item was vocalized by a human or a computer prior to using it in the training of the computing system. The reference challenge items are preferably ranked and sorted according to a score provided by the human test subjects, and further including a step: presenting the one of the set of reference challenge items based on the score.

[0050] A further aspect concerns a method of implementing a CATPCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) to identify a source of data input to a computing system comprising: training the computing system with samples of human voices articulating a set of reference challenge items; receiving a speech utterance from an entity related to one of the set of reference challenge items; determining with the trained computer system whether the speech utterance was vocalized by a machine or a human; wherein the computing system uses one or more speech models that are optimized for identifying humans using the set of reference challenge items.

[0051] A set of human test subjects preferably are used to identify whether a reference challenge item was vocalized by a human or a computer prior to using it in the training of the computing system. The set of reference challenge items preferably represent a selected set of one more contiguous words which when articulated have a difference in acoustical characteristics between a reference human voice and a reference computer synthesized voice that exceeds a target threshold as measured by a reference group of human listeners, and at least some of the acoustical characteristics are used to train the one or more speech models.

[0052] Still another aspect is directed to a method embodied in a computer readable medium for generating challenge data to be used for accessing data and/or resources of an electronic computing system comprising: automatically generating a candidate challenge sentence from a first set of words and phrases using the computing system; automatically generating at least one first utterance from a first machine text to speech system for the candidate challenge sentence using the computing system, the at least one first utterance including first acoustical characteristics; automatically generating at least one second utterance from a human speaker for the candidate challenge sentence using the computing system, the at least one second utterance including second acoustical characteristics; automatically determining a difference in the first and second acoustical characteristics using the computing system to determine a challenge sentence acoustic score for the candidate challenge sentence; automatically storing the at least one first utterance and the at least one second utterance and the candidate challenge sentence in a challenge item database using the computing system for use by an utterance based challenge system when the candidate challenge sentence acoustic score exceeds a target threshold.

[0053] Acoustical features of the first utterance and the second utterance are preferably measured, and regions of greatest difference are identified to be used as discriminators. In other embodiments, differences in articulation are preferably measured and scored on a diphone basis. The diphones can be sorted according to their difference between human and machine articulation. Based on such difference words and phrases are preferably selected from a corpus of text in accordance with a diphone score.

[0054] In some preferred embodiments a machine articulation statistical language model is compiled based on a plurality of first utterances.

[0055] For other preferred embodiments a dialog of multiple challenge sentences based on questions and expected answers to be provided by an entity is compiled. A challenge scenario preferably comprised of the candidate challenge sentence and one or more visual and/or text cues is generated, which challenge scenario is stored in the database.

[0056] A natural language engine can also preferably process the words and phrases to generate the candidate challenge sentence. An evaluation/scoring of the syntax difficulty preferably can be made for challenge sentences to determine appropriate candidates that would be more difficult for a challenge natural engine to decode.

[0057] In other preferred embodiments the candidate challenge sentence are annotated with prosodic elements to generate an expected prosodic pronunciation of the words and phrases. The candidate challenge sentence preferably are annotated with first prosodic elements found in the at least one first utterance and with second prosodic elements found in the least one second utterance. A difference in the first and second prosodic elements preferably is determined to generate a challenge sentence prosodic score for the candidate challenge sentence.

[0058] In other preferred embodiments The candidate challenge sentence preferably is annotated with first content elements for the visual and/or text cues found in the at least one first utterance and with second content elements for the visual

and/or text cues found in the least one second utterance. A difference in the first and second content elements preferably is determined to generate a challenge sentence content score for the candidate challenge sentence.

[0059] In still other preferred embodiments a time required by the human speaker and a machine speaker to generate the first utterance is measured.

[0060] The challenge database can then be used during a processing of input speech by an entity to distinguish between a human and a machine synthesized voice.

[0061] Another aspect concerns a method embodied in a computer readable medium of selecting challenge data to be used for accessing data and/or resources of a computing system comprising: providing a first set of diphones using the computing system; generating an articulation score using the computing system based on a machine text to speech (TTS) system articulation of each of the first set of diphones; and selecting challenge text using the computing system to be used in an utterance based challenge system based on the articulation scores. Thereafter speech input by an entity using the challenge item database can be processed to distinguish between a human and a machine synthesized voice.

[0062] Another aspect concerns a method embodied in a computer readable medium of selecting challenge data to be used for accessing data and/or resources of a computing system comprising: selecting a candidate challenge item which can include text words and/or visual images; measuring first acoustical characteristics of a computer synthesized utterance when articulating challenge content associated with the candidate challenge item; measuring second acoustical characteristics of a human utterance when articulating the challenge content; generating a challenge item score based on measuring a difference in the first and second acoustical characteristics; and designating the candidate challenge item as a reference challenge item when the challenge item score exceeds a target threshold. Thereafter speech input by an entity using the challenge item database can be processed to distinguish between a human and a machine synthesized voice.

[0063] In preferred embodiments the challenge item score is also based on one or more topics associated with the text words and/or visual images and which are identified and measured in the computer synthesized utterance and human utterance respectively. The challenge item score can also be based on prosodic elements associated with the text words and/or visual images and which are identified and measured in the computer synthesized utterance and human utterance respectively. Alternatively or in addition to this, the challenge item score is also based on a collaborative filtering score generated by measuring responses to a sequence of two or more of the candidate challenge items identified in the in the computer synthesized utterances and human utterances respectively. The collaborative filtering score is preferably derived by identifying at least a first reference correlation between at least a first reference text descriptor for a first challenge item and a second reference text descriptor for a second challenge item, including a probability that a human reviewer identifying the first reference text descriptor when presented with the first challenge item also provides the second reference text descriptor when presented with the second challenge item, or vice-versa. Alternatively or in addition to this the collaborative filtering score is derived by identifying at least a first reference correlation between a first challenge item presented in incomplete form, and a predicted response for completing the challenge item.

[0064] Still another aspect concerns a method embodied in a computer readable medium of selecting challenge data to be used for accessing data and/or resources of a computing system comprising: defining a plurality of demographic groups, the demographic groups being based on age, sex and/or domicile; providing a plurality of CAPTCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) challenge items consisting of a combination of images and solicited utterances with the computing system; for each of the challenge items using the computing system to compare a first reference response of a machine entity and a second reference response provided by a representative of the demographic group; for each demographic group selecting an optimal set of CAPTCHA challenge items determined by the computing system to yield the greatest response difference over the machine entity. Thereafter speech input by an entity using the challenge item database can be processed to distinguish between a human and a machine synthesized voice based on identifying a demographic group for an entity.

[0065] Other aspects of the invention concern a challenge apparatus or system for identifying a source of data input to a computing system comprising one or more software routines implemented in a computer readable medium and adapted to cause the challenge system to perform the aforementioned operations o the various aspects described.

DESCRIPTION OF THE DRAWINGS

[0066] FIG. **1** illustrates a general architecture of a preferred embodiment of an articulated utterance based challenge system of the present invention;

[0067] FIG. **2** shows a more detailed block diagram of the main components of a preferred embodiment of an audible-based challenge system of the present invention;

[0068] FIG. **3** shows a detailed block diagram of the main components of a preferred embodiment of a content challenge compilation system of the present invention;

[0069] FIGS. **4**A-**4**C depict various forms that can be used for presenting content to an entity in accordance with the present invention;

[0070] FIG. **5** describes the basic steps required by an entity to decode content and express an utterance related thereto;

[0071] FIG. **6** shows a conventional prior art CAPTCHA system based on a visual challenge.

DETAILED DESCRIPTION

Basic Elements & Concepts Employed in Invention

[0072] FIG. **1** depicts a typical computing system **100** environment which can benefit from embodiments of the present invention, namely, in allowing flexible control over access to computing resources, data, etc., using a spoken language challenge approach. As can be seen in this figure, an "entity" (which may be a natural person or a machine) **101**, **101'** provides input to a data capture interface **110**. In a preferred embodiment the interface can be based within a GUI or a VUI, including within a web/voice browser, a cell phone, a PDA, a desktop computing system, or similar electronic system including consumer electronic devices (such as set top boxes, cameras, conventional phones, etc.)

[0073] Again while the invention relies primarily on extracting audio information from the entity attempting access, other modes of data may be provided at the same time through other mechanisms (i.e., through keyboards, mice, etc.). The invention therefore can be integrated as part of a

multi-modal input device. A Spoken Language Challenge System **120** is responsible for receiving the input from the entity (i.e., typically in the form of a short speech utterance) and determining whether it is a human or a machine. In most instances the challenge logic of the invention would be implemented as one or more software routines executing on a server computing system. However the invention is not restricted in this fashion, and those skilled in the art will appreciate that some components of the challenge logic could be implemented advantageously on a client side as well or be embodied as firmware/hardware depending on the platform.

[0074] Depending on the results of the Spoken Language Challenge System **120** analysis, the entity may be granted access to one or more computing resources **130**, which, again, may typically include such things as being allowed to access resources/data of a computing device, access an online game, set up an online account, access an online ID or account, access an account through an interactive voice response system (such as by VXML coded pages), or access other types of data, such as email. For other applications it may be linked as a condition to: a) posting an item (electronic data or URL tags) on a message board, web log, an auction site, a content reviewing site (i.e., books, movies, etc.) etc., b) posting a recommendation for an item/service (i.e., as a recommendation system protector to reduce the effects of improper shilling attempts by third parties; c) sending an email (i.e., a receiving email system may insist on a voice authentication to confirm that the message was communicated by a human, as a tool for reducing spam); d) selecting an electronic ad presented by an advertising system—i.e., as a tool for reducing click fraud. These are but examples of course and other applications will be apparent to those skilled in the art.

[0075] Again the implementation will vary from application to application depending on the particular needs of a system operator who desires to prevent/restrict automated machines from obtaining access to certain resources. In a typical example, as noted above, CAPTCHA systems are typically employed to prevent entities from signing up for multiple free email accounts; those skilled in the art will appreciate that there are hundreds of other different applications. One critical aspect of all CAPTCHAs is that they must be easy to use by the majority of the population, or they will simply frustrate users and reduce incentives to utilize a particular computing system. Because the large majority of the population is able to read and articulate basic words, the present invention satisfies this criterion as well. In fact, for many cultures, it is likely that visual CAPTCHAs may be inefficient and/or unusable due to inherent features or limitations of a native alphabet. Furthermore in many cases persons may be visually impaired and unable to use a visual based CAPTCHA. In this respect the present invention can complement such existing systems to provide a wider range of access for such persons.

[0076] In general terms, the Spoken Language Challenge System exploits other fundamental strengths that humans have over machines at this time, namely: 1) the ability to rapidly recognize the meaning of a sentence of words; and 2) smoothly articulate the text of such sentence in the form of speech. Humans are trained over several years to understand proper timing, pitch, prosody, phoneme articulation, etc., and articulate/pronounce words with ease. Machines are simply not able practically to duplicate this knowledge or reproduce such fluency. Consequently, as a basic premise it is submitted that a machine "impostor" will be detectable more easily by using human biometric information which must be expressed by an entity attempting to gain access to a computing system, not just simply as "understood" (as in visual and audio CAPTCHAs) by such entity and expressed in text form. The latter limitation of requiring an entity to respond only with text, or mouse clicks, fails to exploit one additional significant human processing/expressive feature which is extremely difficult (if not impossible) to duplicate.

[0077] For example, as shown in FIG. **5**, the basic functional steps performed by a human reader and speaker are shown in an articulation process **500**. These tasks include the following: identifying the words/text of a sentence at **510**; understanding a meaning of the sentence at **520**; evaluating prosodic aspects of the sentence at **530**; and finally at **540** articulating the text of the sentence with correct pronunciation and appropriate prosody.

[0078] Currently conventional computing systems can duplicate such tasks, but only with noticeable/detectable side effects. That is, an optical scanner routine and conventional natural language understanding system can perform steps **510-530** with reasonable performance, but not nearly as quickly as a human. More importantly, in performing step **540** a text to speech (TTS) system can articulate the result, but with very noticeable results. Thus while they are able to perform all of these operations in some fashion, they are unable to perform them all in seriatim in real-time to duplicate a human.

[0079] For instance suppose one or more of the following sentences are presented to an entity within an interface accompanied by a request to vocalize the same as an input to a challenge system:

[0080] I did not permit him to go outside!

[0081] I need a permit for my construction.

[0082] In doing the same tasks in FIG. **5**, a machine imposter must first "see" the text presented (step **510**). This task in itself would require some form of optical scanner and character recognition to identify distinct words. The conventional CAPTCHAs, as noted above, in fact rely primarily on being able to fool machines by rendering such text unrecognizable. However, it can be seen that unlike the present invention, such approach only takes advantage of a small fraction of the processing battle which can be imposed on a machine impostor.

[0083] After identifying the raw text, the machine imposter must then parse the entire sentence to make sense of the meaning of the phrases (step **520**). This, again, typically requires a natural language engine (NLE) which must be fairly advanced and most often "trained" for particular domains to understand pre-defined phrases from a known grammar. In a completely unstructured context devoid of additional cues, a NLE would be impaired and confused trying to "understand" a random sentence of words—particularly sequences which may be designed to exploit frailties in such systems. For example in the sentences above, the system would have to detect very quickly whether the word "permit" was being used as a verb or a noun. This must be identified correctly, of course, to inform a TTS engine of the correct pronunciation. It can be seen that event random text analysis by a NLE without context is extremely challenging, because several aspects of the sentence must be processed very rapidly to perform word/sentence segmentation, abbreviation expansion, numeral expansion, word pronunciation, and homograph disambiguation.

[0084] After understanding the sentence, the machine imposter may also have to annotate the output of the desired articulation with appropriate prosodic elements at step **530**. Acoustical aspects of prosodic structure also include modulation of fundamental frequency (F**0**), energy, relative timing of phonetic segments and pauses, and phonetic reduction or modification. For example a phoneme may have a variable duration which is highly dependent on context, such as preceding and following phonemes, phrase boundaries, word stress, phrase boundaries, etc.

[0085] For the first sentence therefore there may be different prosodies or emphases that can be appropriately placed on different words. For example, depending on the context, any of the bolded words might be emphasized by a speaker to give a more precise meaning:

[0086] I(1) (in the context of who gave permission)

[0087] did not(2) (in the context of affirmation/confirmation)

[0088] permit him (3) (in the context of who was given permission)

[0089] to go outside! (4) (in the context of location)

[0090] Others are of course possible to give different meanings. To reinforce a desired meaning and accompanying prosody, the sentence can be presented with visual clues or other sentences to assist the speaker in determining which context is appropriate. Nonetheless as is apparent from the ambiguity of the sentence, it may not be necessary to place too much emphasis on particular prosody or emphasis given by the entity. As a first layer of defense or threshold it may be sufficient to simply detect if at least one reasonable choice is given as a response for the context in question—i.e., that prosody was indeed used in expressing the sentence. The prosody choice may also be tied to a prior statement, such as: Did you say he could go outside? No, I told him he could go to his room. In the context of location being emphasized in the question, the answer should similarly give prosodic emphasis to that same element.

[0091] Note that in many cases prosody can be determined solely by reference to acoustical characteristics of the speaker, and without requiring the recognition of words. In some instances it may be desirable nonetheless to compute an additional prosody score based on an identity of the articulated words, and comparing said it to a reference prosodic score related to a content of the overall challenge item set of words.

[0092] However step **540** presents the biggest challenge to the machine imposter. To imitate such behavior, a TTS system must know the phonemes to be uttered, which words to accent (i.e., emphasize), and how to intelligently process the sentence to imitate prosodic behavior of a human. Again for each phoneme, the TTS system must decide on the time behavior of the articulation—i.e., duration, intensity stress and intonation, all of which can change the meaning of a statement entirely. Thus the TTS system must be given accurate information by the NL engine. For example in the first sentences above the system would need to know to accent the last syllable of the word "permit" and the second syllable in the second sentence to provide the accurate syntax.

[0093] Finally, TTS systems also have significant problems in the actual articulation of speech as noted in step **540**. Typically pitch range is restrained controlled because a TTS system may be unsure of how to properly adjust pitch, stress, etc. for a particular phoneme/word. This gives the speech a pronounced and detectable mechanical sound.

[0094] Modern TTS systems rely primarily on concatenative synthesis—a technique which relies on extracting model parameters from actual speech waveforms and concatenating them individually to create new utterances. Parts of utterances that have not been previously processed and stored in the database are constructed from smaller units.

[0095] The challenges for this technique include the fact that it is computationally extremely complex to determine which waveform "unit" to select in any particular instance, and then how to modify the same to be prosodically correct (i.e., desired pitch, duration, intonation, etc.) as noted above. Other techniques have similar limitations.

[0096] To address coarticulation, diphones—representing transitions between phones—are typically used. In the diphone approach, all possible diphones in a particular language are stored, and then they are merged to correspond to the phonetization of the input text. A diphone again is a speech unit consisting of two half-phonemes, or of the phonetic transition in between, e.g. "Cat": silence +c−c+a−a+t−t+silence

[0097] Another method, unit selection, collects its speech data from a depository containing units of various lengths, including diphones as well as words and phrases. Each prerecording is stored in multiple occurrences, pronounced in different prosodic contexts. This type of synthesis requires an extensive storage facility, and has only recently become a popular method, since memories and performance of computers have increased.

[0098] From perusing any of a number of speech synthesis sites, including those associated with the most advanced speech synthesis engines (AT&T, IBM, etc.) it is apparent that clipping by TTS systems is very noticeable. That is, transitions between phonemes are often accompanied by sharp distinguishable breaks. Consequently current machine imposters attempting to imitate a human voice merely try to model the human brain's understanding of the parts of speech of a sentence, and make statistical guesses about the proper pronunciation of words. Real-time modeling of phonemes and stress placement are (essentially) impossible tasks at this time.

[0099] The present invention is based on the hypothesis that given the complexities of language, a computer will not be able to imitate a human in a manner that will not be detectable—at least not in a way that is not detectable by another computer trained to "listen" for machine impostors. In this latter respect, therefore, while a TTS system may eventually reach "human-like" performance, it will always include small deficiencies, artifacts and tell-tale signature signs of speech synthesis operations which will be observable and measurable by conventional speech recognition systems of equal or better computing capability.

[0100] Accordingly the present invention exploits the fundamental premise that speech synthesis systems will invariably lag in function, performance, capability, etc., as compared to speech recognition systems, simply because of the relative complexities of the tasks which systems must perform. The latter has the advantage that it can be trained extensively with millions of examples of human speech, as well as computer speech, to very quickly, easily and accurately differentiate a human from a machine. So in a sense it can be said that the best mechanism for catching a machine impostor is by using another machine as a detector, since the speech "decoding" process is inherently capable of detecting

those aspects of a speech "encoding" process which identify a machine synthesized articulation.

[0101] In summary, TTS systems at best can merely try to model the human brain's understanding of the parts of speech of a sentence, and make statistical guesses about the proper pronunciation of words. Consequently they will always suffer to some extent from one or more of the following deficiencies:

[0102] Pauses

[0103] Misplaced stress of phonemes

[0104] Misplaced stress of words

[0105] Discontinuities between phonemes, syllables

[0106] Incorrect prosody

[0107] All of which can be detected by a spoken language challenge system. It will be understood, of course, that a spoken language challenge system **120** may not be appropriate for all applications, because they may lack or have limited means for an audio input. Nonetheless, for some applications, particularly in portable systems (such as cell phones, PDAs, regular phones) where there is very little (or no) physical space in a display (or inadequate resolution) for presenting a visual CAPTCHA, it is expected to be far more useful than conventional approaches. Moreover, in some instances, such as IVR systems, the present invention is one of only a few known mechanisms for preventing unauthorized entry by machine impostors.

[0108] In some applications a system operator may want to use or even combine different forms of CAPTCHAs for extra security so the present invention could be used to supplement conventional visual and audio based techniques. For example, both visual and audible CAPTCHAs could be used in a hybrid challenge system. An entity could be presented simultaneously with a number of visual distinct word CAPTCHAs arranged in the form of a sentence. The entity is then required to read an entire sentence of words that are each visually confounded, thus increasing the chances that a computing system will fail to process such data in a reasonable time frame.

Speech Verification System for Detecting "Liveness" of Human User

[0109] Having considered the limitations of machine imposters, an explanation of a preferred embodiment of an audible-based challenge system is now disclosed.

[0110] In a first preferred embodiment **200** shown generally in FIG. **2** advantage is taken of the fact that the human vocal tract is a highly complex acoustic-mechanical filter that transforms air pressure pulses into recognizable sounds in a manner that is not duplicatable (at this time) by a machine. Thus this embodiment exploits the difference between a conventional TTS loudspeaker based articulation, and a human vocal track articulation.

[0111] The architecture of the system is as follows: a spoken utterance (from a human or a machine) is captured by a routine **210**. The types of routines suitable for capturing speech data from within a browser, or from a cell phone, are explained in detail in U.S. Pat. Nos. 6,615,172, 5,960,399 and 5,956,683 (Qualccomm) respectively. The current ETSI Aurora standard (ES 201 108 Ver. 1.1.3) available at an http site at portal(dot)etsi(dot)org/stq/kta/DSR/dsr.asp also describes such process and is incorporated by reference herein.

[0112] In conventional speaker verification systems, speaker identities are typically confirmed using a biometric evaluation of the speaker's articulation, which includes measuring such human traits as lung capacity, nasal passages and larynx size and storing these as physical biometric parameters within a template. A person's tongue can also influence their articulations. This same technique can be exploited here. Thus a speech utterance is captured, and selected acoustic features are extracted by a routine **210** which best correspond with—and identify—the particular biometric parameters unique to that person. Thus, it is well accepted that the audio spectrum of a voiced sound from a person inherently carries with it sufficient information to uniquely identify a vocal tract of such person and act as a biometric fingerprint.

[0113] The utterance is presented in response to a sentence selection/presentation logic routine **220**, which is responsible for providing the user with appropriate content **221** to articulate. For example, the system may ask the user to speak the following sentence:

[0114] This is how we recognize speech

[0115] Other choices for the content to be articulated are discussed below in connection with FIGS. **3** and **4**A-**4**C. The challenge item/content **221** may in fact be articulated by the challenge system with a request that the entity seeking access repeat the content. In other embodiments it will be desirable to overlay the present system with a conventional visual based system, so that, for example, the sentence above is shown as a sequential set of visually distorted words. By carefully selecting content for the verification based on known machine weaknesses the accuracy of the system in distinguishing between humans and machine imposters can be increased.

[0116] The appropriate acoustic features of the user utterance are then compared by a routine **230** to determine a best match HPn against a voice print **250** of a known human speaker. This aspect of the invention, so far, can simply be based on a conventional speaker verification system (such as one currently offered by Convergys) to perform one phase of the verification. A company known as voiceverified offers a similar solution at www(dot)voiceverified(dot)com.

[0117] The present invention goes beyond that, however, to cover instances where the speaker is unknown, but should still be given access. If the speaker is unknown, or if speed is a consideration, one alternative for a faster determination is to determine a distance to a nominal human voice print reference **280**. Based on these results, the speaker may be either expressly identified (in the case of a human match) or only tentatively classified (in the case of no existing match) as a human speaker.

[0118] The other phase of the comparison simply modifies the prior art architecture to compare the unknown speaker utterance extracted features with a routine **231** against known machine speaker templates **251**. Again if the speaker is unknown, or for speed purposes, one alternative for a faster determination is to determine a distance to a nominal machine voice print reference **280**. Based on these results, the speaker may be either expressly identified (in the case of a machine match) or only tentatively classified (in the case of no existing match) as a machine speaker.

[0119] In a final evaluation, a comparison is made by a routine **260** against the closest human match HPn and the closest machine MPn to make a determination on whether the speaker is human or not. Based on the results of this algorithm, which may be implemented in any number of ways known in the art, a decision/output is provided at **261** to allow or deny access based on such confirmation. For some applications it may be desirable to further include a time element in

the determination, so, that, for example, a user is given a certain period of time in which to articulate or identify the challenge item/text. The amount of time can be based on thresholds established from real world examples so that a human speaker is expected within a certain confidence level to return a response. If the speaker does not articulate the sentence within this period this factor may be used to weight the evaluation or in fact use it as an outright rejection depending on system requirements and specifications.

[0120] Note that in those cases where only a reference print is used, this aspect of the invention modifies the prior art system to discriminate not between individual users, but, rather, to determine simply if an entity providing an utterance has one or more measured vocal tract characteristics above certain thresholds.

[0121] The two verification routines can also be done in parallel, on different computing platforms adapted and optimized for each type of verification. Thus, for example, a first provider may be employed to determine a voice print of an utterance based on an ISP or other profile information for the user associated with the request. It is possible, for example, that certain population groups would be better identified by reference to a database of voice prints **250** unique to a certain online community/population. This would be faster than attempting to store/access a large database online of several million potential matches.

[0122] Some entities may have particular expertise in resolving/identifying human voices based on their expertise in classifying such voice prints. The invention contemplates scenarios, therefore, in which a machine decision criteria routine **240** may determine (based on a cost parameter, an accuracy parameter, a speed parameter, etc.) to allocate the task of classification to one or more verification providers (not shown) who in turn would each have their own voice print databases, verification logic, etc. Incentives could be provided to those entities responding more quickly, or more accurately, or with less cost.

[0123] In other cases it may be desirable to conduct an auction for the right to resolve the verification question. For such situations, a verification entity may opt to bid on specific types of utterances, depending on the language of the utterance, the purported identity of the user, the country of origin of the speaker, the type of sentence selection logic used, and so on. The bid may be in the form of a positive or negative credit, so that in some instances the verification entities have to pay a certain rate for the right to classify the speaker, while in others they may get a credit of some sort for performing such task.

[0124] All of the above would apply equally well to the machine voice print verification process. It is expected that different entities, over time, will develop unique libraries of machine voice prints that can used for such process. In some embodiments it may be desirable, therefore, to submit the task to multiple verifiers to ensure a more accurate response.

[0125] Techniques for creating libraries/databases of voice prints **250**, **251** are well known. These systems are able to create digital voice "prints" **250** capturing the unique physical characteristics of a speaker's vocal tract. The set of voice prints **250** are created using standard enrollment procedures, which can be duplicated in embodiments of the present invention as well to create new voice prints if desired. For example in banking services, the user is prompted later to repeat a certain phrase or password to gain access to an account. Descriptions of such systems can be found in U.S. Pat. No. 6,356,868; and U.S. publication no. 25096906; and WO005020208 all of which are incorporated by reference herein.

[0126] Similarly speech processors which can create voice templates from acoustic features, and verify human identities from their vocal patterns are well-known in the art; for example systems by Scansoft (Speech Secure) and Nuance (Verifier) can perform this function. The aforementioned WO005020208 application uses a different form of voiceprint which does not use spectral features, but is claimed to be more compact and efficient because it employs only sets of rational numbers which are invariant from utterance to utterance. Other suitable examples will be apparent to those skilled in the art.

[0127] Consequently, it is relatively simple to obtain a number of actual samples from live persons, with different enrollment words/phrases to store as reference voice templates **250**, **251**, and to generate reference voice templates **280**, **281**. Because most voice verification systems are text dependent, it is preferable that the number of enrollment words/phrases be relatively large, and augmented or changed on a regular basis, to make it difficult for a machine to "learn" any of the available choices. In any event, at the end of the verification process, the information from the speaker utterance is used to update a human/machine voice print databases **270**, **271** respectively as needed.

[0128] To identify a closest voice print match, the human language comparator **230** and machine articulation comparator **231** may use a distance between vectors of voice features to compare how close they are to a particular person/machine. For example, the integral of the difference between two spectra on a log magnitude may be computed. Alternatively the difference in the spectral slopes can be compared. Other examples will be apparent to those skilled in the art.

[0129] In the absences of access to a library of voice prints, and to train/bootstrap these systems, it may be appropriate to create the human voice print reference **280** (and machine counterpart **281**) by normalizing across a large sample to create a set of average pseudo-or human proxy representative vectors. The new sample utterance is then compared against such reference.

[0130] As seen in FIG. **2**, the speaker is presented with a challenge item (in this instance text) to articulate. In other cases, the computer may ask the speaker to repeat a sentence as part of the challenge item content. The types of spoken challenge items preferably used in the present invention are described below with reference to FIGS. **3** and **4**A-**4**C.

[0131] In some implementations the speaker may be asked to change their position relative to the microphone, such as by placing it closer or nearer in order to better characterize the acoustic features. Consequently a series of articulated challenges at different positions could be used to assess the nature of the speaker.

[0132] It can be seen that the preferred embodiment treats all humans essentially as one class of potential users of a computing system. Similarly, all machines are classified within a second class which is always excluded. Nonetheless it may be desirable in some embodiments to have humans which are excluded, and conversely machines which are expressly included. For example some environments may call for only authorized machines to participate.

[0133] Thus for some embodiments it may be desirable to use the system to exclude even human participants who are detected and matched against a human voice print associated

with a person who is to be denied access. This would be beneficial in many instances for setting up accounts, to prevent users from setting up multiple accounts under different pseudonyms, such as is the common practice in online message board systems.

[0134] It may be advantageous to see that certain users should be excluded, even if they are human, due to undesirable behavior in other applications (i.e., duplicate accounts, spamming, terms of service violations, etc.) Such persons could be "locked out" by requiring them to sign in with their voice print when they attempt access. Since the system will "recognize" their voice by comparing to the prior stored template, the user can be denied access in this fashion.

[0135] Consequently in this variant of the invention even if a particular "authorized" user attempts to re-enroll in a different session with a different enrollment phrase/password, the present invention can implement the type of system described in Magee—U.S. Publication No. 2005/0125226 incorporated by reference herein, in which voice prints in a database are continually compared against each other for detecting fraud. If a "match" is determined within certain controllable thresholds, the user can be rejected from further access, or have privileges restricted or removed from any new account. This technique would allow detecting machine impostors as well attempting to create multiple personalities/accounts.

[0136] This variant of the invention is particularly useful in the area of web blogs, where spammers are now hiring large numbers of persons in third world countries to solve conventional CAPTCHAs implemented at blog sites. The intruders then post large number of spam posts which are undesirable. By using the present invention, a blog site could implement a standard one sentence enrollment test, on a case by case basis, depending on the origin of the enrollee (i.e., a standard CAPTCHA may be used for the remainder of users). After the enrollee "passes" the first test (i.e., he/she is not a machine) the voice print is stored. At a later time, if the same enrollee attempts to create a second account (for the purpose of creating more spam) the system would identify his/her voice, and prohibit such action, because it is difficult to alter one's voice consistently for the same challenge phrase. Alternatively, even if the enrollee is not immediately locked out, it is possible to simply disable access for the second account at a later time after detecting offline that the voice matches to a prior template.

[0137] Similarly the invention could be used to "differentiate" between nationalities, because foreigners speaking English tend to have pronounced accents and similar lack of correct prosody because they are not reared in the particular culture to understand certain nuances of language. Specific reference models could be created on a country by country/culture-by-culture basis. For example the system could be trained to distinguish between far east Indian English speaking persons and Hong Kong English speaking persons. Moreover many of them may have no English fluency whatsoever, so unlike the prior art which relies solely on a keyboard entry, imposing a spoken language test would eliminate such persons as potential intruders, because they could be detected by a number of the same tests imposed on machines noted above. Consequently the invention could be used to automatically lock out (or temporarily quarantine) new accounts originating from users of a certain region of the world which is demonstrated to be primarily spam driven, and further determined to be not reasonably likely prospective audience participants of a particular blog.

[0138] Thus the above methods could be easily employed on an IP address-by-IP address basis, so that only certain known foreign jurisdictions with a history of abusing the system may be scrutinized by the system to detect duplicate accounts. To comply with applicable laws concerning voice recordings the user could be presented with a waiver form, and/or the voice data could be actually "recorded" as a voice print in a computing system located in another jurisdiction.

[0139] Similarly, for the second class of cohorts, typically a challenge system will always deny access to machines. Nonetheless, it may be desirable in some instances to conditionally allow (or only allow) "authorized" machines to obtain access to computing resources, account services, etc. For example a challenge system may offer the entity an opportunity to present an authorization code as part of the process. If the machine can articulate the appropriate passcode, the challenge system may permit access under any appropriate or desirable terms. This may be useful, for example, when a person may want an electronic agent to access information, accounts, etc., and perform transactions on their behalf. In such cases the machine prints are not used to exclude, but rather to confirm the identity of the electronic agent as would be done with a conventional speaker verification system for human speakers.

Spoken Content Challenge Routine

[0140] Of course in any system that is designed to discriminate against a computer imposter one must deal with the inevitable compensation schemes which could be developed to counter the test being used as the determinant. Since the present invention is based on the notion that a computer imposter will not be able to accurately pronounce a randomly selected set of phones, it is useful to consider a few other parameters that can be adjusted to reduce the likelihood of an intruder "beating" the system, such as with a brute force scheme.

[0141] As noted above in some instances it may be desirable to discriminate between a human and a machine solely based on differences in acoustical characteristics in the articulated speech. In other words, without requiring any actual speech "recognition" per se to detect the content of the utterance for its correctness. For other applications it may be desirable nonetheless to detect that the articulated words (or an image) are indeed "correct" for the text in question, or for the challenge item presented as noted below.

[0142] From a basic perspective it is generally accepted that there are least 40 separate English phones, or separate sounds used to articulate words. This means that for a word that contains n phones the number of potential permutations of phones is approximately ${}_{40}P_n$ which is equal to 40!/n! which can be extremely large. Of course it may be easy to simply store a reasonably human facsimile of such phones for all actual English words which is much smaller. However this would have poor performance on transitions between phones compared to a diphone system. Moreover as noted above, the task of picking out the appropriate phone is complex as it depends on syntax, prosody, etc., so this can be used to gain an advantage over an imposter.

[0143] The more advanced TTS systems may use diphones, which are slightly more accurate for reproducing speech. Generally speaking diphones are combinations of phones,

which means that there are approximately 1600 of such in the English language. Nonetheless there are probably only 200-300 "in use" diphones within the grammar of English words. Even with that limitation, however, it can be seen that the for a simple sentence of only a few words that contains N contiguous diphones, the number of unique sounds to be reproduced is at a minimum $20*10^N$. Thus, even when N is only in the range of 10 or higher, this type of articulation challenge very rapidly moves out of the realm of a brute force solution.

[0144] Moreover there may be some diphones which, due to their nature, are much more difficult to reproduce, and thus more susceptible to detection. These can be experimentally identified as noted below, and exploited to impair a machine intruder.

[0145] Accordingly sentence selection/presentation logic **220** is another aspect of the invention which can be tailored specifically to exploit machine weaknesses and human strengths. To wit, this routine is responsible for selecting and presenting the content to the utterance speaker. By careful selection of such content, the chances of detecting a human versus a machine can be increased dramatically.

[0146] A more detailed diagram of the basic architecture for a content selection challenge routine **320** is shown in FIG. **3**. The main component in this instance is a content (in a preferred embodiment a sentence) synthesis engine (CSE) **310**, which receives a variety of data inputs and uses a variety of rules and scenarios to output/construct appropriate content challenge items which can be used by a content challenge routine **220** (FIG. **2**). These candidate sentences are tested against machine and human trainers to develop a suitable body of challenge sentences in a database **395**.

[0147] One body of information which the CSE routine draws upon is a dictionary-grammar **315**. The latter is preferably populated by an entire set of conventional words, phonemes, phrases, etc. for a particular language, as well as a database of homographs, built from a corpus **314**. Homographs are words which are spelled the same, but which have different pronunciations. A human reading a sentence with a homograph (like the words "read," "close" and the like) can be expected to readily perceive the correct context of a word, while a computing machine is likely to require additional time to understand the syntax of a sentence. This time factor, again, can be exploited to differentiate and score the utterance. Thus one basic strategy would be to employ sentences with homographs to handicap a machine intruder.

[0148] The CSE engine **310** also has access to natural language rules **316**, prosody rules **317**, coarticulation/concatenation rules **318** and a scenario builder **319**. These reference rules help the CSE construct appropriate challenge sentences and acceptable pronunciations based on the guidelines noted above.

[0149] For example to develop a set of content challenge items **395**, a variety of sources and strategies could be used. In a first technique, one starts off with a reference corpus **314** of raw prose from any of a number of sources (newspapers, books, Internet, etc) to develop a set of evaluation sentences having a word form W**1**+W**2**+W**3**+ . . . Wk and an associated diphone sequence: {P**1***a*, P**1***b*, P**1***c*} {P**2***a*, P**2***b*} {P**3***a*, P**3***b*, P**3***c*}+ . . . {Pka} and so on. Each of these are then processed with applicable rules **316**, **317**, **318**, etc. to form a candidate sentence **311** that is then evaluated according to various criteria (elaborated below) to determine if it is a useful and suitable discriminator between a live human and a machine imposter. If it turns out to be an appropriate choice it is placed into challenge item set database **395**, and presented as one of the options for a challenge item as seen in FIG. **4**A in the form of explicit text. As noted earlier, for additional security it may be desirable to visually scramble the words using a visual CAPTCHA technique first, and then present the same for articulation. In this hybrid scenario, both acoustic and content features are measured to increase the difficulty of the challenge item for a computing system. To further increase the duration of the challenge (and hence make it unattractive to automated systems) the visual images (which could include depictions of objects and not just distorted text) could be presented in a deliberately paced manner so as to reveal the details over time. Thus the full image would not be revealed until 30 seconds, but would be improved gradually with additional visual details during successive iterations every 4 or 5 seconds. The system could take into account the time it takes an average person to provide the correct answer for the gradually revealed image as part of the scoring.

[0150] Referring again to FIG. **3**, a second technique would involve working backwards, by identifying and compiling a list of diphones and concatenations in a database **396** which are known confounders for TTS systems. From this set of confounders a set of candidate sentences could be automatically constructed or located in a dictionary/grammar **315**. In essence, the strategy here is to measure a set of concatenation gaps/discrepancies for each diphones. Based on developing and sorting such list to identify a preferred set of distinctive diphones **396** that are most likely to confound a TTS system, the system selects sentences from the corpus containing multiple confounding phones, or constructs them from scratch as noted. When constructing sentences from scratch it may be desirable in some cases to use some random word sequences and without considering syntax rules. Thus the end result should be a set of random sentences constructed from word/phoneme patterns which are extremely difficult (i.e. detectable) for a machine to imitate in human form.

[0151] A third technique would combine the challenge sentence with a visual cue of some sort to induce a prosodic response. These can be determined experimentally with reference to a human population by giving an assortment of sentences and visual cues and measuring a response given to the challenge sentence. An example is given in FIG. **4**B in the form of a question that the speaker must respond to, and preferably provide an answer that falls into a category of acceptable responses. Thus the challenge item set database **395** would include such additional cue data as well.

[0152] The visual cues could request information from the speaker such as shown in FIG. **4**B, along with other cues if necessary, such as text that explicitly reads:

[0153] He's hitting a ball

[0154] He's fishing

[0155] He's driving a car

[0156] He's reading a book

[0157] He kicked the cat

[0158] etc., to help give a roadmap to a human speaker.

[0159] Note, too that the visual cues can be tested with a population of humans to determine which ones are most likely to induce a preferred set of reference prosodic characteristics. Since it is desirable to exploit this difference over machines, it would be useful to identify those visual cues that result in a maximal prosodic difference over a reference computer synthesized voice and store these in a database along with their prosodic scores. These examples can be easily determined with routine experimentation, and can be tailored

if desired to a detected gender/demographic of the entity attempting access. Thus the challenge items and associated cues can be customized within the database with fields pertaining to a particular estimated audience that is attempting access. The initial determination of gender, age, demographics, etc., can be done using any number of known techniques.

[0160] A fourth technique would rely on a lack of any explicit text cues. Instead it would offer up a set of pictures/images/cartoons with a prompt for the speaker to give a free form response. The responses are then tabulated and correlated to identify a set of thresholds identifying an acceptable response. For example, in response to picture A, the set of responses included X—50%, Y—35%, Z—5%, other 10% and so on. A statistical profile can then be used against an entity attempting access to see if the response matches one of the human provided answers in the reference set. As before, the images can be tested and optimized by reference to prosodic scores as well so that a database of preferred challenge items is developed.

[0161] An example of this is shown with reference to FIG. **4**C. The challenge item may be in the form of a bubble reader, that asks an open ended question like "what did this person say" or "what is this person doing"? etc. Specific suggestions can be given to the speaker, as well, to induce an accurate human response. Other basic questions might include facts which the speaker will know without reference to another source, such as:

[0162] what is your full name?

[0163] describe for me what color your clothes are

[0164] tell me what you ate for breakfast?

[0165] tell me all the good things that come to mind when you think about your mother?

[0166] The example sets can be augmented with collaborative filtering technology as well, so that correlations can be developed between humans on appropriate responses to challenge items. Thus, if two humans responded with the same answers to multiple challenge items, these individuals' profiles can then be compared to an unknown speaker to present yet another form of human identity detection. For example, if a large percentage of persons who see X say A in response, and such persons also, upon seeing Y say B in response, than a content challenge item can be presented to show both X and Y to detect the same correlation in responses. Of course such test may not determinative but it can act as a further overlay in the access determination along with the required correct pronunciation of the words in question.

[0167] Along the same lines the challenge item could be presented in the form of a game or puzzle in which an image or a set of words could be presented to the entity with a request to predict or guess the next item in sequence. The items could be selected, again, based on a Bayesian type analysis of actual interviews with humans which solicit specific responses based on human associations. For example in the color sequence:

[0168] Green

[0169] Yellow

[0170] (please fill in the blank)

A very common choice would be red, representing the colors on a traffic light. Other examples for other applications will be immediately apparent to those skilled in the art.

[0171] Similar evaluations of prosody scores can be made to identify prosodic characteristics of particular challenge items. By identifying and ranking common prosodic elements presented by humans in articulated sentences, the system can use such scores to compare against an unknown entity for prosodic (dis) similarities.

[0172] As noted, scenario builder **319** is the entity which is responsible for compiling appropriate qualified content to be evaluated, as well as to create the pictorial challenge item sets shown in FIG. **4**B and **4**C. This can be done by reference to a library of images/sequences, reference facts, context rules, question and answer sets, etc. To determine this set a supporting routine may examine logical explicit tags provided by humans on the Internet for image data and the like. For example a website or search engine may contain an image with the tags Cat, Pet provided by human taggers for a picture of a household cat. This tagging data then provides a reasonable starting point for which to provide the explicit cues in the aforementioned figures along with the challenge item text.

[0173] In other instances the scenario builder may suggest a sequence of question/answer sets to be used for the challenge item content, so the accessing entity is effectively engaged in an interactive dialog session. In this manner the timeliness and naturalness of the responses could be evaluated over a series of multiple questions. Simple questions such as "tell me what you are wearing" would get a first answer, and a follow-up could be built dynamically—based on the detection that the entity said ". . . a shirt"—to ask what color the garment is, and so on. An example of another question would be "What movie have you seen recently." If the answer was "Star Wars," for example, the system could follow up with "who played Darth Vader" or "who was your favorite character," or some other pre-stored question/answer associated with the movie, subject to the limitations of maintaining a database of suitable questions and answers. Similar questions could be posed for books, music, sporting events, job occupations, etc. Again the questions and content could be randomly selected and based on a selection of content that is known to present challenges for a machine. The user could be prompted to pick which 1 of a series of topics he/she would like to talk about. Thus a series of Q/A scenarios could be designed and presented to better weed out improper entities, since the likelihood of detection would increase with each utterance.

[0174] Looking again at FIG. **3** the candidate sentences **311** can be presented to one or both of a machine template trainer **330** or a human template trainer **340**. These routines can be embodied in separate computing systems which are adapted to measure and catalog an individual person/machine's articulation of the candidate sentence from a group of such entities. Thus each is responsible for compiling a set of voice prints representing a group set of reference articulations for the candidate sentence. For example, a set of M different machine TTS systems could be presented with the challenge of articulating the candidate sentence **311**, and their responses catalogued and stored. This would aid, for example, in developing a library of known TTS cohorts evaluated with respect to the candidate sentence. Later, during a verification process, an initial determination can be made as to which computing entity best correlates to a particular utterance. Based on such determination a challenge sentence or text can be selected that best discriminates for such machine.

[0175] Similarly a group of N different persons with different demographics could be tested in the same way to develop a set of voice prints for the sentence in question. In a preferred embodiment the sentences could be presented in connection with an online gaming site, so that the articulator data is easily

collected from a wide population base and without having to resort to specialized groups or expense.

[0176] Routine **350** is then responsible for measuring the individual differences between each of the N human audible responses and the M machine audible responses. As noted these differences can be calculated in any number of ways known in the art. This can be compiled in the form of a table/matrix to determine those candidate sentences offering the greatest promise in terms of their ability to differentiate between humans and machines because the distance for such candidates is the greatest. Thus, for example, a threshold evaluation routine **360** may be set to choose only those candidate sentences which represent the highest X % difference in human vs. machine renderings. For example, the top 5% or 10% may be selected to become part of the challenge item set database **395**. The exact criteria will vary according to each application of course according to the desired speed/accuracy/security requirements imposed.

[0177] As a supplement to the scoring routine **350** could present the candidate sentence articulations to a group of human evaluator/raters to collect their vote or opinion on the liveness of the entity (human or machine) responsible for the utterance. The numbers and percentages of acceptance/rejection can be tallied for each candidate sentence as articulated by the M machines. This rejection factor can be used as well for systems which may simply employ a non-automated challenge evaluation, or as an initial filter for reducing a set of candidates for the final challenge item set database. That is, an initial set of candidate sentences may be filtered by a group of human observers, and the ones which nonetheless least human like could be then subjected to more extensive processing to identify more precisely the acoustic differences resulting in the non-human signature. This examination can be done for each machine entity to develop a library of known acoustical differences for different candidate challenge items. This way, after an initial determination is made as to the likely identify of an entity, a randomly selected challenge item can be presented to each that is nonetheless designed to exploit the largest known (significant) deviation from a reference human voice.

[0178] In those situations where a candidate sentence is considered appropriate for the challenge item set, the individual human voice print database and machine voice print databases can be updated. Again, as noted above, it may be useful to compile a simple reference template representing each collective entity (i.e., HPr **280**, MPr **281**) as noted above using aggregate statistics for the entities measured. A notation can be made by a routine **385** to note that the phones/words in question resulted in a useful candidate, and this information again can be used as seed information for finding and retrieving other similar candidates likely to perform well. The individual distinctive diphones can also be stored as desired for later use in a set **396**.

[0179] In some embodiments it may be desirable to construct a statistical language model of machine TTS system behavior to help uniquely identify particular machine entities, and to better characterize overall expected machine articulations. This machine SLM **380** in turn can be used by the machine articulation comparator **231** (discussed above) to further uniquely identify a machine imposter rendition of a challenge sentence. To further train and refine the SLM **380** the system could solicit repeated samples of the same words and phrases to better characterize a signature of a particular unique machine. Moreover to further train the SLM to behave like a human ear, the set of sentences which are uniformly rejected at a high rate by humans can be used as a seed set for training. This will have the effect of biasing the SLM **380** to hear content that is preferred by the human ear for discriminating against a machine imposter.

[0180] Through sufficient training samples the system should be able to identify, with controllable confidence levels, an appropriate set of sentences that are likely to weed out a machine imposter. Moreover candidate sentences which are confusing or take too long for a human can be eliminated as well. Again it is preferable that the challenge sentences include primarily samples that are rapidly processed and articulated as measured against a human reference set. The respective times required by human and machines can also be measured and compiled to determine minimum, maximum, average, mean and threshold times. For example, it may be desirable to select challenge sentences in which the time difference between human and machine articulations is greatest.

[0181] Embodiments of the invention may be used with machines, devices and systems which use a speaker verifier option in accordance with the media resource control protocol (MRCPv2). This protocol allows for speaker verification using a voice authentication methodology that can be used to identify the speaker in order to grant the user access to sensitive information and transactions. It will be apparent from those skilled in the art that this same protocol could be adapted (including with a new dedicated command) to initiate requests for verification of an entity as a human or an authorized machine agent.

[0182] As the proliferation of speech recognition based applications continues, the present invention can be used to detect what might be referred to as "cluck" fraud in the form of improper selections of ads, access to accounts, etc. This would be a superior form of fraud detection since there is no practical mechanism for detecting if a keyboard/mouse selection of an ad is fraudulent.

[0183] Finally, it will be apparent to those skilled in the art that the methods of the present invention, including those illustrated in FIGS. **1**, **2**, **3** and **4** can be implemented using any one of many known programming languages suitable for creating applications that can run on client systems, and large scale computing systems, including servers connected to a network (such as the Internet). The details of the specific implementation of the present invention will vary depending on the programming language(s) used to embody the above principles, and are not material to an understanding of the present invention.

[0184] The above descriptions are intended as merely illustrative embodiments of the proposed inventions. It is understood that the protection afforded the present invention also comprehends and extends to embodiments different from those above, but which fall within the scope of the present claims.

What is claimed is:

1. A method of identifying a source of data input to a computing system: comprising:

a) receiving speech utterance from an entity related to randomly selected challenge text;

wherein said challenge text represents a selected set of one more contiguous words which when articulated have a measurable difference in acoustical character-

istics between a reference human voice and a reference computer synthesized voice that exceeds a target threshold;

b) processing said speech utterance with said computing system to compute first acoustical characteristics of said entity;

c) generating a determination of whether said speech utterance originated from a machine or a human.

2. The method of claim **1** further including the steps: identifying a first computer synthesized voice that best correlates to said entity; and selecting said randomly selected challenge text based on an identity of said first entity so as to maximize a difference in acoustical characteristics.

3. The method of claim **1** further including a step: wherein said challenge text is also selected in part based on a confirmation from a human listener that an articulation of such challenge text originated from a computer synthesized voice.

4. The method of claim **1**, further including steps: soliciting utterances from a plurality of separate computing machines to determine their respective acoustical characteristics; and storing said plurality of associated acoustical characteristics in a database of known computing entities.

5. The method of claim **4**, wherein multiple samples of individual challenge sentences are collected.

6. The method of claim **1** wherein said challenge text is selected in part based on a difference in time for rendering such text into audible form by a human and a computing machine.

7. The method of claim **1** further including a step: granting or denying access to data and/or a data processing device based on the result of step (c), including a signup for an email account or a blog posting.

8. The method of claim **1** further including a step: granting or denying access to an advertisement based on the result of step (c).

9. The method of claim **1** further including a step: performing a separate automated visual challenge test so that both visual processing and articulation processing is considered in one or more of said determinations.

10. The method of claim **1**, wherein a prosody score associated with said speech utterance is also considered during step (c).

11. The method of claim **1** wherein said first test text data consists of a sentence presented in visual form for the entity to articulate.

12. The method of claim **1** wherein said first test text data consists of a sentence presented in audible form for the entity to repeat.

13. The method of claim **1** wherein said first test text data consists of a question presented in visual form and further includes an image cue associated with said question as well as a separate set of acceptable responses presented in visual form.

14. The method of claim **1**, further including a step: selecting one or more second computing systems for separately performing steps (d) and/or (f) based on a performance and/or cost requirement.

15. The method of claim **14**, wherein said one or more second computing systems are selected based on a language spoken by the entity.

16. The method of claim **14** wherein said one or more second computing systems are selected based on an IP address or geographic region associated with the entity.

17. The method of claim **14** wherein said one or more second computing systems are selected based on an auction process in which such one more second computing systems bid for the right to process said speech utterance.

18. The method of claim **1** wherein said challenge text is selected based on a detected accent and/or geographic region associated with the entity.

19. The method of claim **1** further including a step: determining a set of sentences for inclusion as challenge text based on an articulation difficulty score for a computer synthesis engine.

20. The method of claim **19** wherein said set of sentences are automatically generated based on a measured concatenation difficulty for a set of diphones.

21. The method of claim **19** wherein said set of sentences are extracted automatically from a corpus that includes web logs, newspapers, books and/or the Internet.

22. The method of claim **1** wherein discontinuities between phonemes and pauses between words are also measured during step (b).

23. The method of claim **1** wherein processing is distributed so that at least step (a) is performed on a portable computing device and the remaining steps are performed on a server computing device.

24. A method of implementing a CATPCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) to identify a source of data input to a computing system comprising:

a) presenting an image CATPCHA to an entity, which image CAPTCHA includes one or more visually distorted words, phrases or images as a challenge item;

b) receiving a speech utterance from an entity related to said challenge text;

c) processing said speech utterance to generate a determination of whether said speech utterance originated from a machine or a human.

25. The method of claim **24** wherein said image CAPTCHA includes at least two distinct words.

26. The method of claim **24** wherein said determination includes a first score based on computing acoustical characteristics of said speech utterance.

27. The method of claim **24** wherein said determination includes a second score based on recognizing said speech utterance to determine if said one or more visually distorted words, phrases or images are correctly identified.

28. The method of claim **24** wherein said image CAPTCHA is revealed in distinct stages which span a predetermined time period, and with each stage presenting additional visual information.

29. The method of claim **28**, wherein a scoring during part (cc) takes into account a time required for the entity to determine said image CAPTCHA.

30. A method of identifying a source of data input to a computing system comprising:

a) associating a first challenge item with a first set of individual text descriptors; wherein said first set of text descriptors are based on feedback provided by a group of human reviewers;

b) associating a second challenge item with a second set of individual text descriptors; wherein said second set of individual text descriptors are also based on feedback provided by a group of human reviewers;

c) identifying at least a first reference correlation between at least a first reference text descriptor for said first

challenge item and a second reference text descriptor for said second challenge item, including a probability that a human reviewer identifying said first reference text descriptor when presented with said first challenge item also provides said second reference text descriptor when presented with said second challenge item, or vice-versa;

d) presenting the first challenge item to an entity as part of an automated access challenge adapted to distinguish humans from computing machines;

e) receiving speech utterance from said entity related to said first challenge item to determine a first input text descriptor;

f) presenting the second challenge item to said entity as part of said automated access challenge adapted to distinguish humans from computing machines;

g) receiving speech utterance from said entity related to said second challenge item to determine a second input text descriptor;

h) comparing said first and second input text descriptors to identify said reference correlation between them as measured during step (c);

i) generating a determination of whether said speech utterance originated from a machine or a human based on a value of said reference correlation.

31. The method of claim **30** wherein said challenge items include an image.

32. The method of claim **30** wherein said challenge items include a question.

33. The method of claim **30** wherein said first and second challenge items are of different types, including a first type which includes an image, and a second type which includes a question.

34. The method of claim **30**, wherein said first and second challenge items are selected based on a value of said reference correlation.

35. A method of identifying a source of data input to a computing system comprising:

a) receiving speech utterance from an entity related to randomly selected challenge text;

b) processing said speech utterance with said computing system to compute first acoustical characteristics of said entity;

c) comparing said first acoustical characteristics with at least one reference set of acoustical characteristics for a human voice to identify a first score for said speech utterance;

d) comparing said first acoustical characteristics with at least one reference set of acoustical characteristics for a computer synthesized voice in parallel during step (c) to generate a second score for said speech utterance;

e) generating a determination of whether said speech utterance originated from a machine or a human based on said first score and said second score.

36. The method of claim **35** wherein steps (c) and (d) are done by separate entities using separate computing machines.

37. The method of claim **36**, wherein said separate entities participate in an auction to identify said first and/or second scores.

38. A method of identifying a source of data input to a computing system comprising:

a) selecting first test text data to be articulated as a speech utterance by an entity providing input to the computing system;

b) receiving said speech utterance from the entity;

c) generating first recognized speech data from the speech utterance corresponding to said first test text data;

d) processing said first recognized speech data with said computing system to generate an initial determination of whether said speech utterance originated from a machine or a human;

e) optionally repeating steps (a) through (d) based on a confidence score for said initial determination using second test text data, which second test text data is derived dynamically from content presented in said first recognized speech data;

f) processing said second recognized speech data with said computing system to generate a final determination of whether said speech utterance originated from a machine or a human.

39. A challenge system for identifying a source of data input to a computing system comprising:

one or more software routines implemented in a computer readable medium and adapted to cause the challenge system to:

a) receive speech utterance from an entity related to randomly selected challenge text;

wherein said challenge text represents a selected set of one more contiguous words which when articulated have a measurable difference in acoustical characteristics between a reference human voice and a reference computer synthesized voice that exceeds a target threshold;

b) process said speech utterance with said computing system to compute first acoustical characteristics of said entity;

c) generate a determination of whether said speech utterance originated from a machine or a human.

40. A system for implementing a CATPCHA (Completely Automatic Public Turing Test To Tell Humans And Computers Apart) to identify a source of data input to a computing system comprising:

one or more software routines implemented in a computer readable medium and adapted to cause the challenge system to:

a) present an image CATPCHA to an entity, which image CAPTCHA includes one or more visually distorted words, phrases or images as a challenge item;

b) receive a speech utterance from an entity related to said challenge text;

c) process said speech utterance to generate a determination of whether said speech utterance originated from a machine or a human.

41. A system to identify a source of data input to a computing system comprising:

one or more software routines implemented in a computer readable medium and adapted to cause the challenge system to:

a) associate a first challenge item with a first set of individual text descriptors; wherein said first set of text descriptors are based on feedback provided by a group of human reviewers;

b) associate a second challenge item with a second set of individual text descriptors; wherein said second set of individual text descriptors are also based on feedback provided by a group of human reviewers;

c) identify at least a first reference correlation between at least a first reference text descriptor for said first chal-

lenge item and a second reference text descriptor for said second challenge item, including a probability that a human reviewer identifying said first reference text descriptor when presented with said first challenge item also provides said second reference text descriptor when presented with said second challenge item, or vice-versa;

d) present the first challenge item to an entity as part of an automated access challenge adapted to distinguish humans from computing machines;

e) receive speech utterance from said entity related to said first challenge item to determine a first input text descriptor;

f) present the second challenge item to said entity as part of said automated access challenge adapted to distinguish humans from computing machines;

g) receive speech utterance from said entity related to said second challenge item to determine a second input text descriptor;

h) compare said first and second input text descriptors to identify said reference correlation between them as measured during step (c);

i) generate a determination of whether said speech utterance originated from a machine or a human based on a value of said reference correlation.

42. A system to identify a source of data input to a computing system comprising:

one or more software routines implemented in a computer readable medium and adapted to cause the challenge system to:

a) receive speech utterance from an entity related to randomly selected challenge text;

b) process said speech utterance with said computing system to compute first acoustical characteristics of said entity;

c) compare said first acoustical characteristics with at least one reference set of acoustical characteristics for a human voice to identify a first score for said speech utterance;

d) compare said first acoustical characteristics with at least one reference set of acoustical characteristics for a computer synthesized voice in parallel during step (c) to generate a second score for said speech utterance;

e) generate a determination of whether said speech utterance originated from a machine or a human based on said first score and said second score.

* * * * *