US 20140032216A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2014/0032216 A1**

Roth et al. (43) **Pub. Date: Jan. 30, 2014**

(54) **PRONUNCIATION DISCOVERY FOR SPOKEN WORDS**

(71) Applicant: **Nuance Communications, Inc.**, Burlington, MA (US)

(72) Inventors: **Daniel L. Roth**, Boston, MA (US); **Laurence S. Gillick**, Newton, MA (US); **Michael L. Shire**, Cambridge, MA (US)

(73) Assignee: **Nuance Communications, Inc.**, Burlington, MA (US)

(21) Appl. No.: **14/041,857**

(22) Filed: **Sep. 30, 2013**

Related U.S. Application Data

(63) Continuation of application No. 10/939,942, filed on Sep. 13, 2004, now Pat. No. 8,577,681.

(60) Provisional application No. 60/502,084, filed on Sep. 11, 2003.

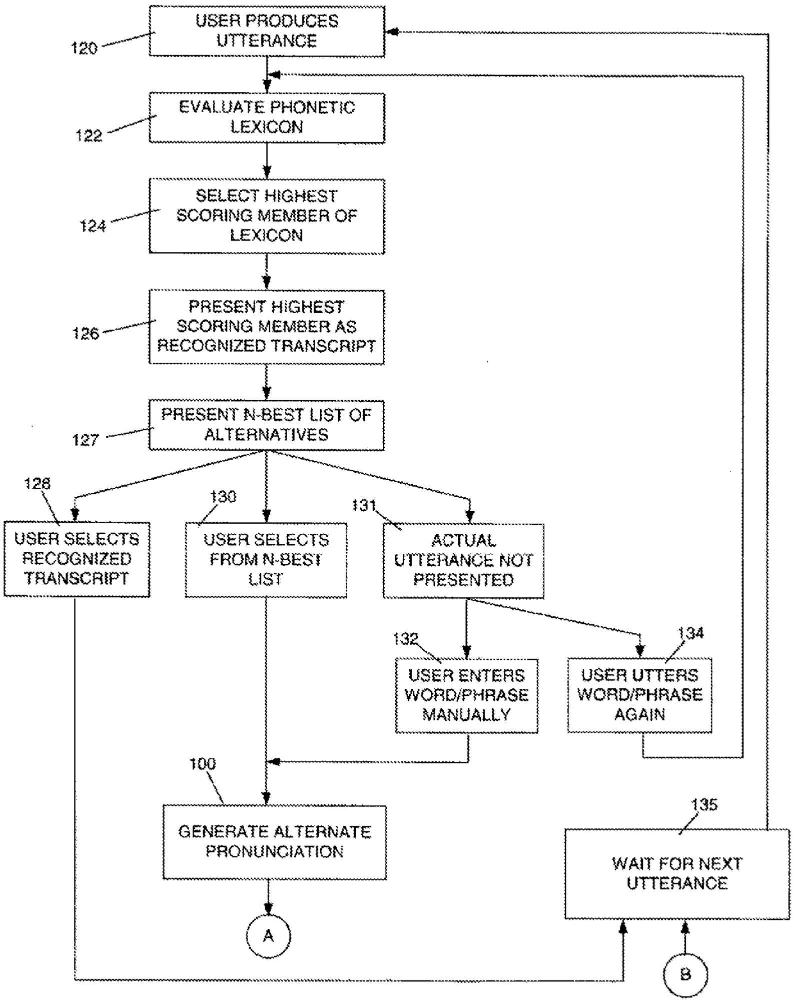
Publication Classification

(51) **Int. Cl.**
G10L 15/06 (2006.01)

(52) **U.S. Cl.**
CPC ***G10L 15/06*** (2013.01)
USPC **704/244**

(57) **ABSTRACT**

A method for a portable device includes receiving a spoken utterance of a word or phrase, generating a plurality of alternative pronunciations of the spoken utterance, scoring one or more pronunciations of the plurality of alternative pronunciations using the spoken utterance, and updating a lexicon with at least one scored pronunciation.


USER PRODUCES UTTERANCE
120
EVALUATE PHONETIC LEXICON
122
SELECT HIGHEST SCORING MEMBER OF LEXICON
124
PRESENT HIGHEST SCORING MEMBER AS RECOGNIZED TRANSCRIPT
126
PRESENT N-BEST LIST OF ALTERNATIVES
127
128
USER SELECTS RECOGNIZED TRANSCRIPT
130
USER SELECTS FROM N-BEST LIST
131
ACTUAL UTTERANCE NOT PRESENTED
132
USER ENTERS WORD/PHRASE MANUALLY
134
USER UTTERS WORD/PHRASE AGAIN
100
GENERATE ALTERNATE PRONUNCIATION
135
WAIT FOR NEXT UTTERANCE
A
B

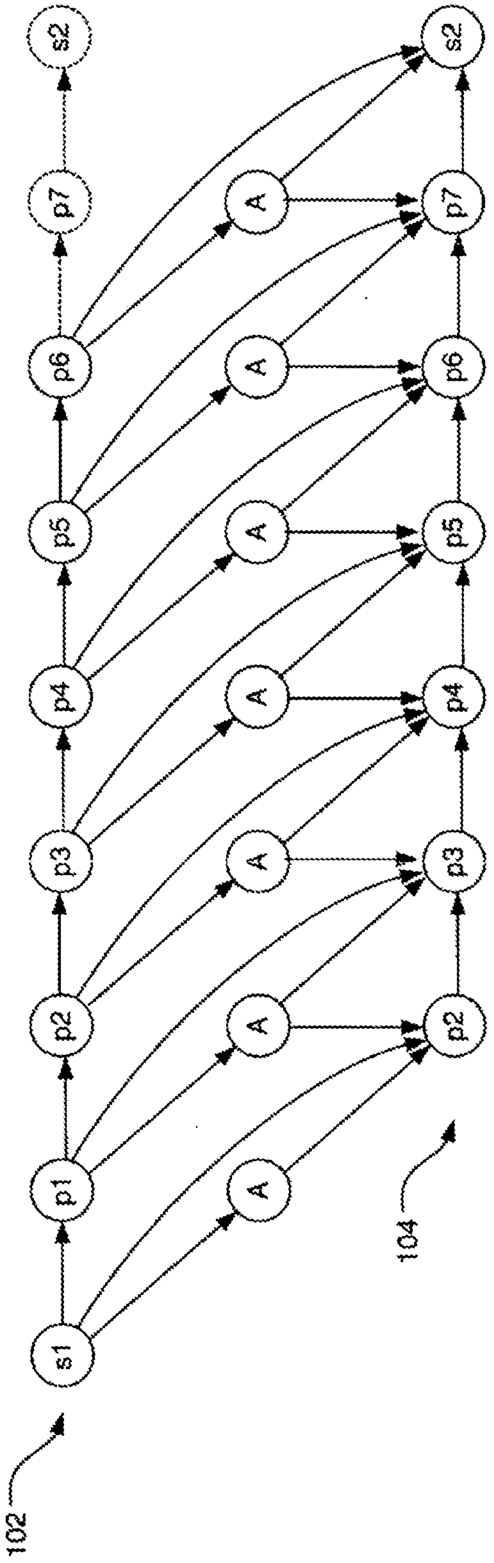


FIG. 1

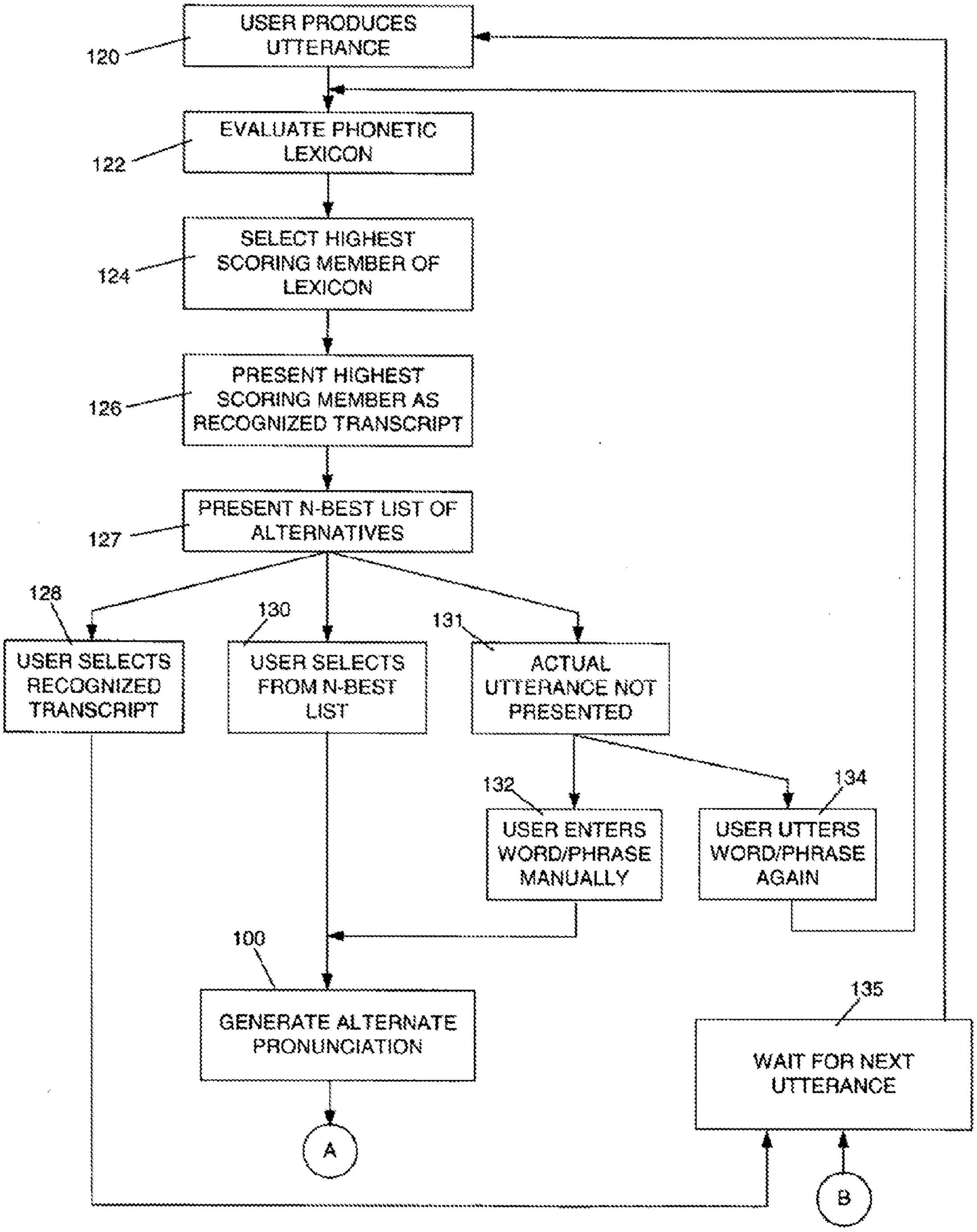

USER PRODUCES UTTERANCE
120
EVALUATE PHONETIC LEXICON
122
SELECT HIGHEST SCORING MEMBER OF LEXICON
124
PRESENT HIGHEST SCORING MEMBER AS RECOGNIZED TRANSCRIPT
126
PRESENT N-BEST LIST OF ALTERNATIVES
127
128
USER SELECTS RECOGNIZED TRANSCRIPT
130
USER SELECTS FROM N-BEST LIST
131
ACTUAL UTTERANCE NOT PRESENTED
132
USER ENTERS WORD/PHRASE MANUALLY
134
USER UTTERS WORD/PHRASE AGAIN
100
GENERATE ALTERNATE PRONUNCIATION
A
135
WAIT FOR NEXT UTTERANCE
B

FIG. 2a

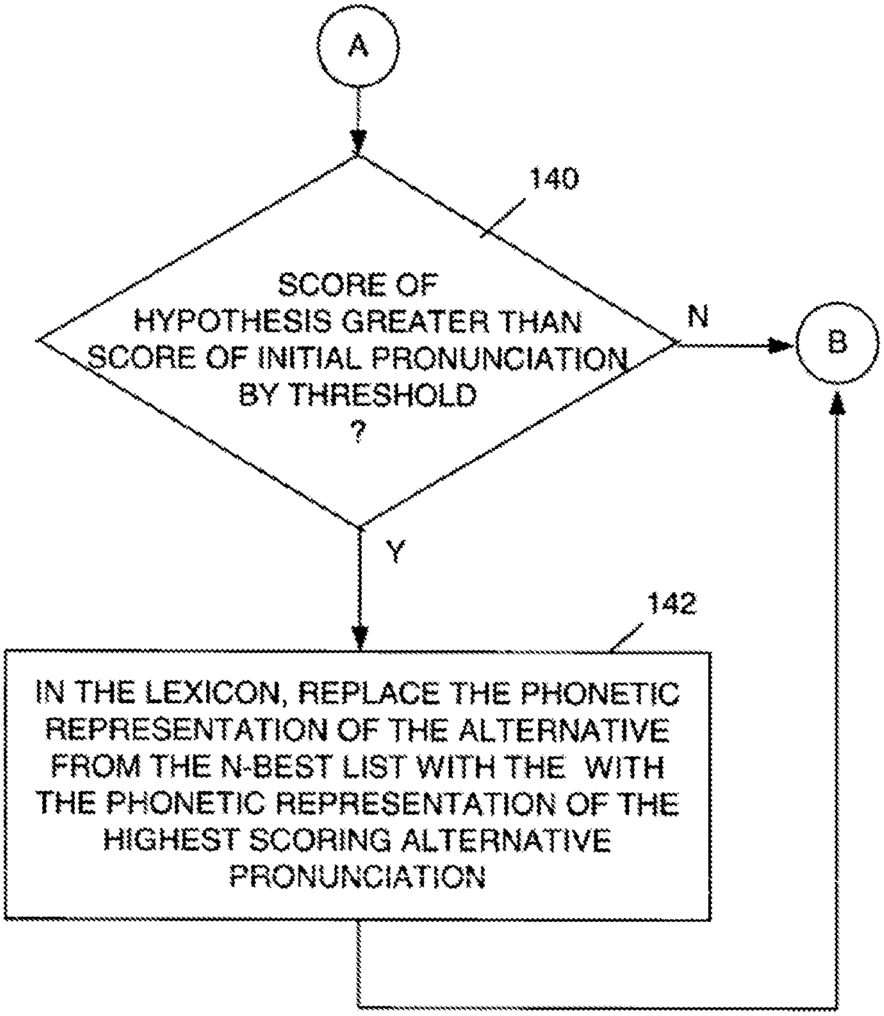

A
140
SCORE OF HYPOTHESIS GREATER THAN SCORE OF INITIAL PRONUNCIATION BY THRESHOLD ?
N
B
Y
142
IN THE LEXICON, REPLACE THE PHONETIC REPRESENTATION OF THE ALTERNATIVE FROM THE N-BEST LIST WITH THE WITH THE PHONETIC REPRESENTATION OF THE HIGHEST SCORING ALTERNATIVE PRONUNCIATION

FIG. 2b

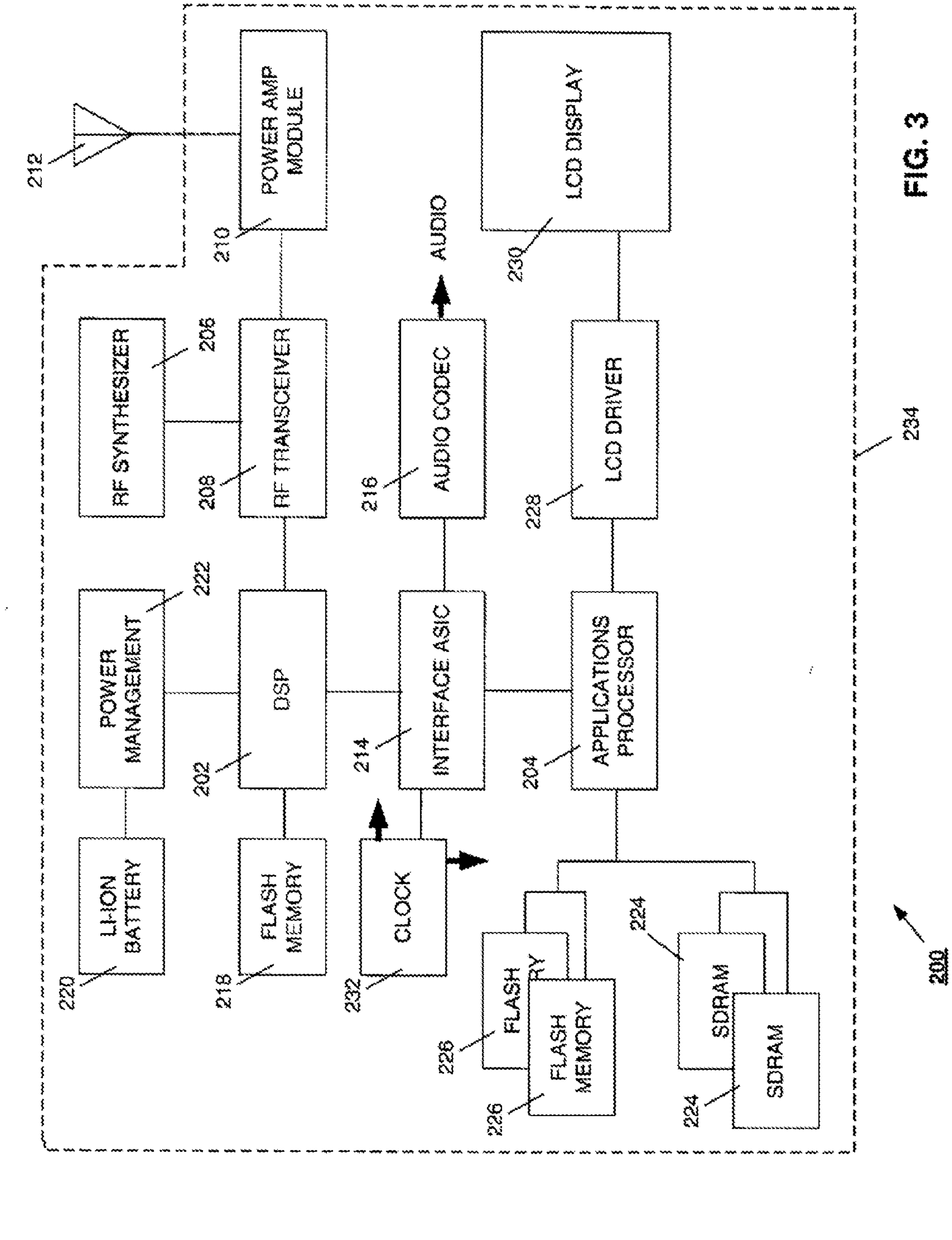

212
POWER AMP MODULE
210
RF SYNTHESIZER
206
RF TRANSCEIVER
208
AUDIO
AUDIO CODEC
216
LCD DISPLAY
230
LCD DRIVER
228
234
POWER MANAGEMENT
222
DSP
202
INTERFACE ASIC
214
APPLICATIONS PROCESSOR
204
LI-ION BATTERY
220
FLASH MEMORY
218
CLOCK
232
FLASH MEMORY
226
FLASH MEMORY
226
SDRAM
224
SDRAM
224
200

FIG. 3

PRONUNCIATION DISCOVERY FOR SPOKEN WORDS

[0001] This application is a continuation of pending U.S. patent application Ser. No. 10/939,942, filed Sep. 13, 2004, which in turn claims priority from U.S. Provisional Patent Application 60/502,084, filed Sep. 11, 2003, which are incorporated herein by reference in their entirety.

TECHNICAL FIELD

[0002] This invention relates generally to wireless communication devices with speech recognition capabilities.

BACKGROUND ART

[0003] Wireless communications devices, such as cellular telephones (cell phones), commonly employ speech recognition tools to simplify the user interface. For example, many cell phones can recognize and execute user commands to initiate an outgoing phone call, or answer an incoming phone call. Many cell phones can recognize a spoken name from a phone book, and automatically initiate a phone call to the number associated with the spoken name.

[0004] Handheld electronic devices (e.g., mobile phones, PDAs, etc., referred to herein as "handhelds") typically provide for user input via a keypad or similar interface, through which the user manually enters commands and/or alphanumeric data. Manually entering information may require the user to divert his attention from other important activities such as driving. One solution to this problem is to equip the handheld with an embedded speech recognizer.

[0005] Due to numerous factors, the speech recognizer may occasionally incorrectly decode the utterance from the user. To deal with such errors, some speech recognizers generate a list of N alternatives for the recognized transcript (i.e., the word or words corresponding to what the user uttered), referred to herein as the choice list (also known in the art as an N-best list), from which the user may choose the correct version. One factor contributing to incorrect recognitions that is particularly relevant in the following description is variations in user pronunciation. A user with a certain dialect or accent may utter a word that does not score well with the phonetic representation of that word stored in the lexicon of the speech recognizer.

SUMMARY

[0006] The described embodiment generates an alternative phonetic representation (i.e., alternative pronunciation) of an initial pronunciation of a word (or phrase). In general, the initial pronunciation of the word is not the highest-scoring word provided by the speech recognizer, but is rather a word chosen by the user from an N-best list of alternatives or entered manually. The alternative phonetic representation is then stored as either a replacement for, or in addition to, the existing phonetic representation in the phonetic lexicon.

[0007] In the described embodiment, a speech recognizer processes an utterance from a user and generates a recognized transcript, along with an N-best list of alternatives. For an initial transcript, the user chooses one of the alternatives to the recognized transcript, or enters an alternative transcript manually (if the correct transcript is not available from the speech recognizer). The speech recognizer is constrained to recognize a hypothesis that differs from the initial transcript by no more than one phoneme. The score of this hypothesis thus represents the best scoring alternate pronunciation with respect to the utterance that is different from the initial pronunciation by at most one phoneme. If the score of this alternate pronunciation is higher (by some threshold) than that of the initial pronunciation by some threshold, the speech recognizer updates its lexicon by replacing the initial pronunciation currently in the lexicon with the alternate pronunciation. Alternatively, instead of replacing the pronunciation, the speech recognizer may add the new pronunciation, so that both pronunciations are in the lexicon.

[0008] If the score of the new pronunciation is not higher (by some threshold) than the score of the initial pronunciation by more than some threshold, the speech recognizer does not update its lexicon.

[0009] In one aspect, a method of generating an alternative pronunciation for a word or phrase given an initial pronunciation and a spoken example of the word or phrase includes providing the initial pronunciation of the word or phrase, generating the alternative pronunciation by searching a neighborhood of pronunciations about the initial pronunciation, and selecting a highest scoring pronunciation within the neighborhood of pronunciations. The neighborhood may include pronunciations that differ from the neighborhood by some limited number or amount of speech sub-units, such as phonemes, syllables, diphones, triphones, or other such subunits of speech known in the art.

[0010] The method includes searching the neighborhood of pronunciations that differ from the initial pronunciation by at most one phoneme, for example by using a speech recognition system to perform phoneme recognition with a constraint.

[0011] The method further includes using a phonetic recognizer to associate a score with each of the initial and/or the alternative pronunciations, and using one or both of these scores to decide whether to add the new pronunciation to the lexicon.

[0012] The method includes updating the associated lexicon by replacing the initial pronunciation in the lexicon with the highest-scoring alternative pronunciation, or by augmenting the lexicon by adding the alternative pronunciation. The user may have an option of allowing or disallowing the update of the lexicon.

[0013] In another aspect, a method of generating an alternative pronunciation of an initial pronunciation includes generating an initial pronunciation corresponding to a spoken utterance, generating one or more potential alternative pronunciations by changing the initial pronunciation by one phoneme, and selecting a highest scoring potential alternative pronunciation with respect to the spoken utterance as the alternative pronunciation of the initial pronunciation.

[0014] In another aspect, a computer readable medium with stored instructions adapted for generating an alternative pronunciation of an initial pronunciation includes instructions for generating an initial pronunciation corresponding to a spoken utterance. The medium further includes instructions for generating one or more potential alternative pronunciations by changing the initial pronunciation by one phoneme, and instructions for selecting a highest scoring potential alternative pronunciation with respect to the spoken utterance as the alternative pronunciation of the initial pronunciation.

[0015] In another aspect, a method of updating a lexicon used by a speech recognizer includes selecting a phonetic representation of a spoken utterance, generating a set of alternate phonetic representations by changing one or more phonemes in the phonetic representation, and scoring the set of alternate phonetic representations as to how well each one matches the spoken utterance, so as to produce a highest-scoring phonetic representation. The method further includes updating the lexicon with the highest scoring phonetic representation.

[0016] In another aspect, a method of generating an alternative pronunciation for a word or phrase given an initial pronunciation and a spoken example of the word or phrase includes providing the initial pronunciation of the word or phrase. The method further includes generating the alternative pronunciation by searching a neighborhood of pronunciations about the initial pronunciation via a constrained search. The neighborhood includes pronunciations that differ from the initial pronunciation by at most one phoneme. The method also includes selecting a highest scoring pronunciation within the neighborhood of pronunciations.

[0017] In another aspect, a method of generating an alternative pronunciation of an initial pronunciation includes generating an initial pronunciation corresponding to a spoken utterance, generating one or more potential alternative pronunciations by constructing one or more hypotheses constrained so as to match the initial pronunciation except for phoneme, and selecting a highest scoring potential alternative pronunciation with respect to the spoken utterance as the alternative pronunciation of the initial pronunciation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] FIG. **1** shows a constraint (finite-state machine) used in phoneme recognition to find the best-scoring pronunciation that differs from the original pronunciation by at most one phoneme.

[0019] FIGS. **2***a* and **2***b* show, in flow diagram form, the operation of the described embodiment.

[0020] FIG. **3** shows a high-level block diagram of a smartphone.

DETAILED DESCRIPTION

[0021] The described embodiment is a cell phone with embedded speech recognition functionality that allows a user to bypass the manual keypad and enter commands and data via spoken words. Embedded application software in the cellular telephone provides the speech recognition functionality (also referred to a speech recognizer). The speech recognizer includes a process for updating its phonetic lexicon to better match a user's pronunciation.

[0022] When the user utters a word or phrase, the speech recognizer searches a lexicon of phonetic representations for the highest scoring match of the acoustic utterance, and provides a recognized transcript corresponding to that highest scoring phonetic representation. The speech recognizer also provides the user with a list of alternatives to the recognized transcript (i.e., the N-best list). The N-best list corresponds to the next N highest scoring phonetic representations (with respect to the utterance) in the lexicon.

[0023] If the user selects an alternative from the N-best list instead of the recognized transcript, or if the user manually enters an alternative because the correct choice is not available in the recognized transcript or the N-best list, the speech recognizer may update its phonetic lexicon with an alternative pronunciation that is within a neighborhood of the alternative transcript (referred to herein as the "initial transcript") chosen by the user.

[0024] The speech recognizer searches the space of all pronunciations that differ from the initial pronunciation by no more than one phoneme. If the score of the pronunciation output by the speech recognizer is greater than the score of the initial pronunciation (by a predetermined threshold), the speech recognizer updates the lexicon with the new pronunciation. The particular value of the threshold is selected to result in desired performance without changing the lexicon for insignificant variations of pronunciation. The threshold thus allows for filtering small pronunciation changes that do not provide a beneficial impact. Updating the lexicon includes replacing the initial pronunciation. Updating the lexicon may alternatively include augmenting the lexicon with the new pronunciation, without removing or otherwise replacing the initial pronunciation.

[0025] FIGS. **1**, **2***a*, and **2***b* show flow diagrams describing how the described embodiment updates its lexicon as generally set forth above. We then present a description of a typical cell phone system in which the general functionality can be implemented.

[0026] In the most general sense, each of the embodiments described herein takes an utterance, i.e., a spoken example of a word or phrase, along with an initial pronunciation of that utterance (e.g., a pronunciation corresponding to a recognized transcript or an alternative to that transcript, or some other source of a pronunciation), and generates an alternative pronunciation that is within a "neighborhood" of the initial pronunciation. In the described embodiment, this neighborhood is defined by a variation in the phonemes of the initial pronunciation (e.g., one phoneme different), but in general the neighborhood could be defined by any variation of the initial pronunciation that changes how well the changed pronunciation matches the utterance. Any pronunciation subunit, e.g., syllables, diphones, triphones, etc., as an alternative to phonemes, may be used to define these variations. Further, the neighborhood could be defined by a combination of such variations. Also in this embodiment the initial pronunciation comes from a cell phone user's choice of an alternative recognized transcript, but in general the initial pronunciation could come from other sources. The concepts described herein merely require an initial pronunciation and a corresponding spoken example of that pronunciation. For a cell phone with a phonetic lexicon, all that is required is a spoken example of a word or phrase and a spelling of that word or phrase that can be used to find a pronunciation in the lexicon.

[0027] FIG. **1** shows the constraint (finite-state machine) used in the phoneme recognition including a first row **102** of states with the states constrained to phonemes p_1 through p_7 as shown, and an initial silence state s_1 and a final silence state s_2. The phonemes p_1 through p_7 represent the initial pronunciation described above. Below the first row of states **102** is a second row of states **104**, which is essentially a duplicate of the initial pronunciation states in the first row **102** starting with the second phoneme. Between the first row **102** and the second row **104** are a number of "any phoneme" states (A) that can take on any particular phoneme identity. Potential transition paths are shown with arrowed lines. The first row thus represents the sequence of phonemes in the initial pronunciation with no changes, and the second row **104** represents the sequence of phonemes with one phoneme different. In the second row **104**, where a node has more than one input, the recognizer chooses the highest scoring input, i.e., the path that best matches the spoken utterance. Possible hypothesis paths into the n^{th} node of the second row **104** include (i) the

$(n-1)^{th}$ state of the second row **104**, (ii) the $(n-1)^{th}$ "any phoneme" state, so that a different phoneme replaces the $(n-1)^{th}$ phoneme the initial pronunciation, (iii) the $(n-2)^{th}$ phoneme of the initial pronunciation, effectively deleting the previous phoneme, or (iv) the n^{th} "any phoneme" state, thereby inserting an additional phoneme into the hypothesis.

[0028] With this architecture, regardless of the path taken from the initial silence s_1 to the end of the second row **104**, the recognized hypothesis will include at most one phoneme change (substitution, insertion, or deletion), and will represent the highest scoring hypothesis with at most one phoneme different. The score at s_2 therefore corresponds to the best scoring pronunciation with at most one phoneme different from the initial pronunciation, which is used as the alterative pronunciation. States p_7 and s_2 are shown in broken lines, because they have no input to the second row **104** result. In the preferred embodiment, insertions are excluded at the beginning and end of the utterance.

[0029] The process for updating the speech recognizer lexicon in the described embodiment is shown in FIGS. **2***a* and **2***b*. The process begins when the user utters a word or phrase **120** (i.e., an utterance). The speech recognizer evaluates **122** its phonetic lexicon of standard pronunciations with respect to the utterance using a phonetic recognizer, and selects **124** the highest-scoring member. The speech recognizer presents **126** the highest scoring member to the user as the recognized transcript, and also presents **127** the next N highest scoring members as an N-best list of alternatives to the recognized transcript.

[0030] The user typically selects either (i) the recognized transcript **128** or (ii) one of the members of the N-best list **130** of alternatives, as what he actually uttered. However in some cases, neither the recognized transcript nor the N-best list includes **131** what the user actually uttered. In those cases, the user may either enter the word/phrase manually **132**, effectively bypassing the speech recognition functionality, or simply utter **134** the word or phrase again.

[0031] If the user selects the recognized transcript **128**, the speech processor does not update its lexicon, and waits for the next utterance. If the user selects an alternative from the N-best list **130** or manually enters the word/phrase, the speech recognizer generates **100** an alternative pronunciation from the initial pronunciation as described above.

[0032] The speech recognizer compares the score of the user's alternative (i.e., the initial pronunciation) to the score of the alternate pronunciation. If **140** the score of the alternate pronunciation is greater than the score of the initial pronunciation by a threshold, the speech recognizer replaces **142** the phonetic representation of the initial pronunciation in the lexicon with the alternative pronunciation generated **100** by the speech recognizer

[0033] Updating the lexicon to replace the initial pronunciation as described above removes that initial phonetic representation from future consideration by the speech processor. Other users of the cell phone, however, may pronounce words in such a way that would produce a better score on the original phonetic representation that was replaced than on the updated phonetic representation. Therefore another way to update the lexicon in the above-described procedure is to add the highest scoring phonetic representation to the lexicon without eliminating the original pronunciation, so that both pronunciations are included in the lexicon for future consideration by the speech processor.

[0034] In either case of updating the lexicon (i.e., by replacement or augmentation), the cell phone may provide the user with the option of whether or not to allow update. This option may be on a case-by-case basis, so that each time a potential update is available, the user may affirmatively allow or disallow the update via a keystroke or spoken command. This option can also be selected as an enable/disable function, so that the all updates are allowed when the user enables the function, and all updates are disallowed when the user disables the function.

[0035] The speech recognizer may be able to further improve the pronunciation through an iterative process. For example, if the score of the alternative pronunciation is better than the initial pronunciation by a predetermined threshold, the speech recognizer generates yet another pronunciation by taking the previously determined alternative pronunciation and finding a new, higher-scoring alternative pronunciation that differs from the previously determined alternative pronunciation by only one phoneme. This iterative process continues until the improvement drops below the predetermined threshold, indicating that the improvement is leveling off

[0036] A smartphone **200**, as shown in FIG. **3**, is a typical platform that can provide such speech recognition functionality via embedded application software. In fact, the described method of updating the phonetic lexicon may also be implemented in other portable phones, and in other hand held devices in general.

[0037] Smartphone **200** is a Microsoft PocketPC-powered phone which includes at its core a baseband DSP **202** (digital signal processor) for handling the cellular communication functions (including for example voiceband and channel coding functions) and an applications processor **204** (e.g. Intel StrongArm SA-110) on which the PocketPC operating system runs. The phone supports GSM voice calls, SMS (Short Messaging Service) text messaging, wireless email, and desktop-like web browsing along with more traditional PDA features.

[0038] An RF synthesizer **206** and an RF radio transceiver **208**, followed by a power amplifier module **210**, implement the transmit and receive functions. The power amplifier module handles the final-stage RF transmit duties through an antenna **212**. An interface ASIC **214** and an audio CODEC **216** provide interfaces to a speaker, a microphone, and other input/output devices provided in the phone such as a numeric or alphanumeric keypad (not shown) for entering commands and information.

[0039] DSP **202** uses a flash memory **218** for code store. A Li-Ion (lithium-ion) battery **220** powers the phone and a power management module **222** coupled to DSP **202** manages power consumption within the phone. SDRAM **224** and flash memory **226** provide volatile and non-volatile memory, respectively, for applications processor **214**. This arrangement of memory holds the code for the operating system, the code for customizable features such as the phone directory, and the code for any embedded applications software in the smartphone, including the voice recognition software described above. The visual display device for the smartphone includes LCD driver chip **228** that drives LCD display **230**. Clock module **232** provides the clock signals for the other devices within the phone and provides an indicator of real time. All of the above-described components are packages within an appropriately designed housing **234**.

[0040] Smartphone **200** described above represents the general internal structure of a number of different commercially available smartphones, and the internal circuit design of those phones is generally known in the art.

[0041] In the described embodiment, an application running on the applications processor **104** performs the process of updating the phonetic lexicon as described in FIGS. **1**, **2***a*, and **2***b*.

[0042] Other aspects, modifications, and embodiments are within the scope of the following claims.

What is claimed is:

1. A method comprising:

receiving a spoken utterance of a word or phrase;

generating a plurality of alternative pronunciations of the spoken utterance;

scoring one or more pronunciations of the plurality of alternative pronunciations using the spoken utterance; and

updating a lexicon with at least one scored pronunciation.

2. The method of claim **1**, wherein the at least one scored pronunciation is the highest scoring pronunciation of the scored pronunciations.

3. The method of claim **1**, further comprising using a finite-state machine to generate the plurality of alternative pronunciations.

4. The method of claim **1**, wherein updating the lexicon includes replacing an existing pronunciation with the at least one scored pronunciation.

5. The method of claim **1**, wherein updating the lexicon includes adding a phonetic representation of the at least one scored pronunciation to an existing representation.

6. The method of claim **1**, wherein the alternative pronunciations are generated by searching a neighborhood of pronunciations about an initial pronunciation of the spoken utterance.

7. A system comprising:

at least one processor; and

a memory device operatively connected to the at least one processor;

wherein, responsive to execution of program instructions accessible to the at least one processor, the at least one processor is configured to:

receive a spoken utterance of a word or phrase;

generate a plurality of alternative pronunciations of the spoken utterance;

score one or more pronunciations of the plurality of alternative pronunciations using the spoken utterance; and

update a lexicon with at least one scored pronunciation.

8. The system of claim **7**, wherein the at least one scored pronunciation is the highest scoring pronunciation of the scored pronunciations.

9. The system of claim **7**, wherein the at least one processor is configured to use a finite-state machine to generate the plurality of alternative pronunciations.

10. The system of claim **7**, wherein the at least one processor is configured to update the lexicon by replacing an existing pronunciation with the at least one scored pronunciation.

11. The system of claim **7**, wherein the at least one processor is configured to update the lexicon by adding a phonetic representation of the at least one scored pronunciation to an existing representation.

12. The system of claim **7**, wherein the at least one processor is configured to generate the alternative pronunciations by searching a neighborhood of pronunciations about an initial pronunciation of the spoken utterance.

13. A computer program product encoded in a non-transitory computer-readable medium, which when executed by a computer causes the computer to perform the following operations:

receiving a spoken utterance of a word or phrase;

generating a plurality of alternative pronunciations of the spoken utterance;

scoring one or more pronunciations of the plurality of alternative pronunciations using the spoken utterance; and

updating a lexicon with at least one scored pronunciation.

14. The computer program product of claim **13**, wherein the at least one scored pronunciation is the highest scoring pronunciation of the scored pronunciations.

15. The computer program product of claim **13**, wherein the computer uses a finite-state machine to generate the plurality of alternative pronunciations.

16. The computer program product of claim **13**, wherein the computer updates the lexicon by replacing an existing pronunciation with the at least one scored pronunciation.

17. The computer program product of claim **13**, wherein the computer updates the lexicon by adding a phonetic representation of the at least one scored pronunciation to an existing representation.

18. The computer program product of claim **13**, wherein the computer generates the alternative pronunciations by searching a neighborhood of pronunciations about an initial pronunciation of the spoken utterance.

* * * * *