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

**(19) World Intellectual Property Organization**
International Bureau

**(43) International Publication Date**
17 June 2010 (17.06.2010)

**PCT**

**(10) International Publication Number**
**WO 2010/067118 Al**

**(51) International Patent Classification:**
***GlOL 15/28*** (2006.01) ***GlOL 15/26*** (2006.01)

**(21) International Application Number:**
PCT/GB2009/05 1684

**(22) International Filing Date:**
10 December 2009 (10.12.2009)

**(25) Filing Language:** English

**(26) Publication Language:** English

**(30) Priority Data:**
61/121,903 11 December 2008 (11.12.2008) US

**(71) Applicant** *(for all designated States except US):* **NOVAURIS TECHNOLOGIES LIMITED** [GB/GB]; Millbank, Stoke Road, Bishops Cleeve, Cheltenham, Gloucestershire GL52 8RW (GB).

**(72) Inventors; and**
**(75) Inventors/Applicants** *(for US only):* **HUNT, Melvyn** [GB/GB]; 8 Dewey Close, Woodmancote, Cheltenham, Gloucestershire GL52 9UF (GB). **BRIDLE, John** [GB/GB]; Millbank, Stoke Road, Bishops Cleeve, Cheltenham, Gloucestershire GL52 8RW (GB).

**(74) Agents:** NEWELL, William, **Joseph** et al; Wynne-jones, Laine & James LLP, Essex Place, 22 Rodney Road, Cheltenham, Gloucestershire GL50 **U J** (GB).

**(81) Designated States** *(unless otherwise indicated, for every kind cf national protection available):* AE, AG, AL, AM, **AO, AT, AU, AZ, BA,** BB, BG, **BH,** BR, **BW, BY, BZ, CA, CH, CL, CN, CO,** CR, **CU, CZ,** DE, **DK, DM, DO, DZ,** EC, EE, EG, **ES, FI,** GB, GD, GE, **GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.**

**(84) Designated States** *(unless otherwise indicated, for every kind cf regional protection available):* ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, **IT,** LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *with international search report (Art. 21(3))*

**(54) Title:** SPEECH RECOGNITION INVOLVING A MOBILE DEVICE

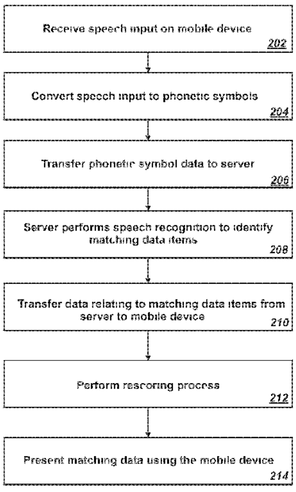

Receive speech input on mobile device
202
Convert speech input to phonetic symbols
204
Transfer phonetic symbol data to server
206
Server performs speech recognition to identify matching data items
208
Transfer data relating to matching data items from server to mobile device
210
Perform rescoring process
212
Present matching data using the mobile device
214

*Fig. 2*

**(57) Abstract:** A system and method of speech recognition involving a mobile device. Speech input is received (202) on a mobile device (102) and converted (204) to a set of phonetic symbols. Data relating to the phonetic symbols is transferred (206) from the mobile device over a communications network (104) to a remote processing device (106) where it is used (208) to identify at least one matching data item from a set of data items (114). Data relating to the at least one matching data item is transferred (210) from the remote processing device to the mobile device and presented (214) thereon.

# Speech Recognition involving a Mobile Device

The present invention relates to speech recognition involving a mobile device.

Speech recognition involving mobile devices, such as mobile telephones, is conventionally carried out either entirely locally on the processor fitted in the device/phone itself or by sending the speech waveform over the network to a server. Servers have to be used when the processor (if any) in the phone has inadequate power or inadequate memory resources to carry out the speech recognition satisfactorily. Servers also have to be used when the information needed for the speech recognition must be held centrally because it is too sensitive to distribute to the phone subscribers (e.g. a list of all account holders at a bank), is too large or needs to be updated very frequently. Server-based arrangements incur costs, normally borne by the subscriber or by the network service provider, for the transmission of the encoded speech waveforms over the network. Also, especially when the network connection is poor, undesirable delays (latencies) are introduced into the speech recognition response. Transmission of the speech waveform over the network also normally requires the signal to be restricted to the telephone bandwidth (roughly 300Hz to 3.3 kHz) or at least the region below 4 KHz, and the necessary encoding of the waveform inevitably introduces distortions.

Some work has been carried out on "distributed speech recognition", in which the power spectrum parameters, typically mel-frequency cepstrum components (MFCCs), are computed locally and transmitted to the server where recognition is carried out. The sets of MFCCs, which are usually called "frames", are computed at a rate of around 100 per second. The main motivation for this arrangement compared with the conventional server-based arrangement is the avoidance of coding distortions in transmitting the speech waveform, as well as the possibility of analyzing a wider bandwidth than that normally used for telephone communications. There has been an attempt to define an international standard for distributed speech recognition in the Aurora project of the ETSI standards organisation (see http://portal.etsi.org/stq/kta/dsr/dsr.asp, for instance). However, apart from some use in Japan, the technique has not been widely used in practical situations.

Embodiments of the present invention provide an alternative approach to distributed speech recognition and have advantages over the conventional methods described above.

Embodiments of the present inventors' approach are novel and unusual in that the speech to be recognized is first converted to a (necessarily errorful) sequence of phonetic symbols. This sequence is significantly more compact than the encoded speech waveform or the MFCCs. Thus, transmitting the phonetic sequence from a mobile device over the network to a server is much faster and less expensive. The phonetic sequence can then be used on the server to search through very large lists of items and one or more good matches can be returned to the mobile device. One example of a complete speech recognition process that uses an initial conversion to a sequence or lattice of phonetic symbols is described in US patent number 714631 9 ("Phonetically Based Speech Recognition System and Method"), the contents of which are incorporated herein by reference, although the techniques described in this specification may be implemented using other speech recognition processes. In some embodiments, a detailed representation of the speech is retained on the mobile device in order to allow further refinement to be carried out locally to make a more accurate decision on the best matching items.

Embodiments of the system/process have the advantage, shared by other methods involving servers, that the full database being searched and the pronouncing dictionary never has to be downloaded to the mobile device/phone, but, as noted, it avoids heavy data transmission with the consequential undesirable latency. There is a particular advantage with small form factor devices and smart phones, such as the Apple iPhone™, which have a reasonably powerful processor, but make it difficult to incorporate into local speech recognition applications large databases (e.g. the set of all streets in the US that is needed to have a spoken address be recognized for an in-car navigation aid) and thus make it difficult to carry out the complete recognition process locally.

According to a first aspect of the present invention there is provided a method of speech recognition involving a mobile device, the method including:

receiving speech input on a mobile device;

converting the speech input to a set of phonetic symbols on the mobile device;

transferring data relating to the phonetic symbols from the mobile device over a communications network to a remote processing device;

using the data relating to the phonetic symbols on the remote processing device to identify at least one matching data item from a set of data items;

transferring data relating to the at least one matching data item from the remote processing device to the mobile device, and

presenting data relating to the at least one transferred matching data item using the mobile device.

The identifying of at least one matching data item may include matching the data relating to the phonetic symbols against a set of phonetic reference forms corresponding to the set of data items.

The set of phonetic symbols may comprise at least one sequence (or lattice) of phonetic symbols. The lattice can comprise a directed acyclic graph used as a compact representation of a set of similar sequences.

A number of the matching data item(s) transferred from the remote processing device to the mobile device can correspond to a lower of: a maximum number of data items to be presented on the mobile device, or a number of data items arranged in order of decreasing posterior probability of corresponding to the phonetic symbols down to a predetermined threshold value. The posterior probability for a said data item can be computed by taking a match score of the phonetic symbols against the phonetic reference form of the data item and normalizing the match score.

The step of presenting data relating to the at least one transferred matching data item can comprise displaying an orthographic representation of the at least one transferred matching data item. Alternatively, the data presented may correspond to a map coordinate of a particular location represented by the matching data item. Alternatively, the data presented may correspond to an identification code for a piece of music or media represented by the transferred matching data item. The step of presenting data relating to the at least one transferred matching data item can comprise using a speech synthesis technique executing on the mobile device to generate a spoken form of the at least one transferred matching data item.

The method may further include the mobile device storing data representing the speech input (e.g. as a speech waveform or a sequence of frames of MFCCs or other representation of a short-term acoustic power spectrum corresponding to the speech input) and the mobile device performing a rescohng process using the at least one transferred matching data item and the stored speech input data. The rescohng process can include (for each said matching data item to be rescored) generating sequences or lattices of acoustic hidden Markov models (HMMs) of phonetic units corresponding to a phonetic specification of reference pronunciations corresponding to a said transferred matching data item. Each of the sequences or lattices can be matched against a sequence of frames of spectrum parameters corresponding to the speech input data stored in the mobile device to produce a match score. The matching of the sequences or lattices may involve Viterbi time alignment or a full forward probability method. The rescoring process may involve producing a network including data representing phonetic specifications of all of the data items in the set.

The method may further include the mobile device sharing common elements in the phonetic specification in order to reduce an amount of data transferred to the mobile device for the rescoring process. The method may further include transferring data specifying a general-purpose sub-grammar representing a set of alternative number-word sequences to the mobile device (rather than transferring data fully describing the alternative number-word sequences) for the rescoring process, and the mobile device may use the sub-grammar to determine a most likely one of the number-words sequences to be presented.

The rescoring process may include:

transferring phonetic specifications corresponding to the matching data items from the remote processing device the to the mobile device with corresponding indices instead of a full description of the matching data items;

the mobile device selecting which of the phonetic specification(s) is/are to be presented;

transferring the index or indices corresponding to the selected phonetic specification(s) from the mobile device to the remote processing device;

the remote processing device transferring data representing a full description(s) of the data item(s) corresponding to the index or indices transferred by the mobile device back to the mobile device, and

the mobile device presenting the data representing the full description(s).

The method may include transferring from the remote processing device the matching data items corresponding to a predetermined number of best matches of the phonetic symbols and the data items in the set.

According to a further aspect of the present invention there is provided a computer program product configured to execute methods substantially as described herein.

According to another aspect of the present invention there is provided a system adapted to perform speech recognition involving a mobile device, the system including:

a mobile device configured to receive speech input;

the mobile device including a device configured to convert the speech input to a set of phonetic symbols on the mobile device;

the mobile device further configured to transfer data relating to the phonetic symbols to a remote processing device over a communications network;

the remote processing device including a device configured to use the data relating to the phonetic symbols to identify at least one matching data item from a set of data items;

the remote device including a communications device configured to transfer data relating to the at least one matching data item to the mobile device over the communications network, and

the mobile device including a device configured to present data relating to the at least one transferred matching data item.

The mobile device may comprise a mobile telephone. The communications network may include a GSM, CDMA or UMTS network.

According to another aspect of the present invention there is provided a mobile device adapted to perform speech recognition including:

a device configured to receive speech input;

a device configured to convert the speech input to a set of phonetic symbols on the mobile device;

a device configured to transfer data relating to the phonetic symbols to a processing device over a communications network.

According to another aspect of the present invention there is provided a processing device adapted to perform speech recognition involving a mobile device, the processing device including:

a device configured to receive data relating to a set of phonetic symbols from a mobile device over a communications network;

a device configured to use the data relating to the phonetic symbols to identify at least one matching data item from a set of data items;

a communications device configured to transfer data relating to the at least one matching data item to the mobile device over the communications network.

According to yet another aspect of the present invention there is provided a mobile device configured to execute a method substantially as described herein.

According to further aspect of the present invention there is provided a processing device adapted to perform speech recognition involving a mobile device, the processing device configured to execute a method substantially as described herein.

According to yet another aspect of the present invention there is provided a speech recognition system in which the speech to be recognized is spoken by the user into a mobile device/phone where it is reduced to a sequence or lattice of phonetic symbols, which are then transmitted over a data communications network to a server which matches the phonetic information against a set of phonetic reference forms corresponding to a set of items to be recognized and identifying information suitable for presenting to the user for one or more of the best matching items is transmitted back to the mobile phone and presented to the user.

Information on a multiplicity of best matching items can be returned to the mobile device and may be subjected to a refinement process using a stored representation of the original input speech. The information on the multiplicity of best matching items can include phonetic specifications of the one or more expected pronunciations of each of the items. The refinement may comprise matching a sequence of acoustic parameter frames derived from the input speech against a sequence or lattice of acoustic models representing phonetic units corresponding to the possible pronunciations of the items being matched. The acoustic models may comprise hidden Markov models.

The identifying information may comprise a textual form of the spoken input, or some close equivalent. The identifying information may comprise data needed by a speech synthesis system located in the mobile telephone to generate speech allowing the user to know which item has or multiplicity of items have been recognized. Identifying indices corresponding to the item or plurality of items found to match best by the refinement process can be transmitted back to the server, which then returns data suitable to inform the user as to the identity of the items.

While the invention has been described above, it extends to any inventive combination of features set out above or in the following description. Although illustrative embodiments of the invention are described in detail herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to these precise embodiments. As such, many modifications and variations will be apparent to practitioners skilled in the art. Furthermore, it is contemplated that a particular feature described either individually or as part of an embodiment can be combined with other individually described features, or parts of other embodiments, even if the other features and embodiments make no mention of the particular feature. Thus, the invention extends to such specific combinations not already described.

The invention may be performed in various ways, and, by way of example only, embodiments thereof will now be described, reference being made to the accompanying drawings in which:

Figure 1 illustrates schematically an example system where a mobile device communicates with a server;

Figure 2 illustrates schematically example steps performed by the mobile device and the server, including a re-scoring process, and

Figure 3 illustrates schematically example steps performed during the re-scoring process.

Figure 1 shows an example system where a mobile device 102 can communicate via a network 104 with a server 106 in order to perform a speech recognition process. The mobile device 102 may be a mobile telephone, such as an Apple iPhone™, or another type of device, such as a Personal Digital Assistant or a portable computer with audio input/output (e.g. microphone 102A and speaker 102B) and wireless communications capabilities. The mobile device can further include a processor 103A and memory 103B for executing applications, which can involve speech recognition, as well as a display 103C.

The network 104 may comprise a Global System for Mobile (GSM) communications network, but use of other types of networks is possible. The server 106 can comprise a general purpose computer configured to communicate with other devices over the network using an interface 107 and further comprises a processor 108 and memory 110. The memory 110 of the server 106 includes a speech recognition application 112, which may be based on a speech recognition process such as that described in US patent number 714631 9, for example. The memory further includes a database 114, which can be used/searched by the speech recognition application to find data item(s) that best matches user input. The database can comprise a set of data items representing any type of information, such as a telephone number directory, music track information, etc. In an alternative embodiment the database 114 can stored be on another remote device that is in communication with the server 106.

Referring to Figure 2, steps occurring during example use of the components of Figure 1 are shown. At step 202 speech input from a user is received by the mobile device 102. This can, for instance, involve a telephone number directory application executing on the processor 103B of the mobile device prompting the user to say the name and/or other detail, e.g. partial address, of a person/entity whose telephone number is desired into the microphone 102A. It can also involve known techniques such as analogue to digital conversion or spectrum analysis, which will be familiar to the skilled person.

At step 204 the speech input is converted into data representing a set of phonetic symbols. The set of phonetic symbols can comprise one or more sequences of phonetic symbols that are considered to comprise close contenders for matching the speech input. Here, the term "set" is to be interpreted broadly and is not limited to data having any particular constrains in terms of order, uniqueness, etc. Because they all describe the same speech input, the sequences can be efficiently represented as a directed acyclic graph, or lattice. In some cases, a single sequence is sufficient. Here, a "phonetic symbol" can correspond to what phoneticians call a "segment", which itself normally corresponds to a single phoneme, though the use is not excluded of phonetic units somewhat smaller than a phoneme, such as the closure, release and aspiration components of a voiceless plosive, or units somewhat larger than a phoneme, such as /sk/. The phonetic symbols can be considered to comprise a computer-generated approximate phonetic transcription of the speech input. The conversion of audio data to phonetic symbols can be performed by a phonetic decoder algorithm that is called by, or is part of, the telephone directory application executing on the mobile device. The publicly available toolkit, HTK, can, for example, be used to construct a suitable phonetic decoder, and to build the models that it needs from a corpus of training speech. The decoder may be based on Bellman's Dynamic Programming. The HTK toolkit can currently be obtained from http://htk.eng.cam.ac.uk, which also provides access to "The HTK Book", by S. J. Young et al.

At step 206 data relating to the phonetic symbols is transmitted from the mobile device 102 via the network 104 to the server 106. The data transmitted may comprise a direct representation of the sequence(s), or can comprise a version of the sequence(s) that has been compressed, encrypted and/or coded/processed/re-ordered in some other way. Transmitting the phonetic symbol data over the network is much faster and less expensive than transmitting an encoded speech waveform or MFCCs, as with conventional server-based approaches. Upon receipt, the server may process the transmitted data to perform any necessary de-compression, de-encryption, etc, operations.

At step 208 the phonetic symbol data is used by the server speech recognition application 112 to search at least one very large list of data items in the database 114 using known techniques, including cheap symbol-matching operations, such as described in US patent no. 7146319 and US patent no. 7403941. Thus, the server process can determine a list of one or more items that best match the input spoken by the user of the mobile device 102. The length of the list can be limited to what can be presented to the user on the mobile device. Data relating to one or more good matches in the list found by the search can then be transmitted over the network 104 to the mobile device at step 210.

The length of the list returned to the mobile device in some embodiments can be the shorter of: the maximum length of list that can be presented on the mobile device, and the list with items in order of decreasing posterior probability of corresponding to the input down to some threshold minimum value (for example, 5%). The posterior probability for a particular item would typically be computed by taking the match score of the phonetic representation of the input speech against the reference representation of that item and "normalizing" that score by dividing it by the sum of a large number of well matching scores, where an appropriate large number is 100.

The information to be presented to the user (at step 214) can comprise an orthographic representation of the items in the returned list, but in some cases the information may be delivered in some other form, such as map coordinates of a particular location, the identification code for a piece of music or media, or information that causes a text-to-speech synthesis system to generate spoken forms of the item(s) to be presented to the user.

In a more complex embodiment that is intended to produce more accurate results, a list of information corresponding to a set of well-matching items is returned to the mobile device, with the mobile phone processor then performing a refinement process known as "rescoring" (optional step 212 of Figure 2). The information for each item consists of one or a plurality of sequences or lattices specifying the reference pronunciations of that item and a tag identifying the item. By retaining a detailed representation of the speech in the memory 103B of the mobile device (e.g. a speech waveform or a sequence of frames of MFCCs or other representation of the short-term acoustic power spectrum), this optional further refinement can be carried out locally to make a more accurate decision on the best matching items. The refinement process requires an increase in the amount of data transferred from the server to the mobile device, but this increase will still result in a smaller transfer of data than use of conventional distributed speech recognition methods. Furthermore, data transmission speeds are normally several times greater in the download (server to mobile device) direction than in the upload (mobile device to server) direction.

Referring to Figure 3, for each item to be rescored, a process executing on the processor 103A of the mobile device 102 generates (step 302) sequences or lattices of acoustic hidden Markov models (HMMs) of phonetic units corresponding to the phonetic specification of the reference pronunciations received from the server. Each of these model sequences or lattices is then matched (step 304) against the sequence of frames of spectrum parameters corresponding to the input stored in mobile device memory 103B to produce a more reliable match score. The matching process can be identical to that used in standard HMM-based speech recognition and may involve so-called Viterbi time alignment (finding the single best alignment and summing the match log probability along it), or may comprise the so-called full forward probability method, which determines the probability of the particular sequence of HMMs having given rise to the input taking all possible time alignments into account. These steps are carried out for all the items returned to the mobile device from the server 106 and the best matching item, or the best N matching items (where N is a small number greater than 1 and less than some maximum, for which a reasonable value might be 5) can be presented to the user (e.g. step 306).

An efficient way of carrying out rescohng is to compile the phonetic specifications of all items in the set to be rescored into a network. In this case, it requires less computation to obtain the top-scoring item than to obtain a set of N top-scoring items. The list to be presented to the user can then comprise the top rescored item together with the N-1 remaining top-scoring items from the phonetic matching process. There is an obvious generalization, in which the top M rescored items (M > 1 and <= N) is displayed followed by the remaining N-M top-scoring items from the phonetic matching process.

In some embodiments, the total amount of data exchanged between the server 106 and the mobile device 102 is reduced or minimized. The system recognizes that the transmission of phonetic symbols massively reduces the amount of data transmitted from the mobile phone to the server. However, in the case of rescoring, the amount of data needing to be transmitted from the server to the mobile phone can be large. Such may be the case when many items are sent back for rescoring, and especially if each one is accompanied by the associated information to be presented to the user.

One technique to reduce the amount of data to be transmitted to the mobile phone is to exploit overlap between the various items. Since the items being sent for rescoring all match the spoken input reasonably well, it is usually the case that many of them resemble each other quite closely. For example, in the case of addresses, typically many items to be rescored share the same city and state. By sharing common elements in the phonetic specification, the amount of data to be transmitted can be significantly compressed. This compression technique can also be applied to the information to be displayed.

Further, in some applications there can be positions in the phrases where it may be advantageous to use general-purpose sub-grammars. An example is house numbers in street addresses. There is likely to be considerable ambiguity remaining after symbolic scoring - a problem that is aggravated by the fact that there are many ways to say some numbers (e.g. 2021 as "two oh two one", or "twenty-two twenty-one" or "two thousand [and] twenty one" etc.). Rather than sending a possibly large set of alternative number-word sequences to the mobile device 102, it may be more efficient for the server 106 to transmit the name of an appropriate sub-grammar, and leave it to the rescoring process to determine the most likely sequence of number words, which can then be transformed to a suitable representation such as the corresponding digit sequence.

Another technique for reducing the total amount of data to be transmitted to the mobile device 102 reduces the amount of presentation data needing to be transmitted by adding a further exchange of transmissions. Instead of a full description of the item(s) to be potentially presented to the mobile device user being transmitted at each and every exchange between the mobile device and the server, the phonetic specifications are transmitted with simple indices replacing a full description of the item(s) to be presented to the user. When the set of items to be displayed has been finally determined, their indices are transmitted back to the server 106, which then sends a full description of the presentation information for only those items to the mobile device.

In practice, the top-scoring item or items determined by rescoring is usually in the top few items determined by the phonetic symbol matching process. A further alternative is therefore to transmit in the first transmission from the server the presentation information for just the top few items. The mobile device will then only request additional presentation information if the rescoring determines that items not in the top few should be presented.

With regard to embodiments described herein, each may be implemented using logic or instructions executed by processors or processing resources, with access to memory and data structures from which results may be obtained or generated. Servers and other devices described that carry out functions described may do so using processors and memory resources.

# CLAIMS

1. A method of speech recognition involving a mobile device, the method including:

receiving (202) speech input on a mobile device (102);

converting (204) the speech input to a set of phonetic symbols on the mobile device;

transferring (206) data relating to the phonetic symbols from the mobile device over a communications network (104) to a remote processing device (106);

using (208) the data relating to the phonetic symbols on the remote processing device to identify at least one matching data item from a set of data items (114);

transferring (210) data relating to the at least one matching data item from the remote processing device to the mobile device, and

presenting (214) data relating to the at least one transferred matching data item using the mobile device.

2. A method according to claim 1, wherein the step of identifying (208) at least one matching data item includes matching the data relating to the phonetic symbols against a set of phonetic reference forms corresponding to the set of data items.

3. A method according to claim 1 or 2, wherein the set of phonetic symbols comprises at least one sequence (or lattice) of phonetic symbols.

4. A method according to claim 2 or 3, wherein a number of the matching data item(s) transferred from the remote processing device (106) to the mobile device (102) corresponds to a lower of: a maximum number of said data items to be presented on the mobile device, or a number of data items arranged in order

of decreasing posterior probability of corresponding to the phonetic symbols down to a predetermined threshold value.

5. A method according to claim 4, wherein the posterior probability for a said data item is computed by taking a match score of the phonetic symbols against the phonetic reference form of the data item and normalizing the match score.

6. A method according to any one of the preceding claims, wherein the step of presenting (214) data relating to the at least one transferred matching data item comprises displaying an orthographic representation of the at least one transferred matching data item.

7. A method according to any one of the preceding claims, wherein the step of presenting (214) data relating to the at least one transferred matching data item comprises outputting data corresponding to a map coordinate of a particular location represented by a said transferred matching data item.

8. A method according to any one of the preceding claims, wherein the step of presenting (214) data relating to the at least one transferred matching data item comprises outputting data corresponding to an identification code for a piece of music or media represented by the transferred matching data item.

9. A method according to any one of the preceding claims, wherein the step of presenting (214) data relating to the at least one transferred matching data item comprises using a speech synthesis technique executing on the mobile device (102) to generate a spoken form of the at least one transferred matching data item.

10. A method according to any one of the preceding claims, further including the mobile device (102) storing data representing the speech input and the mobile device performing a rescoring process using the at least one transferred matching data item and the stored speech input data.

11. A method according to claim 10, wherein the rescohng process includes (for each said transferred matching data item to be rescored) generating (302) sequences or lattices of acoustic hidden Markov models (HMMs) of phonetic units corresponding to a phonetic specification of reference pronunciations corresponding to a said transferred matching data item.

12. A method according to claim 11, wherein each of the sequences or lattices is matched (304) against a sequence of frames of spectrum parameters corresponding to the speech input data stored in the mobile device (102) to produce a match score.

13. A method according to claim 12, wherein the matching of the sequences or lattices involves Viterbi time alignment or a full forward probability method.

14. A method according to any one of claims 10 to 13, wherein the rescohng process involves producing a network including data representing phonetic specifications of all of the data items in the set (114).

15. A method according to any one of claims 11 to 13, wherein the rescoring process includes the mobile device (102) sharing common elements in the phonetic specification in order to reduce an amount of data transferred to the mobile device for the rescoring process.

16. A method according to any one of claims 11 to 13, wherein the rescoring process includes:

transferring data specifying a general-purpose sub-grammar representing a set of alternative number-word sequences to the mobile device (102) for the rescoring process, and

the mobile device using the sub-grammar to determine a most likely one of the number-words sequences to be presented (214).

17. A method according to any one of claims 11 to 13, wherein the rescoring process includes:

transferring phonetic specifications corresponding to the matching data items from the remote processing device (106) the to the mobile device (102) with corresponding indices instead of a full description of the matching data items;

the mobile device selecting which of the phonetic specification(s) is/are to be presented;

transferring the index or indices corresponding to the selected phonetic specification(s) from the mobile device to the remote processing device;

the remote processing device transferring data representing a full description(s) of the data item(s) corresponding to the index or indices transferred by the mobile device back to the mobile device, and

the mobile device presenting the data representing the full deschption(s).

18. A method according to any one of claims 11 to 13, wherein the rescohng process includes:

transferring from the remote processing device (106) the matching data items corresponding to a predetermined number of best matches of the phonetic symbols and the data items in the set.

19. A computer program product configured to execute a method according to any one of the preceding claims.

20. A system adapted to perform speech recognition involving a mobile device, the system including:

a mobile device (102) configured to receive (102A) speech input;

the mobile device including a device (103A) configured to convert the speech input to a set of phonetic symbols on the mobile device;

the mobile device further configured to transfer data relating to the phonetic symbols to a remote processing device (106) over a communications network (104);

the remote processing device including a device (108, 110) configured to use the data relating to the phonetic symbols to identify at least one matching data item from a set of data items (114);

the remote device including a communications device (107) configured to transfer data relating to the at least one matching data item to the mobile device over the communications network, and

the mobile device including a device (102B, 103C) configured to present data relating to the at least one transferred matching data item.

21. A system according to claim 20, wherein the mobile device comprises a mobile telephone (102).

22. A system according to claim 20 or 21, wherein the communications network comprises a GSM, CDMA or UMTS network (104).

23. A mobile device (102) adapted to perform speech recognition including:

a device (102A) configured to receive speech input;

a device (103A) configured to convert the speech input to a set of phonetic symbols on the mobile device;

a device (103A) configured to transfer data relating to the phonetic symbols to a remote processing device (106) over a communications network (104).

24. A processing device (106) adapted to perform speech recognition involving a mobile device, the processing device including:

a device (107) configured to receive data relating to a set of phonetic symbols from a mobile device (102) over a communications network (104);

a device (108) configured to use the data relating to the phonetic symbols to identify at least one matching data item from a set of data items (114);

a Communications device (107) configured to transfer data relating to the at least one matching data item to the mobile device over the communications network.

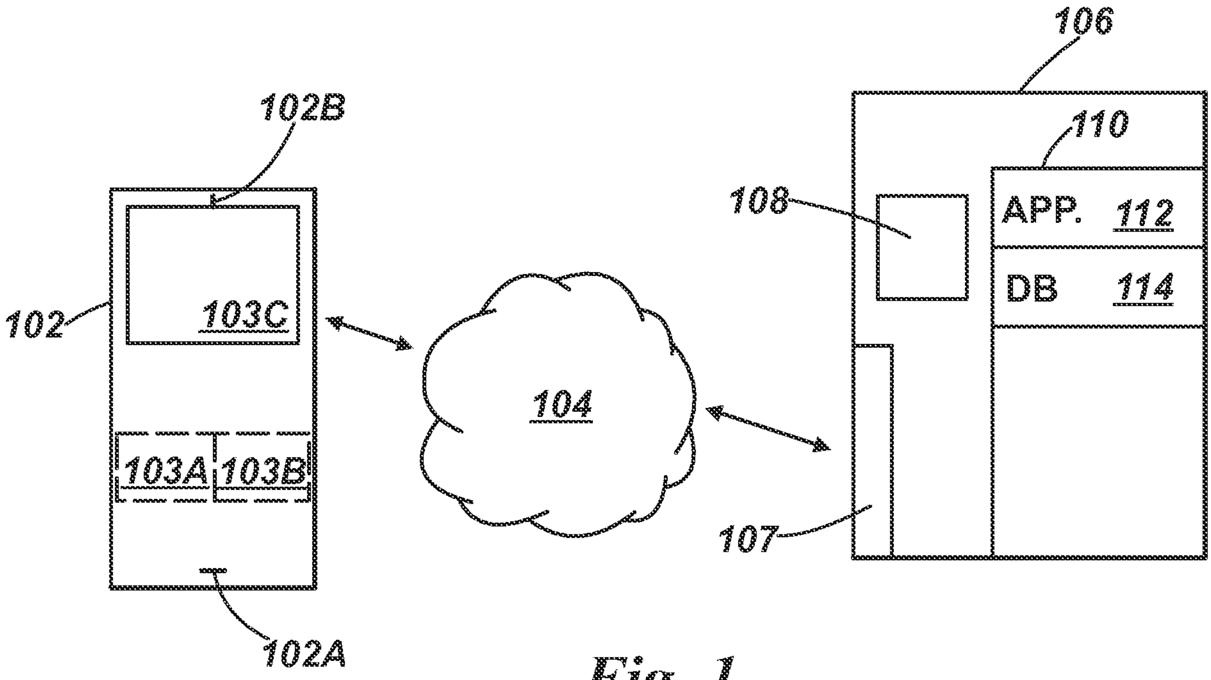
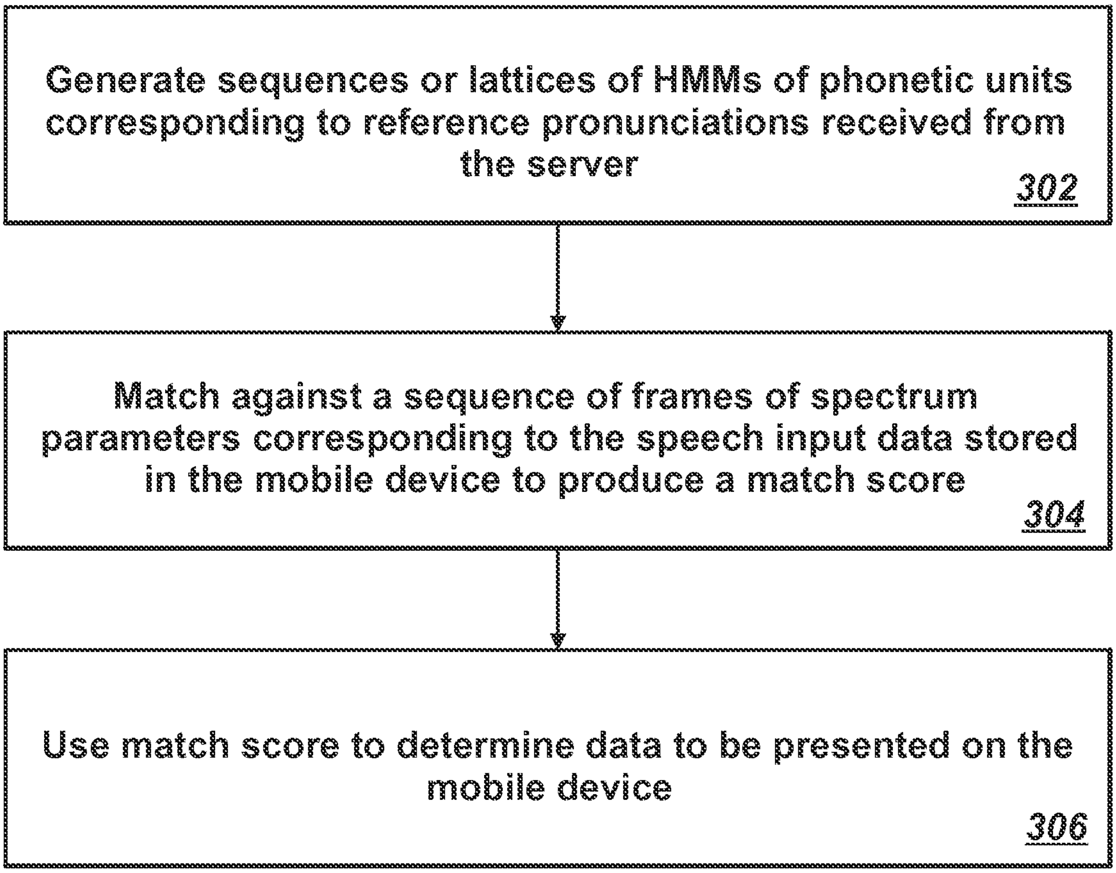


Fig. 1


Generate sequences or lattices of HMMs of phonetic units corresponding to reference pronunciations received from the server 302

Match against a sequence of frames of spectrum parameters corresponding to the speech input data stored in the mobile device to produce a match score 304

Use match score to determine data to be presented on the mobile device 306

Fig. 3

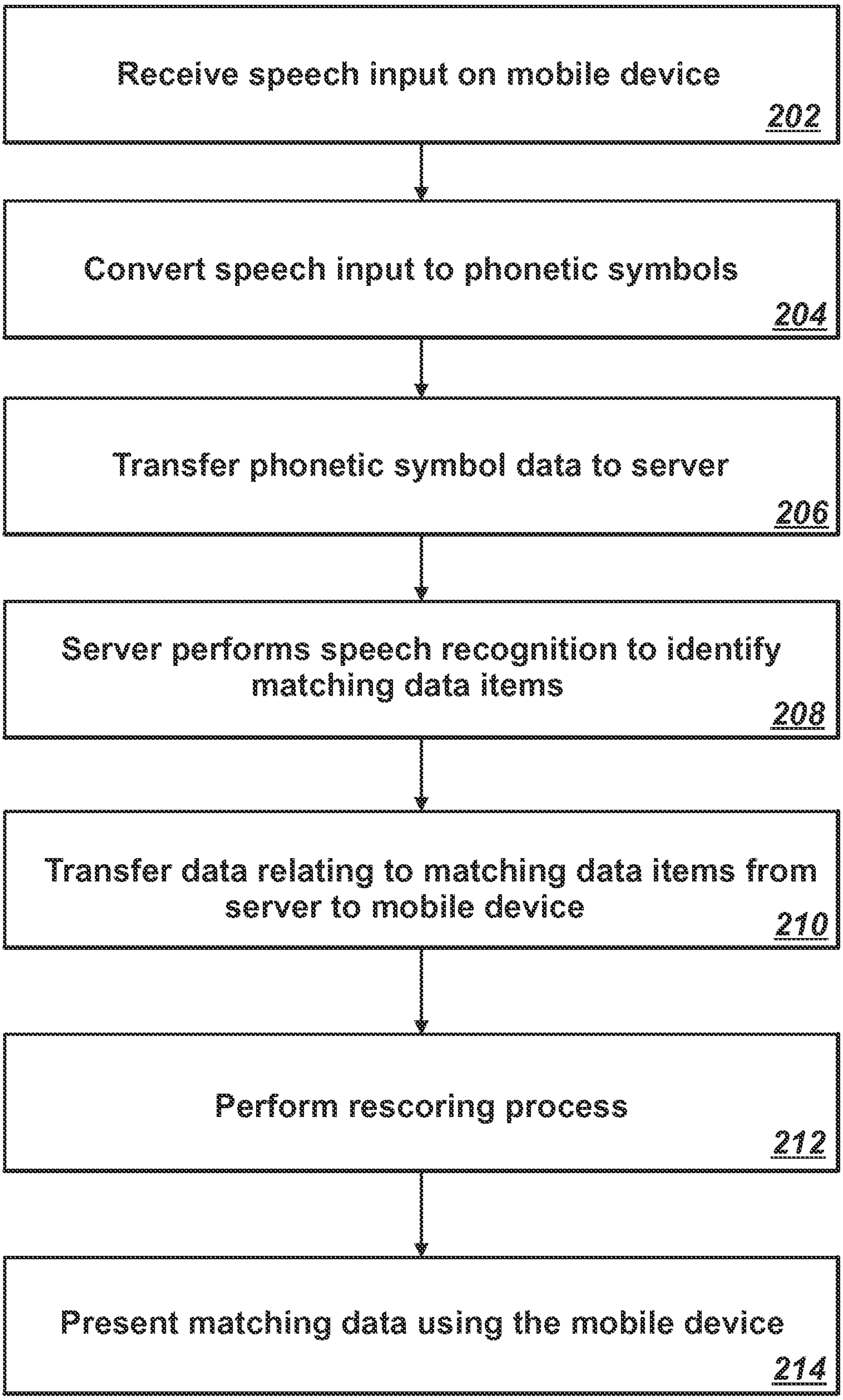

Receive speech input on mobile device
202
Convert speech input to phonetic symbols
204
Transfer phonetic symbol data to server
206
Server performs speech recognition to identify matching data items
208
Transfer data relating to matching data items from server to mobile device
210
Perform rescoring process
212
Present matching data using the mobile device
214

*Fig. 2*

# INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2009/051684

A CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/28 G10L15/26

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
GIOL

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

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

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where approp πate, of the relevant passages | Relevant to claim No |
|---|---|---|
| X<br>Y | JP 2003 044091 A (NTT DOCOMO INC)<br>14 February 2003 (2003-02-14)<br>the whole document<br>----- | 1-9,<br>19-24<br>10-18 |
| Y | EP 0 984 430 A (MATSUSHITA ELECTRIC IND CO LTD [JP]) 8 March 2000 (2000-03-08)<br>the whole document<br>----- | 10-18 |
| A | WO 95/17746 A (QUALCOMM INC [US])<br>29 June 1995 (1995-06-29)<br>figures 2,3<br>page 6, line 6 - page 9, line 19<br>----- | 1-24 |
| A | US 2006/195323 Al (MONNE JEAN [FR] ET AL)<br>31 August 2006 (2006-08-31)<br>the whole document<br>-----<br>-/-- | 1-24 |

☒ Further documents are listed in the continuation of Box C

☒ See patent family annex

' Special categones of cited documents

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

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

¹L" document which may throw doubts on pnoπty claιm(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

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

"P¹ document published prior to the international filing date but later than the pnority date claimed

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

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

"Y¹ document of particular relevance, the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such docu¬ ments, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| | |
|---|---|
| Date of the actual completion of the international search<br>3 March 2010 | Date of mailing of the international search report<br>12/03/2010 |
| Name and mailing address of the ISA/<br>European Patent Office, P B 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel (+31-70) 340-2040,<br>Fax (+31-70) 340-3016 | Authoπzed officer<br>Chetry, Nicol as |

# INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2009/051684

C(Contınuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropπate of the relevant passages | Relevant to claim No |
|---|---|---|
| A | US 2005/043946 Al (UEYAMA TERUHIKO [JP] ET AL) 24 February 2005 (2005-02-24)<br>figures 5,6<br>paragraph [0100] - paragraph [0138]<br>----- | 1-24 |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2009/051684

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| JP 2003044091 A | 14-02-2003 | NONE | |
| EP 0984430 A | 08-03-2000 | DE 69922104 Dl | 30-12-2004 |
| | | DE 69922104 T2 | 15-12-2005 |
| | | ES 2233002 T3 | 01-06-2005 |
| | | US 6684185 Bl | 27-01-2004 |
| WO 9517746 A | 29-06-1995 | AT 261172 T | 15-03-2004 |
| | | AU 692820 B2 | 18-06-1998 |
| | | AU 1375395 A | 10-07-1995 |
| | | BR 9408413 A | 05-08-1997 |
| | | CN 1138386 A | 18-12-1996 |
| | | DE 69433593 Dl | 08-04-2004 |
| | | DE 69433593 T2 | 03-02-2005 |
| | | EP 0736211 Al | 09-10-1996 |
| | | FI 962572 A | 20-08-1996 |
| | | HK 1011109 Al | 14-01-2005 |
| | | JP 9507105 T | 15-07-1997 |
| | | JP 3661874 B2 | 22-06-2005 |
| | | ZA 9408426 A | 30-06-1995 |
| US 2006195323 Al | 31-08-2006 | AT 441175 T | 15-09-2009 |
| | | CN 1764945 A | 26-04-2006 |
| | | EP 1606795 Al | 21-12-2005 |
| | | ES 2331698 T3 | 13-01-2010 |
| | | FR 2853127 Al | 01-10-2004 |
| | | WO 2004088636 Al | 14-10-2004 |
| US 2005043946 Al | 24-02-2005 | NONE | |