OPIC
Office de la propriété intellectuelle du Canada

CIPO
Canadian Intellectual Property Office

(12) (19) (CA) **Demande-Application**

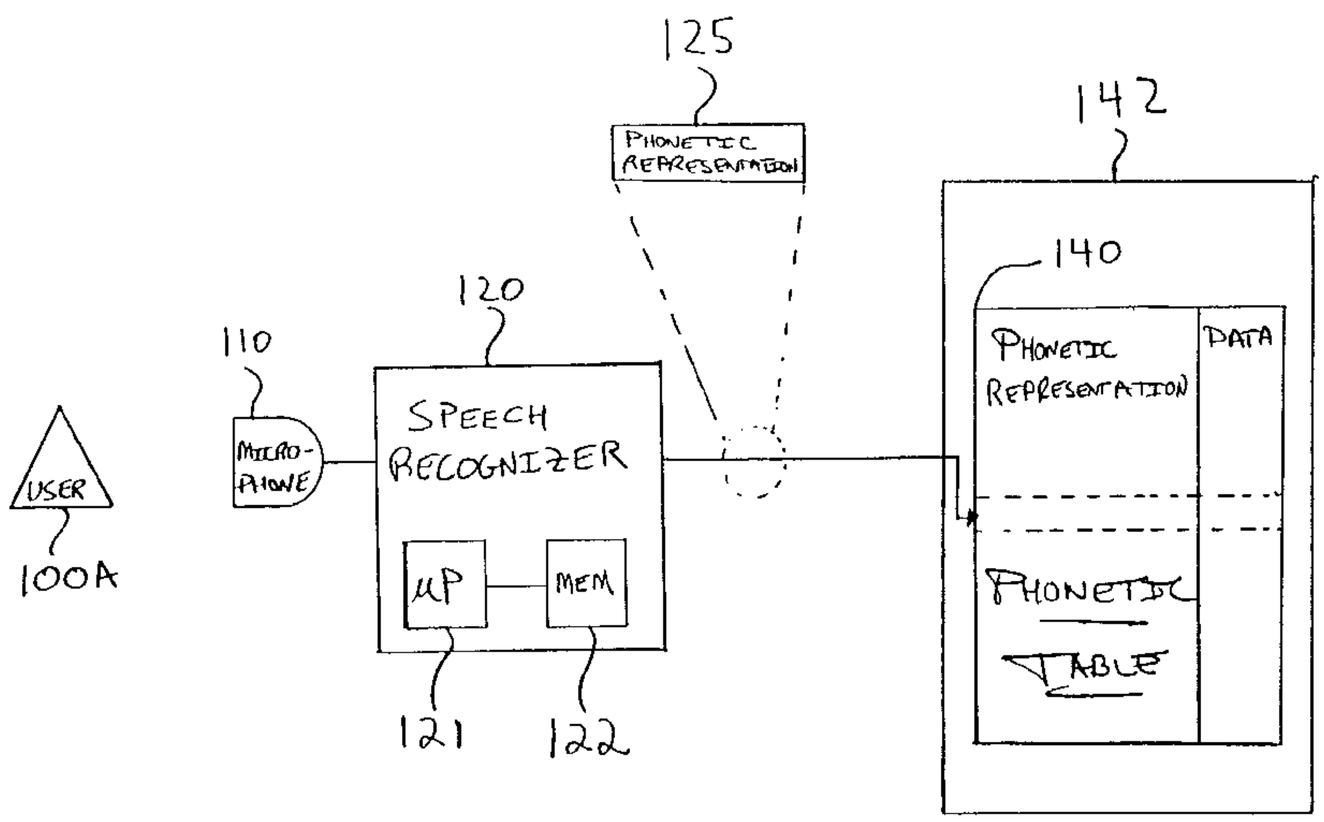
(21) (A1) **2,308,951**
(22) 2000/05/19
(43) 2000/11/21

(72) GEILHUFE, MICHAEL, US
(72) MACMILLAN, DAVID, US
(72) BAREL, AVRAHAM, IL
(72) BROWN, AMOS, IL
(72) BOOTSMA, KARIN LISSETTE, US
(72) GADDY, LAWRENCE KENT, US
(72) PYO, PHILLIP PAUL, US
(72) SHEN, WADE, US
(71) INFORMATION STORAGE DEVICES, INC., US

(51) Int.Cl.[7] G10L 15/22, G06F 13/38, G06F 3/16
(30) 1999/05/21 (60/135,301) US
(30) 2000/05/09 (09/567,858) US

(54) **METHODE ET DISPOSITIF S'APPLIQUANT A DES DISPOSITIFS A COMMANDE VOCALE AVEC STOCKAGE, UTILISATION, CONVERSION, TRANSFERT ET RECONNAISSANCE AMELIORES D'EXPRESSIONS**

(54) **METHOD AND APPARATUS FOR VOICE CONTROLLED DEVICES WITH IMPROVED PHRASE STORAGE, USE, CONVERSION, TRANSFER, AND RECOGNITION**



(57) The present invention provides for the storage of speech phrases. Speech phrases are processed by a speaker-independent speech recognition engine of a voice controlled device. This engine returns a speaker-independent representation of the phrase. The speaker-independent representation is stored. Method of converting text to speaker-independent representations of speech and speaker-independent representations of speech into text.

ABSTRACT OF THE DISCLOSURE

The present invention provides for the storage of speech phrases. Speech phrases are processed by a speaker-independent speech recognition engine of a voice controlled device. This engine returns a speaker-independent representation of the phrase. The speaker-independent representation is stored. Method of converting text to speaker-independent representations of speech and speaker-independent representations of speech into text.

METHOD AND APPARATUS FOR VOICE CONTROLLED DEVICES WITH IMPROVED PHRASE STORAGE, USE, CONVERSION, TRANSFER, AND RECOGNITION

CROSS-REFERENCES TO RELATED APPLICATIONS

This non-provisional United States (US) patent application claims the benefit of US Provisional Application No. 60/135,301 filed on May 21, 1999 by inventors GEILHUFE et al.

This application is related to United States patent application Serial No. 09/316,332, filed by inventors GEILHUFE et al, Attorney Docket No. 042236.P050, entitled "METHOD AND APPARATUS FOR STANDARD VOICE USER INTERFACE AND VOICE CONTROLLED DEVICES" and to be assigned to Information Storage Devices, Inc. the disclosure of which is hereby incorporated by reference, verbatim and with the same effect as though it were fully and completely set forth herein.

This application is also related to United States patent application Serial No. 09/316,643, filed by inventors GEILHUFE et al, Attorney Docket No. 042236.P051, entitled "METHOD AND APPARATUS FOR CONTROLLING VOICE CONTROLLED DEVICES" and to be assigned to Information Storage Devices, Inc. the disclosure of which is hereby incorporated by reference, verbatim and with the same effect as though it were fully and completely set forth herein.

This application is also related to United States patent application Serial No. 09/316,604, filed by inventors GEILHUFE et al, Attorney Docket No. 042236.P052,

entitled "METHOD AND APPARATUS FOR ENHANCING ACTIVATION OF VOICE CONTROLLED DEVICES" and to be assigned to Information Storage Devices, Inc. the disclosure of which is hereby incorporated by reference, verbatim and with the same effect as though it were fully and completely set forth herein.

This application is also related to United States patent application Serial No. 09/316,334, filed by inventors GEILHUFE et al, Attorney Docket No. 042236.P053, entitled "METHOD AND APPARATUS FOR IDENTIFYING VOICED CONTROLLED DEVICES" and to be assigned to Information Storage Devices, Inc. the disclosure of which is hereby incorporated by reference, verbatim and with the same effect as though it were fully and completely set forth herein.

This application is also related to United States patent application Serial No. 09/316,666, filed by inventors GEILHUFE et al, Attorney Docket No. 042236.P055, entitled "METHOD AND APPARATUS FOR MACHINE TO MACHINE COMMUNICATION USING SPEECH" and to be assigned to Information Storage Devices, Inc. the disclosure of which is hereby incorporated by reference, verbatim and with the same effect as though it were fully and completely set forth herein.

To the extent that a conflict arises through the incorporation of the preceding documents, the description of the present invention herein shall control.

FIELD OF THE INVENTION

This invention relates generally to machine interfaces. More particularly, the invention relates to storage, use, conversion and recognition of phrases.

BACKGROUND OF THE INVENTION

Previously, electronic devices were controlled by manual input from a human. More recently, voice controlled devices have been introduced including computers having voice control software for control by a human user's speech.

In voice controlled devices, it is desirable to store phrases under voice control. As defined herein, a phrase is defined as a single word, or a group of words treated as a unit. This storing might be to set options or create personalized settings. Once stored, a phrase can later be used as a command or with a command as a data field or other object. A command is usually provided by a user to control a device. For example, in a voice-controlled telephone, it is desirable to store people's names and phone numbers under voice control into a personalized phone book. At a later time, this phone book can be used to call people by speaking their name (e.g. "Cellphone call John Smith", or "Cellphone call Mother").

Prior art approaches to storing the phrase ("John Smith") operate by storing the phrase in a compressed, uncompressed, or transformed manner that attempts to preserve the actual sound. Detection of the phrase in a command (i.e. detecting that John is to be called in the example above) then relies on a sound-based comparison between the original stored speech sound and the spoken command. Sometimes the stored waveform is transformed into the frequency domain and / or is time adjusted to facilitate the match, but in any case the fundamental operation being performed is one that compares the actual sounds. The stored sound representation and comparison for detection suffers from a number of disadvantages. If

a speaker's voice changes, perhaps due to a cold, stress, fatigue, noisy or distorting connection by telephone, or other factors, the comparison typically is not successful and stored words are not recognized. Because the word or phrase is stored as a sound representation, there is no way to extract a text-based representation of the word or phrase. A sound stored phrase is strictly sound based. Additionally, storing a sound representation results in a speaker dependent system. It is unlikely that another user could speak the same word or phrase using the same sounds in a command and have it be correctly recognized. It would not be reliable, for example, for a secretary to store phonebook entries and a manager to make calls using those entries. It is desirable to provide a speaker independent storage means.

Additionally, if the words or phrases are stored as sound representations, the stored phrases can not be used in another speech recognition device unless the same waveform processing algorithms are used by both devices. Thus, transferring data associated with the stored sound phrases between devices, such as phone numbers in a phonebook, a cellphone or electronic organizer, is impractical unless the devices use the exact same speech recognition engine. It is desirable to recognize spoken words or phrases and store them in a representation such that, once stored, the phrases can be used for speaker independent recognition, can be used by multiple devices, and can be merged with the representations of other devices. Additionally, it is desirable to store information in text form, and to use it later in voice commands. For example, a text-based phonebook from a personal computer or organizer might be loaded into a cellphone with the text-based representation of the name

John Smith and his phone number. In this case, it is desirable for any arbitrary speaker to be able to place a call using voice control (e.g. "Cellphone call John Smith").

SUMMARY OF THE INVENTION

The present invention includes a method, apparatus and system for storage of phrases using a speaker-independent representation, and for speech recognition that uses this representation, as described in the claims. Briefly, the present invention provides for the initial storage of a spoken phrase (for example, when making a new phonebook entry under voice control). This is processed by the speaker-independent speech recognition engine of the voice controlled device. This engine returns a speaker-independent representation of the phrase. This speaker-independent representation is what is stored.

When a spoken command is issued to the voice controlled device, it is also processed by the speaker-independent speech recognition engine of the present invention. This could be the same speaker-independent engine used for storing the original entries, or a completely different speaker-independent engine. In either case, the engine returns a speaker-independent representation of the phrase. This speaker-independent representation can be compared to earlier stored representations to determine whether this phrase is recognizable.

By using a speaker-independent representation for the stored entries and the phrases spoken later, a number of advantages result. Command recognition will be reliable even if a user's voice has changed due to a cold, stress,

fatigue, transmission over a noisy or distorting phone link, or other factors. Additionally, if a way to convert text to speaker-independent representations is provided, text-based information can reliably be used in spoken commands. Furthermore, by storing speaker independent representations of speech, recognition can be speaker-independent and is reliable even if someone else had stored the original voice entry. Stored phrases originating from multiple text sources and from different speakers can be merged and reliably used in speaker-independent recognition. The use of speaker independent stored entries facilitates upgrading to improved speaker-independent engines as they become available. New speaker-independent engines can use existing stored information without impacting reliability or requiring re-storage, since all stored entries are held in speaker-independent form. Using the speaker-independent stored entries can provide downward compatibility. New information, stored using the new improved speech recognition engines, can be used as commands in voice controlled devices having older speech recognition engines. Old and new generations of equipment can inter-operate without prior coordination by using the speaker independent representations. This allows, for example, two PDAs to exchange voice-stored phonebook entries and provide reliable recognition to the new users of that information. Finally, there are no legacy restrictions to hold back or restrict future development of speaker-independent recognition engines as long as they can create speaker-independent outputs, unlike waveform-storage based systems, which must always be able to perform exactly the same legacy waveform transformations.

OBJECTIVES OF THE INVENTION

A first objective of the present invention is to allow a phrase to be stored by one user, and later have that phrase spoken by the same user and properly recognized by the voice controlled device, even if the sound of the user's speech is different. The users' speech may be different for any reason, including sickness, stress, fatigue, or transmission over a noisy or distorting telecommunications link.

A second objective of the present invention is to allow a phrase to be stored by one user, and later have that phrase spoken by a different user and recognized correctly by the voice controlled device.

A third objective of the present invention is to allow a phrase to be stored on a first device by one user and then have the phrase be transferred to other devices, where it can be correctly recognized whether it is spoken by the same or another user.

A fourth objective of the present invention is to allow phrases stored using one speech recognition engine to be used for recognition with a different version or different model of speech recognition engine.

A fifth objective of the present invention is to allow manufacturers to continue to develop speech recognition engines in parallel, independent of each other. This can occur because phrases stored on early models of recognizers can be recognizable on later models of recognizers. In addition, phrases stored on later models can be recognized on earlier models. Furthermore, phrases stored by one manufacturer's recognizer can be recognized by another's. Included in this objective is providing an invention that preserves this capability even between speech recognition engines not yet invented.

A sixth objective of the present invention is to permit the consolidation of phrases entered by speech, and phrases entered by text (including Caller-ID, text-based personal or public phone books, keypad entry or other means), into a single set of phrases that can be recognized.

A seventh objective of the present invention is to allow the capture of directory assistance numbers which can alter be retrieved by a person speaking the associated name.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram of a voice controlled device illustrating phrase storage.

Figure 2 is a block diagram of a voice controlled device illustrating phrase recognition.

Figure 3 is a block diagram of an alternate embodiment of the recognizer and comparator of FIG. 2.

Figure 4 is a block diagram of an alternate method by which recognition result may be represented.

Figure 5 is a block diagram of two voice controlled devices illustrating exchange of phrases to be recognized.

Figure 6 is a block diagram illustrating the method by which text-based representations can be incorporated into a phonetic speech recognition system.

Figure 7 is a block diagram of the method of formation of a mapping dictionary.

Figure 8 is a block diagram of the method of capturing data by a speech interface.

DETAILED DESCRIPTION

Speech recognition software is now available that can perform speaker-independent recognition and which generates speaker-independent representations of the spoken speech. The present invention uses speaker-independent recognition software in combination with other elements for storing phrases, using phrases, converting phrases and recognizing phrases. As defined herein, a phrase is defined as a single word, or a group of words treated as a unit. The present invention communicates using audible and non-audible speech. Speech as defined herein for the present invention encompasses a) a signal or information, such that if the signal or information were passed through a suitable device to convert it to variations in air pressure, the signal or information could be heard by a human being and would be considered language, and b) a signal or information comprising actual variations in air pressure, such that if a human being were to hear the signal, the human would consider it language. In the preferred embodiment, the speaker-independent representation of speech is a phonetic representation of speech. Other speaker-independent representations of speech may also be used in accordance with the present invention.

Referring to FIG. 1, consider the storage of phrases. During the storing of a phrase, a user 100A speaks the phrase to be stored into microphone 110. The phrase is processed by the speech recognizer 120, which generates the speaker-independent phonetic representation of the speech 125. This representation is entered into Table 140, possibly with additional data. Speech recognizer 120 includes a processor 121 and software code stored in

storage medium 122. Table 140 is resident in storage medium 142. Storage medium 122 and 142 may be in the same physical storage device or medium, or separate devices or medium.

Consider for example, a telephone application with a dial-by-spoken-name feature, in which the telephone is issued the command "Dial John Smith". The additional data in Table 140 might be one or more phone numbers for John Smith. In this example, the phrase stored as a phonetic representation is "John Smith".

Speech recognizer 120 can also be used to capture the phone number information, converting from a spoken phone number to digital digits, or the phone number can be entered by a keypad or other input means.

Depending on the type of device using the present invention, there may be a check prior to storing the phrase in Table 140 to see if there is already an entry with the same or similar phonetic representation, such that errors in recognition might occur if both phrases were stored. If there is, the user can be prompted to take appropriate action, depending on the type of device utilized. Considering the telephone example, a user might try to store a second phone number for the same specific person. This might be acceptable if there is a way to determine which number was intended to be called (e.g. the user says "Call John Smith", and the machine prompts "Say 'first' to call 234-5678 and say 'second' to call 987-6543.") Alternatively, if there is no way provided to select from multiple phone numbers for the same person when calling, the user may be prompted while storing a second entry to rename the new entry. For example, the user can store it under "John Smith Office" to

differentiate it from the phrase "John Smith", assuming "John Smith" is already associated with another number such as a home number. There are a variety of ways to handle these situations.

Once there are one or more entries stored in Table 140, a user can make use of them. As shown in Figure 2, the user 100A can speak a phrase, which is picked up by microphone 110 and converted to a phonetic representation by speech recognizer 120. The phonetic representation is compared to the entries in Table 140 by a comparator 130. In Figure 2, the comparator 130 is shown separate from the speech recognizer 120. It is possible to incorporate the comparator function into the speech recognizer. This variation is shown in Figure 3 as speech recognizer and comparator 120'. In either event, as a result of the comparison, results are returned which can be a result set of no matches, one match, or multiple matches. The computer may look for perfect matches or also good matches as is customary in speech recognition. This result set is represented by Results 135, which can include the phonetic representation(s) found, the data element(s) found, both, or some other results indicator appropriate for the intended application. As an example of an alternative results indicator, Figure 4 shows results 135' consisting of a pointer or index or multiple pointers or indexes into Table 140 indicating which elements were matched. In Figure 4, rows numbered 2, 3, and 5 in Table 140 are illustrated as being successfully matched by results 135' by pointers 401 over the set of rows 402.

As stated, it is possible to have the recognizer 120 or 120' identify multiple possible matches. In this case, a variety of alternatives can be used to narrow the selection down to a single matching result prior to

further processing, if a single matching result is required. Alternatively, multiple selections can each be processed further, which may or may not include a later selection of a single matching result from multiple matching results. Possible ways to determine which of multiple results should be selected as a best match include asking the user to specify which was the best match, having the recognizer pick one of the multiple matches, or have the recognizer request the user to re-speak the phrase. Having the user re-speak the phrase may result in a single match due to slightly different voicing by the speaker. Alternatively, the first and subsequent speaking of the phrase may be used collectively to help identify a single best match for the result.

The representation of phrases in Table 140 against which the speech engine compares the microphone input is phonetic in nature. Some speech recognition systems attempt to identify a phrase by comparing against the actual sounds spoken by a user for target phrases. The Dynamic Time Warping method is one of these that makes use of the approach of comparing actual sounds. However, the present invention relies on a speaker-independent system, which in the preferred embodiment operates on a phonetic representation of target phrases stored in table 140.

There are multiple ways of representing phonetic pronunciations, including representation of phonemes, representation of smaller sound elements than phonemes, or representation of larger sound elements than phonemes. One can also have a phonetic representation that consisted of combinations of these elements. The present invention encompasses these variations, along with other speaker-independent representations.

With some phrases, there are multiple acceptable ways of saying them. For example, the word "lead" can be pronounced "led" or "leed". There are many other examples such as this in English and other languages. In these cases, there can be multiple entries for a target phrase in Table 140 to account for the multiple pronunciations and / or the code of the speech recognizer can be programmed to account for these variations.

Because Table 140 stores a speaker independent representation of a spoken phrase, specific advantages are obtained. First, the recognition process is insensitive to changes in the speakers voice, as might be caused by sickness, stress, fatigue, transmission over a noisy or distorting phone link, or other factors.

Second, it is possible for one user, for example user 100A in Figure 1, to store a phrase, and another user, for example user 100B in Figure 2, to use the phrase in a spoken command. This is useful, for example, in allowing a secretary to store phone numbers in a cellphone for dial-by-name use, which is later used by a manager to make calls.

Third, with reference to Figure 5, the present invention allows information in a first device 501 that has a Table 140, containing phonetic representations and data, to be transferred by transfer means 90 and 90' to a second device 502 that has a table 240 containing phonetic representations and data. Table 140 and Table 240 may have different types of data and different phonetic representations (or, in the general case, different speaker-independent representations), in which case a conversion means 95 can be used to provide translation. The phonetic representations and data of Table 140 being

transferred into Table 240 may replace all existing information in Table 240, may replace or update selected records, or may be appended to the existing records in Table 240. It is also possible to include a date-time stamp or other information within the data, which can be used, in addition to the phonetic representation, to determine how the Table 140 and Table 240 data are combined. For example, where there are duplicate phonetic representations, the data associated with the newest representation can be preserved. The specific actions taken will depend on the type of device, characteristics of the information, and the variety of known methods of merging, updating, replacing and synchronizing tables of information which is used and widely documented in public literature. The present invention encompasses these variations.

Once the phonetic representations and data from Table 140 are incorporated into Table 240, the entire set of Table 240 entries can be used for recognition by any individual. The phrases stored in Table 240 might be spoken by the same user 100A who originally recorded the information in Table 140. Alternatively, the phrases stored in Table 240 might be spoken by the user 100B who recorded the original information in Table 240 before the transfer of information from Table 140. Alternatively, the phrases might be spoken by a user 100C who recorded none of the Table 140 or Table 240 information. Because the information stored in Table 240 is phonetically based, the speech recognizer can use it to recognize speech from any of these users 100A, 100B, or 101C.

Furthermore, since the data in Table 240 is phonetically based, it is not necessary for speech recognizer 220 to have the same speech recognition

software as speech recognizer 120. For example, one version of the speech recognition software may be released later having more sophisticated speech recognition algorithms than the other. Alternatively, they might be speech recognizers from different manufacturers with phonetic recognizers. By using conversion means 95 to perform conversion from Table 140's phonetic representation to Table 240's representation, it is not even necessary for the phonetic representations or data of Table 140 and Table 240 to be identical. It is only necessary that there be some mapping that can be performed by between Table 140's phonetic representation and Table 240's phonetic representation, and mappings for any portions of the data fields to be transferred. These mappings are performed by the conversion means 95. The goal is to have the same phonetic information (but not necessarily the same representation of that phonetic information) in Table 140 and Table 240 for the transferred records, and the same data field information (but not necessarily the same representation of that data field information) in Table 140 and Table 240 for the transferred records. More generally, with respect to the stored phrases, the goal is to have the same speaker-independent information (but not necessarily the same speaker-independent representation) in Table 140 and Table 240 for the transferred records. Moreover, while perfect conversion is preferable, if minor conversion errors are introduced by conversion means 95; perhaps due to difficulties in mapping between the two representations, algorithm errors or other issues so that the phonetic information is similar but not precisely the same; acceptable recognition is achievable.

This ability to transfer phonetic representations and associated data between devices with different versions of speech recognizers or recognizers from different manufacturers is an important issue. This means recognizers can be improved without being held back by legacy phonetic representation and data representation issues. Development can also proceed in parallel at multiple manufacturers, each working independent of the other. The approach also is valid even for those speech recognizers not yet conceived or invented. In all cases, the only requirement is that there be some mapping that can be implemented by conversion means 95 between the current phonetic representation and that used in the future system. This is in contrast to speech recognition systems that use dynamic time warping and other approaches that attempt to match actual sound patterns, where it is very difficult to change or improve the algorithm without losing the ability to recognize previously captured phrases.

Returning for a moment to conversion means 95, and transfer means 90 and 90'. They can be unidirectional, in which case data can be transferred from Table 140 to Table 240, but not from Table 240 to 140. Or they can be bi-directional, in which case data can move from Table 140 to table 240, or from Table 240 to Table 140.

So far the description of the present invention has focussed on the phonetic representation of the phrase. In many cases, it is also helpful to have a corresponding text representation. The present invention provides a variety of ways that a text representation can be matched to a phonetic representation.

A first method by which a text representation can be matched to a phonetic representation is by direct translation. Figure 6 shows a Text Table 170 that contains one or more records that each consist of a text representation of a phrase and possibly additional data. To allow recognition of a text representation of a phrase, it is necessary to convert the text representation into a speaker independent representation of Table 140, which in the preferred embodiment is a phonetic representation. The Spelling to Pronunciation Converter 160 in Figure 6, which is commercially sold by Conversational Computing Corporation as part of their speech-controlled web browsing product, has such a capability. By means of the Spelling to Pronunciation Converter 160, records 175 from table 170 can be converted to Records 150 which are phonetic representations and can be loaded into Table 140. Once a phonetic representation record 150 is created, it can be incorporated into table 140 in a variety of ways. The choice of incorporation method depends on the particular device and result desired. Some alternatives include always appending the record to Table 140, making the storage of the record in the Table 140 conditional on the results of first searching for a pre-existing identical or similar record in the table, making storage of the record conditional on additional information, or performing some type of merging of the new record with one or more pre-existing records. The specific actions taken will depend on the type of device and characteristics of the information, and a variety of methods of merging, updating, replacing and synchronizing tables of information which are widely available and documented in public literature. The present invention encompasses all these variations.

In any event, once the information has been moved from text Table 170 to phonetic Table 140, the speech recognition process described elsewhere in this document can be used to compare incoming speech to the phrases in Table 140.

A second method by which a text representation can be matched to a phonetic representation is by converting phonetic representations to text representations. This can be done by phonetic-to-text software algorithms or through use of a dictionary formed using spelling-to-pronunciation translation. The dictionary method is shown in Figure 7. The dictionary 700 consists of a series of records, each of which has space for a text representation 701 and a phonetic representation 702. Initially, the set of all possible text representations is loaded into the text representation 701 field of the records, and the phonetic representation field 702 is left blank.

Each text representation 701 is then processed by spelling to pronunciation conversion software 160 to create the corresponding phonetic representation, which is stored in the associated phonetic representation field. At the end of this process, the dictionary shows the text representation for each possible phonetic representation. Given a phonetic representation, the dictionary can be scanned to determine whether there is none, one or more than one text representation 701 for the given phonetic representation 702. The dictionary can be sorted or indexed on the phonetic representation field 702 to speed this lookup. In the event there is more than one text representation 701, a variety of options are possible, including returning all possible text representations 701 entries, or making some type of machine-based or user-assisted selection of which to use.

In the two above methods for matching text representations to phonetic representations, while perfect conversion of the spelling-to-pronunciation is preferable, even if minor conversion errors occur, acceptable recognition or matching to existing records is still be achievable, providing the matching process looks for the best match rather than a perfect match.

A third method by which text representations and phonetic representations can be matched is by comparing the contents of the data fields. In this approach, with reference to Figure 6, the data field of Phonetic Table 140 and the data field of Text Table 170 both share common data elements that help uniquely identify which text representation can be paired with a phonetic representation. For example, in a telephone application, the data fields of both Phonetic Table 140 and Text Table 170 might have the phone number for a person. The name of this person is stored phonetically in the phonetic representation field of Phonetic Table 140 alongside the data field with their phone number, perhaps due to the user storing them as described above for Figure 1. The name of this person is also stored textually in the text representation field, along with their phone number in the data field of Text Table 170. By identifying matching phone numbers in Phonetic Table 140 and Text Table 170, matching phonetic representations and text representations can be identified. A best match comparison rather than a perfect match comparison, perhaps with a required limit on the allowed degree of mismatch, may be needed for some applications to accommodate for possible minor differences, for example errors arising from typos, in the data field keys.

A fourth method by which text representations and phonetic representations can be matched is by way of a spoken spelling interface, in which the user spells the text that corresponds to a phonetic representation. A description of how spoken spelling information can be converted to a text representation is contained in US patent 5,799,065. A best match, rather than a perfect match, can be used to accommodate minor spelling errors.

Reconsider the telephone example. If there are a multiplicity of phone numbers associated with one record and no other, and a match is made between the phonetic representation and text representation, then both the phonetic representation and text representation can be associated with the multiplicity of phone numbers.

More specifically, assume for example that the information in Table 170 has been created by loading information from a personal digital assistant, and it includes office, home and cellular phone numbers. Assume the user 100A has stored, by the process depicted in Figure 1, the office phone number in Table 140. By locating the match between the Table 140 and Table 170 records, the user can speak the target's name and have access to all three numbers - office, home and cellular.

In the more general case, once matching relationships are found between phonetic Table 140 entries and text Table 170 entries, the entries in both tables can be provided including the union of all information in the data fields of the matching records of both Tables.

Consider now how text entries for text Table 170 might originate for a telephone application. Referring to Figure 6, alternatives for text entries include text information from a telephone's Caller-ID system 186, from

public phone books 185, computer information 180 such as from personal computer or PDA (Personal Digital Assistant) databases or applications, or information from a keyboard 190 or other input means attached to a device, such as microprocessor 195, having access to the Text Table 170.

Alternatively, a user can store a phone number by speech such as by saying the number and an identifier (e.g., 234-5678 for John Smith). In this case, the phonetic representation of the name "John Smith" is stored in the phonetic representation field of a record in Table 140, and the phone number is stored in the data field. Assume the telephone is equipped with a Caller-ID capability 186 that provides the phone number and text name of the party on the other end of the telephone call. The Caller ID name can be processed with spelling to phonetic software. If there is a reasonable match to the phonetic representation, it can be stored as a text representation in Table 170 with the phone number in the data field and the text name of the caller in the text representation field. (If there is not a close match, it may indicate that the phone's registration and hence Caller ID information is in another person's name, in which case it may not be desirable to store the text representation.) By comparing the phone numbers in the data fields of Table 140 and Table 170, it is possible to create a representation of the complete information for the caller, including the text name, the phonetic representation, and the combination of the data fields. If John Smith later calls the telephone from another location, that Caller ID record can be added so that there are now two phone numbers accessible by speaking John Smith's name. Loading records from or synchronization of records with other data sources, including those shown in

Figure 6, can further enhance the information available by voice control.

Another method of capturing additional data is by a speech interface. For example, consider a telephony application, as shown in Figure 8, consisting of a user 100A, a voice controlled telephone 800, a directory assistance service 810, and a communications medium 805 that connects voice controlled telephone 800 to directory assistance service 810. The directory assistance service 810 can be a human, a machine based system, or a system using a mixture of human and machine interactions. When a user speaks the command to call a name not currently in the voice controlled Telephone 800's Table 140, the telephone can automatically dial directory information and in response to hearing the phrase "name" (as in the operators inquiry "what name please?"), automatically pass the requested name to the operator. The speech recognizer can also listen for the report from directory assistance service 810 of the number for the desired person, and capture this. The spoken name's phonetic representation can be stored in the phonetic representation portion of a record in table 140, and the phone number stored in the data portion. If that person later calls the phone, their Caller ID number or name can be sued to identify a matching phonetic representation stored. The Caller ID name can then be stored as the text representation in Table 170.

In general, one or more external data sources can be used to populate entries within text Table 170. The specific choices of external data sources depend on the nature of the application. Further elucidating the possible implementations, Table 170 can be located in the

same device that comprises Table 140, or they can be in separate devices connected via a communications method. The information of how Table 140 and Table 170 entries are matched can be stored in a variety of forms, including being stored by being copied or moved into a new table, stored in a Table 140 in which the text representation is included in the data field, stored in a Table 170 in which the phonetic information is included in the data field, stored using a third table that stores pointers to the records in tables 140 and Tables 170 that correspond, or through other means. The present invention is intended to address all these variations.

Finally, once the text and corresponding phonetic information is matched up, it can be transferred to other devices as described above and as displayed in Figure 5.

Additionally, the present invention is applicable to any language, including English, because it is based on speaker independent representations, including phonetic representations, which are applicable to any language.

In the preceding descriptions, it was stated that speech originated from a user, depicted in the Figures as 100A, 100B, 100C, and the like. It is within the scope of the invention that these users can be all humans, all machines with speech interfaces that interact with the machine of the present invention through speech, or any mixture of humans and machines. This includes machine-to-machine speech communication over wired and wireless links which may not include any audible speech.

Audible speech refers to speech that a human can hear unassisted. Non-audible speech refers to any encodings or representations of speech that are not included under the definition of audible speech, including that which may be

communicated outside the hearing range of humans and transmission media other than air.

The machine-to-machine speech communication possibilities includes the scenario where a plurality of communicating machines incorporate the present invention. The present invention includes the cases where machine-to-machine speech communication involves more than two machines, as might occur where there are multiple interacting devices within one room or on one telephone line.

The preferred embodiments of the present invention for "METHOD AND APPARATUS FOR VOICE CONTROLLED DEVICES WITH IMPROVED PHRASE STORAGE, USE, CONVERSION, TRANSFER, AND RECOGNITION" are thus described. While the present invention has been described in particular embodiments, the present invention should not be construed as limited by such embodiments, but rather construed according to the claims that follow below.

CLAIMS

What is claimed is:

1. A method for recognizing at least one phrase, the method comprising:

receiving at least one first phrase and converting the at least one first phrase into at least one speaker independent representation of speech;

receiving at least one second phrase which may or may not be the same as the at least one first phrase; and

comparing the at least one second phrase with the at least one speaker independent representation of speech.

2. The method of claim 1, wherein,

the speaker independent representation of speech is a phonetic representation of speech.

3. The method of claim 1, wherein,

the comparing generates match results and the match results are of the set of no match, one match, or plurality of matches.

4. The method of claim 1, wherein,

the comparing generates match results and the match results is a plurality of matches, the method further comprises:

selecting one of the plurality of matches as the best match.

5. The method of claim 4, wherein,

the selecting one of the plurality of matches as the

best match includes requesting a user to specify which is the best match.

6. The method of claim 4, wherein,

the selecting one of the plurality of matches as the best match includes randomly selecting a match of the plurality of matches as the best match.

7. The method of claim 4, wherein,

the selecting one of the plurality of matches as the best match includes receiving at least one additional phrase and the at least one additional phrase is used for additional determination of which speaker independent representation of speech is the best match.

8. The method of claim 1, wherein,

a first user communicates the at least one first phrase and the at least one second phrase.

9. The method of claim 1, wherein,

a first user communicates the at least one first phrase and a second user communicates the at least one second phrase.

10. The method of claim 1, wherein,

in the receiving the at least one first phrase and the at least one second phrase, the first and second phrases are audible speech which are received.

11. The method of claim 1, wherein,

in the receiving the at least one first phrase and the at least one second phrase, the first and second

phrases are non-audible speech which are received.

12. The method of claim 1, wherein,

in the receiving the at least one first phrase and the at least one second phrase, the at least one first phrase is non-audible speech and the at least one second phrase is audible speech which are received.

13. The method of claim 1, wherein,

in the receiving the at least one first phrase and the at least one second phrase, the at least one first phrase is audible speech and the at least one second phrase is non-audible speech which are received.

14. A method for recognizing at least one phrase, the method comprising:

receiving at least one first phrase and converting the at least one first phrase into at least one speaker independent representation of speech;

receiving at least one second phrase which may or may not be the same as the at least one first phrase;

converting the at least one second phrase into at least one speaker independent representation of speech for recognition; and

comparing the at least one speaker independent representation of speech for recognition with the at least one speaker independent representation of speech.

15. The method of claim 14, wherein,

the converting of the at least one first phrase into at least one speaker independent representation of speech is performed by a first speaker independent speech recognizer; and

the converting the at least one second phrase into at least one speaker independent representation of speech for recognition is performed by the first speaker independent speech recognizer.

16. The method of claim 14, wherein,

the converting of the at least one first phrase into at least one speaker independent representation of speech is performed by a first speaker independent speech recognizer; and

the converting of the at least one second phrase into at least one speaker independent representation of speech for recognition is performed by a second speaker independent speech recognizer.

17. The method of claim 14, wherein,

the first speaker independent speech recognizer and the second speaker independent speech recognizer are of the same design.

18. The method of claim 14, wherein,

the first speaker independent speech recognizer and the second speaker independent speech recognizer are of different designs.

19. A speech receiving device, comprising:

a processor;

a processor readable storage medium; and

code recorded in the processor readable storage medium for converting speech into speaker independent representations of speech.

20. The device of claim 19, further comprising:

code recorded in the processor readable storage medium for comparing speaker independent representations from converting speech, to stored speaker independent representations.

21. The device of claim 19, further comprising:

code recorded in the processor readable storage medium for converting speaker independent representations into text.

22. The device of claim 19, further comprising:

code recorded in the processor readable storage medium to send speaker independent representation of speech to other devices.

23. The device of claim 19, further comprising:

code recorded in the processor readable storage medium to receive speaker independent representations of speech from other devices.

24. The device of claim 19, further comprising:

code recorded in the processor readable storage medium to convert one type of speaker independent representations of speech to another type.

25. A speech receiving device, comprising:

a processor;

a processor readable storage medium;

code recorded in the processor readable storage medium for converting speech into speaker independent representations of speech;

code recorded in the processor readable storage medium for storing speaker independent representations of speech into a storage medium; and

code recorded in the processor readable storage medium for comparing speech with the speaker independent representations of speech stored in the storage medium to generate match results.

26. The device of claim 25, wherein,

the speaker independent representations of speech are phonetic representations of speech.

27. The device of claim 25, wherein,

the speech is audible speech.

28. The device of claim 25, wherein,

the speech is non-audible speech.

29. The device of claim 25, further comprising:

code recorded in the processor readable storage medium to select a best match of the match results.

30. The device of claim 29, wherein,

the code recorded in the processor readable storage medium to select a best match of the match results reports to a user the match results and requests the user to specify which is the best match of the match results.

31. The device of claim 29, wherein,

the code recorded in the processor readable storage medium to select a best match of the match results randomly selects a match of the match results as the best

match.

32. The device of claim 29, further comprising:

code recorded in the processor readable storage medium to use at least one additional phrase for additional determination as to which speaker independent representation of speech stored in storage is the best match.

33. A method of transferring speaker independent representations of speech between devices, the method comprising:

receiving at least one phrase at a first device;

converting the at least one phrase into at least one speaker independent representation of speech; and,

transferring the at least one speaker independent representation of speech to a second device.

34. The method of claim 33, wherein,

the at least one speaker independent representation of speech is a phonetic representation of speech.

35. The method of claim 33, wherein,

the first device utilizes one type of speaker independent representation of speech and the second device uses a different type of speaker independent representation of speech and the method further comprises:

converting the at least one speaker independent representation of speech into a type of speaker independent representation of speech compatible with the second device.

36. A method of transferring speaker independent representations of phrases between devices, the method comprising:

receiving at least one phrase at a first device;

converting the at least one phrase into at least one speaker independent representation of speech;

providing a set of speaker independent representations of speech associated with a second device; and

transferring the at least one speaker independent representation of speech from the first device to the second device for combining with the set of speaker independent representations of speech.

37. The method of claim 36, wherein,

the set of speaker independent representations of speech is empty.

38. The method of claim 36, wherein,

the set of speaker independent representations of speech has one speaker independent representation of speech.

39. The method of claim 36, wherein,

the set of speaker independent representations of speech has more than one speaker independent representation of speech.

40. The method of claim 36, wherein,

the at least one speaker independent representation of speech is a phonetic representation of speech.

41. The method of claim 36, wherein,

the at least one speaker independent representation of speech of the first device transferred to the second device is combined with the set of speaker independent representations of speech in the second device by merging the at least one speaker independent representation of speech with the second set of speaker independent representations of speech.

42. The method of claim 36, wherein,

the at least one speaker independent representation of speech of the first device transferred to the second device is combined with the set of speaker independent representations of speech in the second device by replacing the set of speaker independent representations of speech in its entirety with the at least one speaker independent representation of speech.

43. The method of claim 36, wherein,

the at least one speaker independent representation of speech of the first device transferred to the second device is combined with the set of speaker independent representations of speech in the second device by selectively replacing elements of the set of speaker independent representations with elements of the at least one independent representation of speech.

44. The method of claim 36, wherein,

there are date stamps associated with the at least one speaker independent representation of speech and with the set of speaker independent representations of speech, and the at least one speaker independent representation of speech is combined with the second set speaker independent

representations of speech using the date stamps.

45. The method of claim 36, wherein,

the first device has a first speaker independent speech recognizer;

the second device has a second speaker independent speech recognizer; and

the first speaker independent speech recognizer is of the same design as the second speaker independent speech recognizer.

46. The method of claim 36, wherein,

the first device has a first speaker independent speech recognizer;

the second device has a second speaker independent speech recognizer; and

the first speaker independent speech recognizer is of a different design than the second speaker independent speech recognizer.

47. The method of claim 36, wherein,

the first device and the second device use different speaker independent representations of speech, and

the method further comprises:

converting the at least one speaker independent representation of speech into a type of speaker independent representation of speech compatible with the second device.

48. The method of claim 47, wherein,

the type of speaker independent representation of speech compatible with the second device is a phonetic representation of speech.

49. A method of providing interoperability for speaker independent speech recognizers, the method comprising:

providing a first speech recognizer;

the first speech recognizer receiving at least one phrase;

converting the at least one phrase into at least one speaker independent representation of speech; and

providing the at least one speaker independent representation of speech to a second speech recognizer.

50. The method of claim 49, wherein,

the first speech recognizer is located within a device at a first moment;

the second speech recognizer is located within the device at a second moment; and

the first speech recognizer and the second speech recognizer are never located within the device at the same point in time.

51. The method of claim 49, wherein,

the first speech recognizer is located within a device at a first moment;

the second speech recognizer is located within the device at a second moment; and

at some point in time, the first speech recognizer and the second speech recognizer are both located within the device.

52. The method of claim 49, wherein,

the first speech recognizer is within a first device and the second speech recognizer is within a second device.

53. The method of claim 49, wherein,

the first speech recognizer and the second speech recognizer use the same type of speaker independent representations of speech.

54. The method of claim 53, wherein,

the first speech recognizer and the second speech recognizer are the same design.

55. The method of claim 53, wherein,

the first speech recognizer and the second speech recognizer are of different designs.

56. The method of claim 53, wherein,

the speaker independent representations of speech are phonetic representations of speech.

57. The method of claim 49, wherein,

the first speech recognizer and the second speech recognizer use different types of speaker independent representations of speech; and

the method further comprises:

converting the speaker independent representation of speech of the first recognizer into a type of speaker independent representation of speech compatible with the second speech recognizer.

58. The method of claim 57, wherein,

the at least one speaker independent representation of speech is a phonetic representation of speech.

59. The method of claim 57, wherein,

the type of speaker independent representation of speech compatible with the second speech recognizer is a phonetic representation of speech.

60. A method of transferring speaker independent representations of phrases between devices, the method comprising:

providing a first speech recognizer that operates with a first type of speaker independent representation of speech;

the first speech recognizer receiving at least one phrase;

converting the at least one phrase into at least one speaker independent representation;

providing a second speech recognizer that operates with a second type of speaker independent representation of speech;

providing a conversion means for converting the first type of speaker independent representation of speech into a type compatible with the second speech recognizer; and

providing a conversion means for converting the second type of speaker independent representation of speech into a type compatible with the first speech recognizer.

61. The method of claim 60, wherein,

at least one of the first or second type of speaker independent representations of speech is a phonetic representation of speech.

62. A method for converting speech into text, the method comprising:

receiving at least one phrase;

converting the at least one phrase into at least one speaker independent representation of speech;

communicating at least one second phrase; converting the at least one second phrase into at least one speaker independent representation of speech for recognition;

comparing the at least one speaker independent representation of speech for recognition with the at least one speaker independent representation of speech to generate a match result; and

converting the at least one speaker independent representation of speech into text.

63. The method of claim 62, wherein,

the at least one speaker independent representation of speech stored in storage is a phonetic representation of speech.

64. The method of claim 62, wherein,

the converting is performed by a speech-to-text converter.

65. The method of claim 62, wherein,

the converting provides one or more possible text results and the method further comprises:

processing at least one of the possible text results to create at least one possible speaker independent representation of speech; and

comparing the at least one speaker independent representation of speech stored in storage with the at least one possible speaker independent representation of speech to generate text.

66. The method of claim 62, further comprising:

providing a first additional data associated with the speaker independent representations of speech stored in storage;

providing one or more possible text results;

providing a second additional data associated with the possible text results; and

using relationships between the first additional data and the second additional data to assist in the converting.

67. The method of claim 66, wherein,
the first additional data and the second additional data include phone number information.

68. The method of claim 66, wherein,

the converting includes having a user speak the spelling of at least part of the text. 69. A method for developing speech recognizers that can use stored phrases from other speech recognizers, the method comprising:

identifying stored speaker independent representations of speech associated with a first speaker independent recognizer, at least some of the stored speaker independent representations of speech being created by providing at least one phrase to the first speaker independent recognizer; and

providing a second speaker independent speech recognizer that can use the stored speaker independent representations of speech associated with the first speaker independent recognizer.

69. A method for developing speech recognizers that can use stored phrases from other speech recognizers, the method comprising:

identifying stored speaker independent representations of speech associated with a first speaker independent speech recognizer, where at least some of the stored speaker independent representations of speech being created by providing at least one phrase to the first speaker independent speech recognizer; and

providing a second speaker independent speech recognizer that can use the stored speaker independent representations of speech associated with the first speaker independent speech recognizer.

70. The method of claim 69, wherein,

the first speaker independent speech recognizer uses a first type of speaker independent representation of speech;

the second speaker independent speech recognizer uses a second type of speaker independent representation of speech; and

to recognize phrases, the second speaker independent speech recognizer uses the stored speaker independent representations of speech associated with the first speaker independent speech recognizer by converting the stored speaker independent representations of speech associated with the first speaker independent speech recognizer into a type compatible with the second speaker independent speech recognizer.

71. A method for developing speech recognizers that can provide stored phrases to other speech recognizers, the method comprising:

providing a first speaker independent speech recognizer that creates speaker independent representations of speech from phrases; and

identifying a second speaker independent speech recognizer that can use the speaker independent representations of speech from the first speaker independent speech recognizer.

72. The method of claim 71, wherein,

the first speaker independent speech recognizer uses a first type of speaker independent representation of speech;

the second speaker independent speech recognizer uses a second type of speaker independent representation of speech; and

to recognize phrases, the second speaker independent speech recognizer uses the stored speaker independent representations of speech associated with the first speaker independent speech recognizer by converting the stored speaker independent representations of speech associated with the first speaker independent speech recognizer into a type compatible with the second speaker independent speech recognizer.

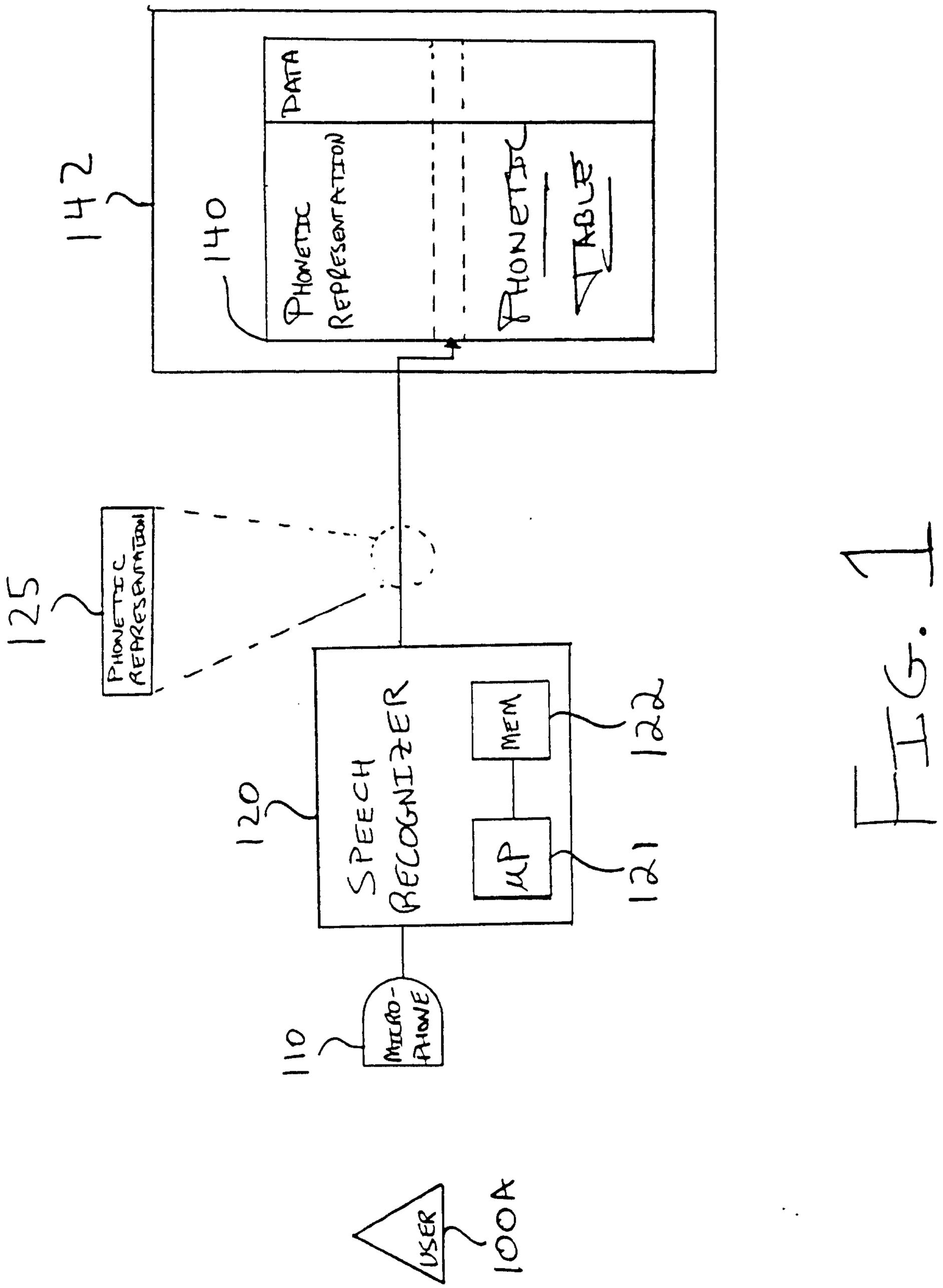

100A
USER
110
MICRO-PHONE
120
SPEECH RECOGNIZER
121
μP
122
MEMORY
125
PHONETIC REPRESENTATION
140
DATA
PHONETIC TABLE
142

FIG. 1

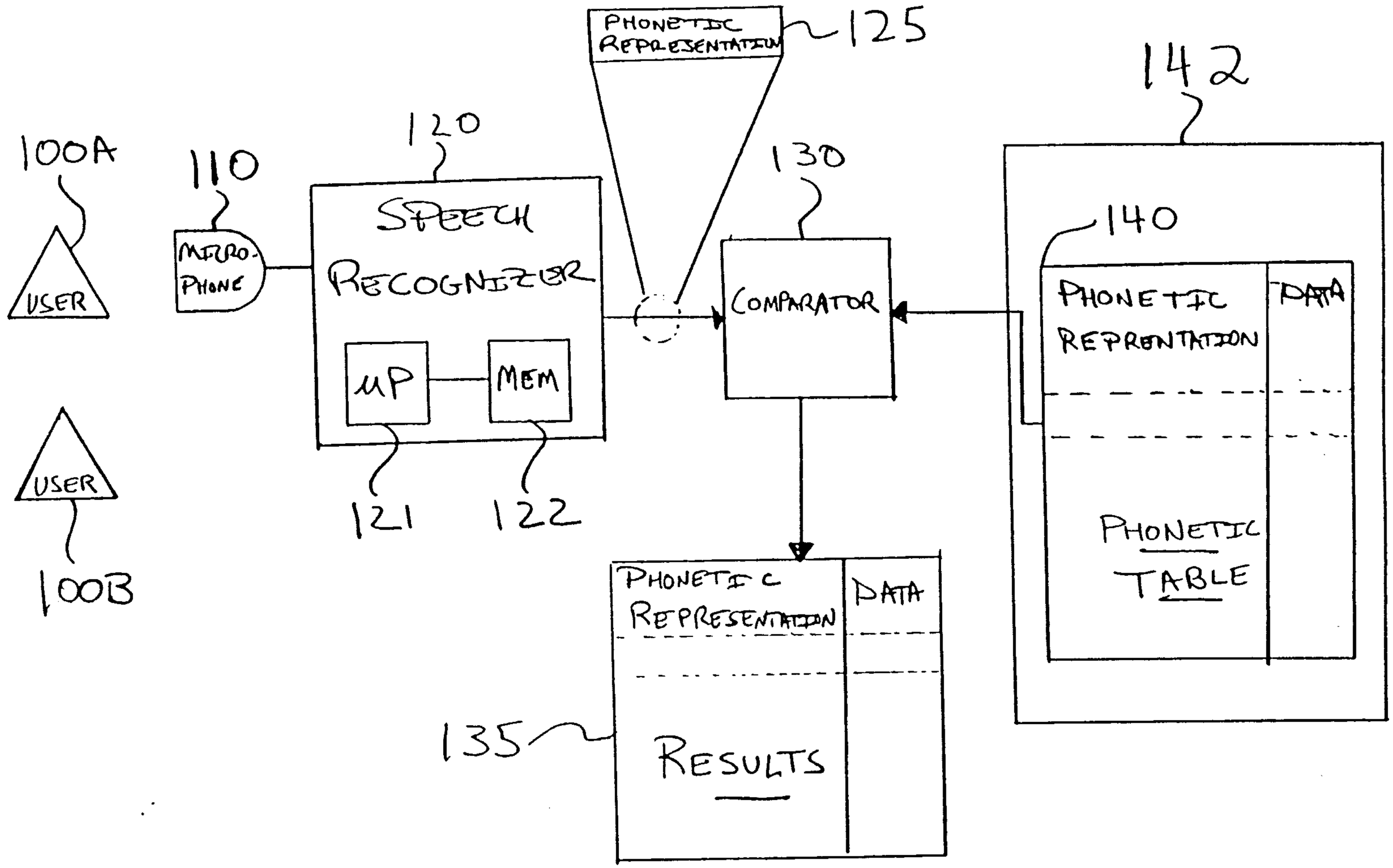


FIG. 2

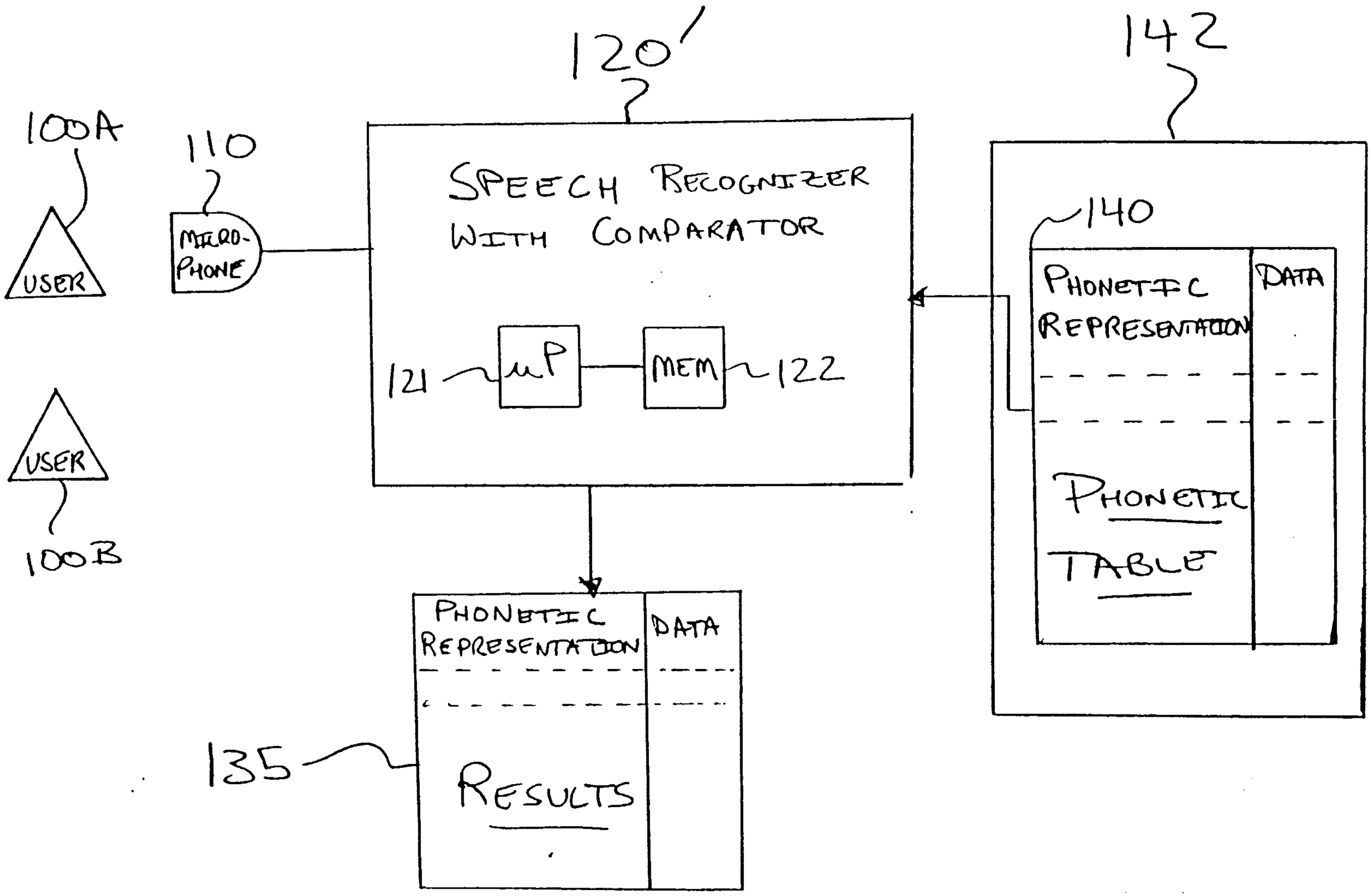


FIG. 3

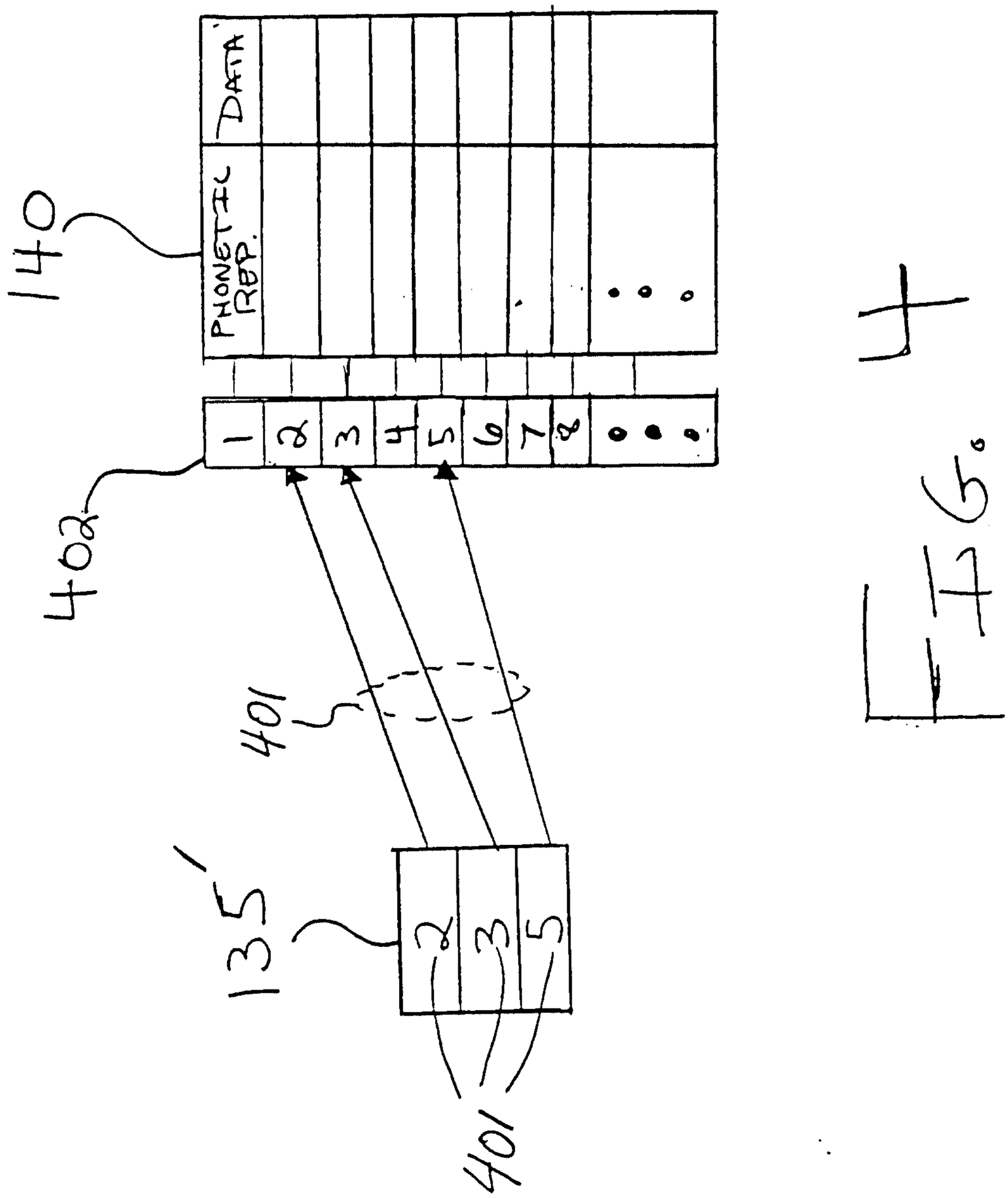


FIG 6.

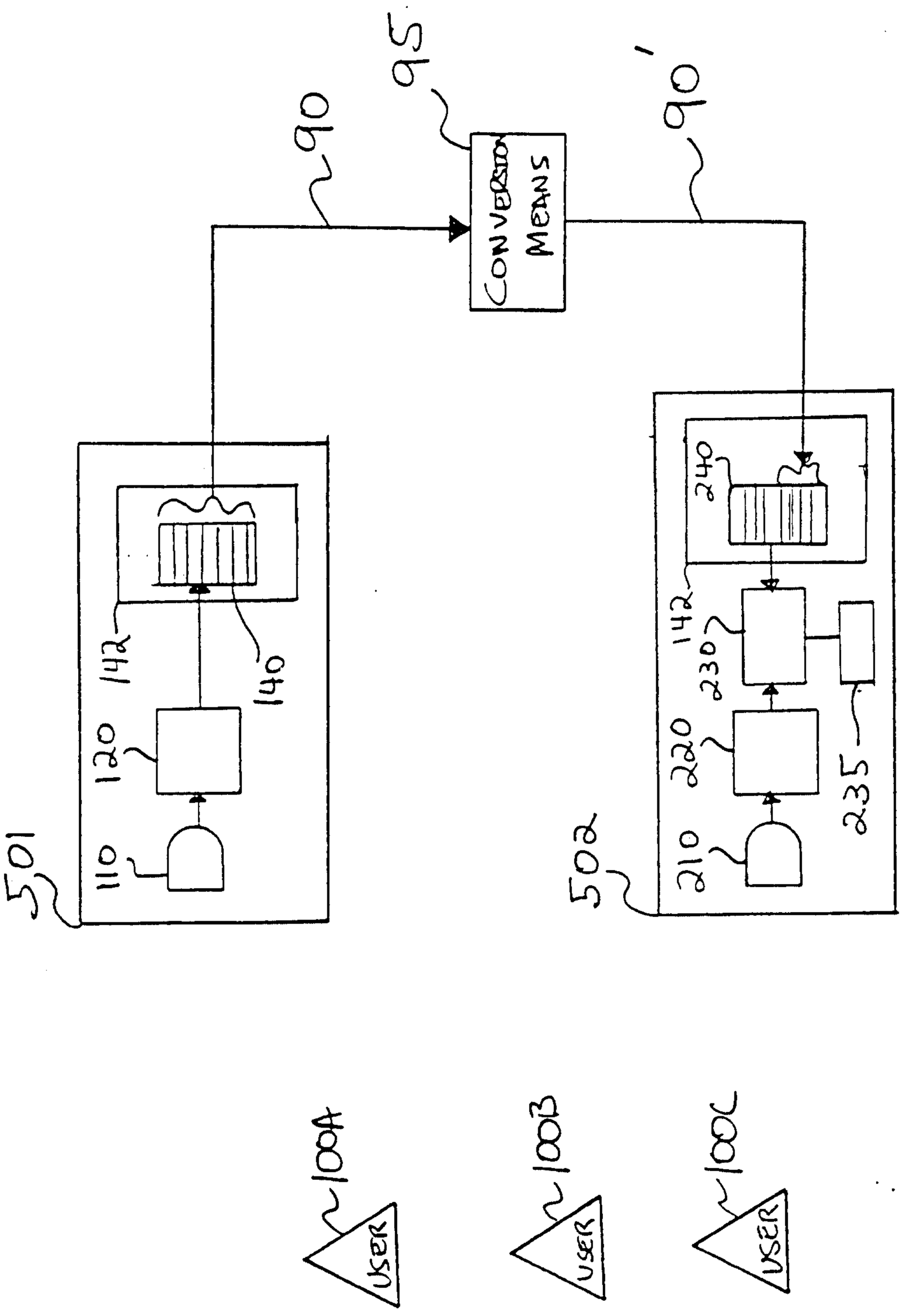

501
110
120
140
142
90
95
90'
CONVERSION MEANS
502
210
220
230
235
240
142
1004
1003
100c
USER
USER
USER

FIG. 5

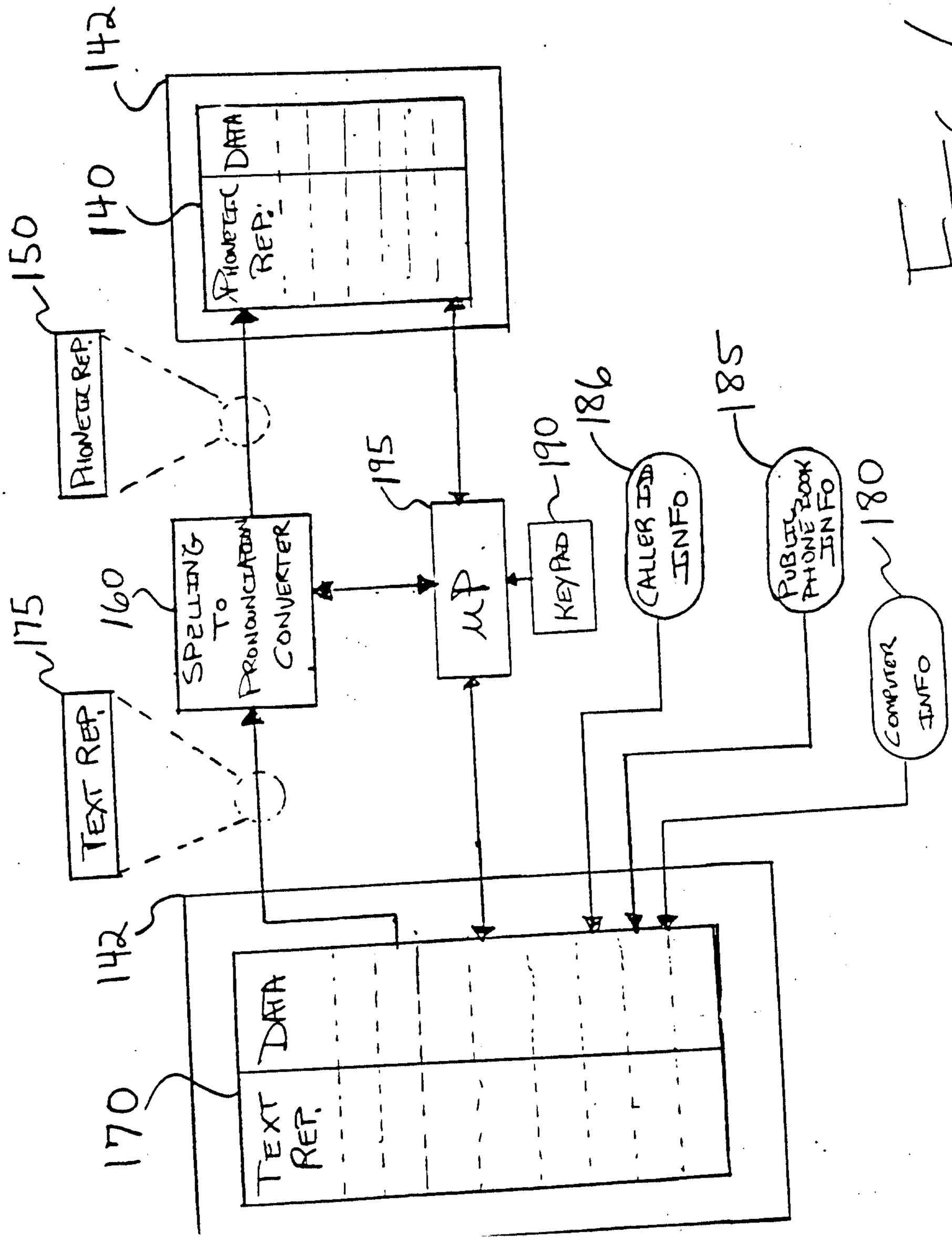

142
140
PIONEERED DATA REP.
150
PIONEERED REP.
175
TEXT REP.
160
SPELLING TO PRONUNCIATION CONVERTER
195
μP
190
KEYPAD
186
CALLER ID INFO
185
PUBLIC PHONE BOOK INFO
180
COMPUTER INFO
142
170
DATA
TEXT REP.

FIG. 6

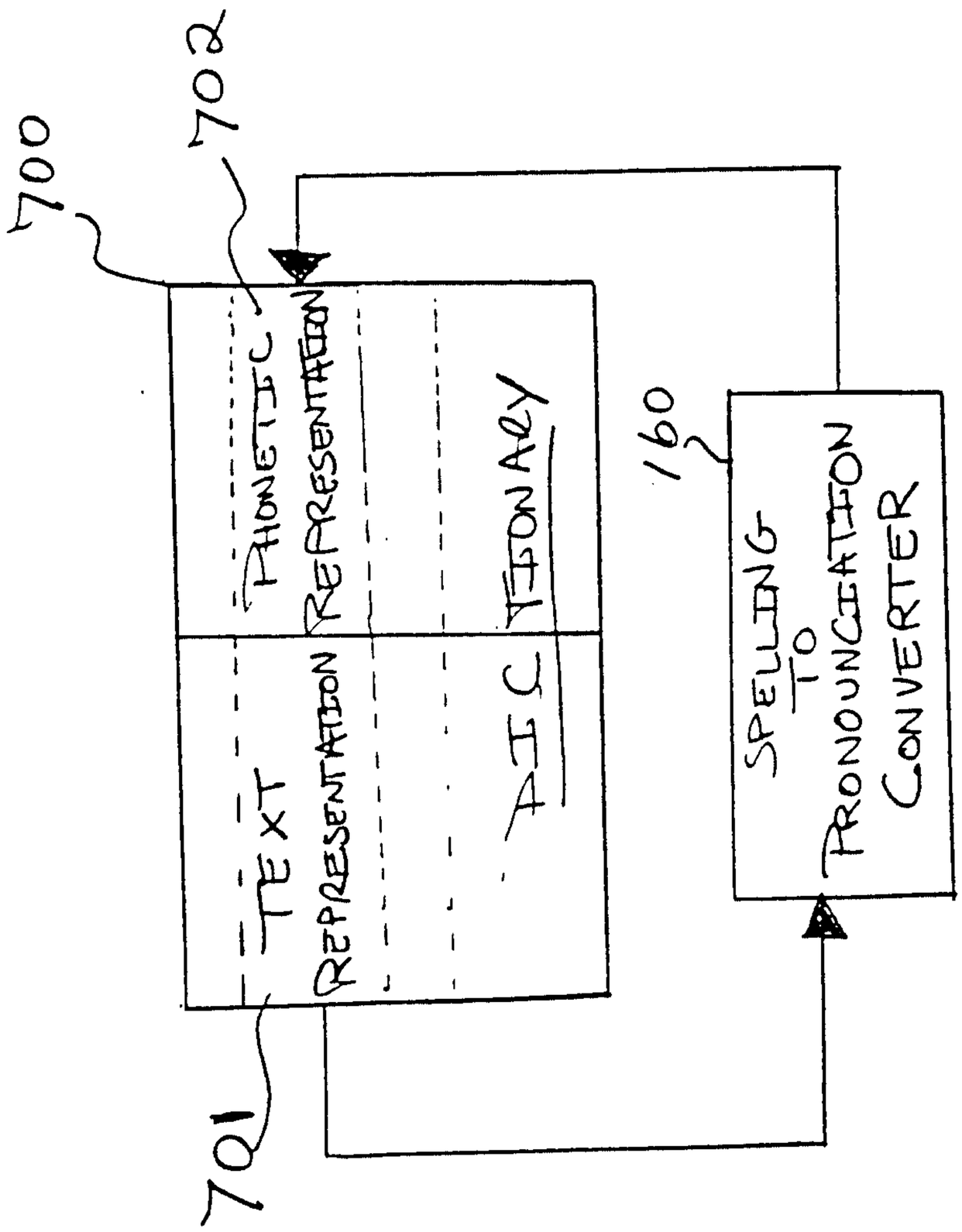

700
702
PHONETIC REPRESENTATION
TEXT REPRESENTATION
ACTION ARY
160
SPELLING TO PRONUNCIATION CONVERTER
Q1

FIG. 7

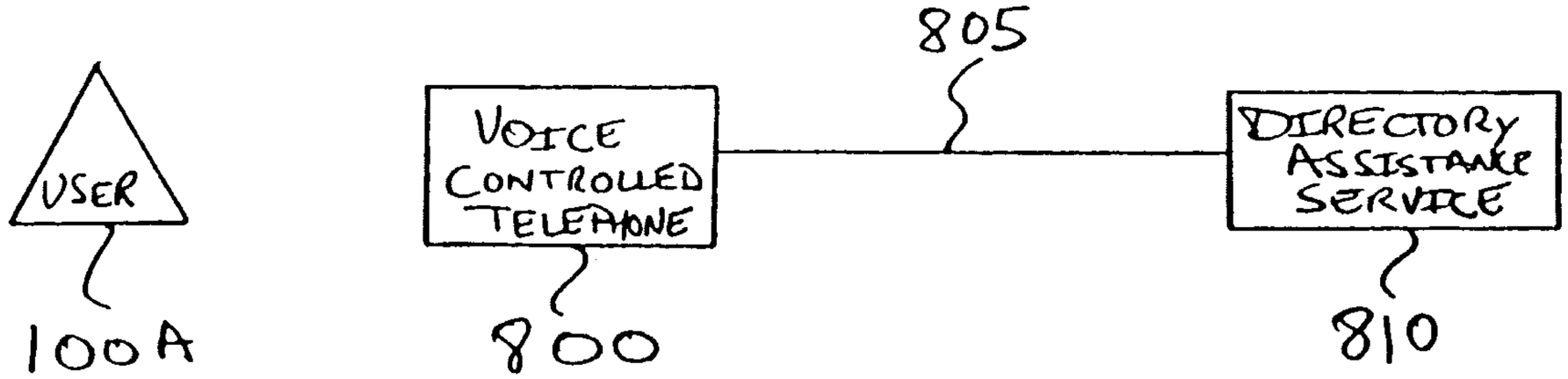
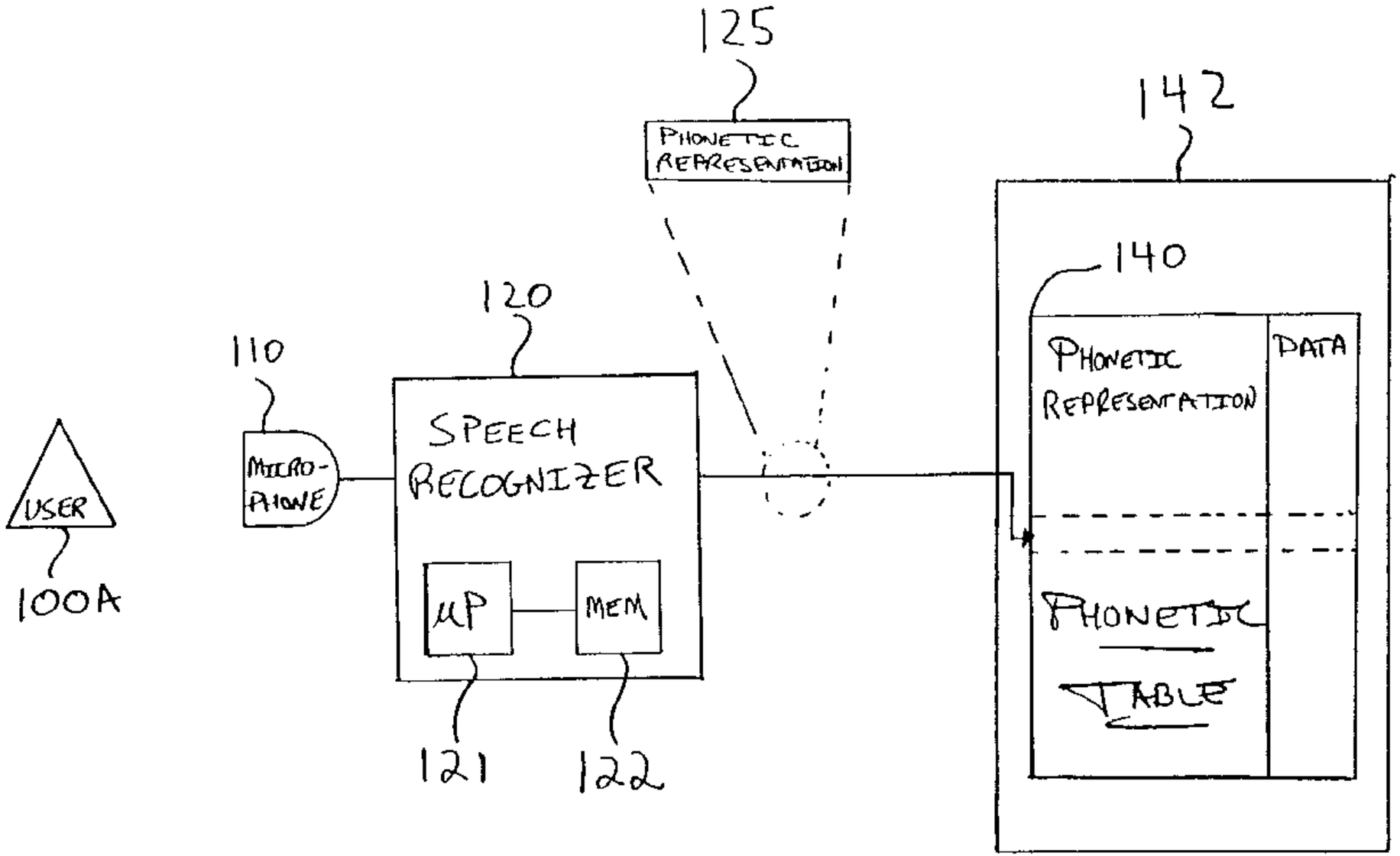


FIG. 8

CA 02308951 2000-05-19


125
PHONETIC REPRESENTATION
142
140
120
110
PHONETIC REPRESENTATION
DATA
SPEECH RECOGNIZER
MICRO-PHONE
USER
100A
μP
MEM
PHONETIC TABLE
121
122