PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G09B 1/00, 5/00, 19/00, 19/04, 19/06

A1

(11) International Publication Number: **WO 98/02862**

(43) International Publication Date: 22 January 1998 (22.01.98)

(21) International Application Number: PCT/IL97/00143

(22) International Filing Date: 4 May 1997 (04.05.97)

(30) Priority Data:
08/678,229 11 July 1996 (11.07.96) US

(71) Applicant *(for all designated States except US)*: DIGISPEECH (ISRAEL) LTD. [IL/IL]; Kehilat Saloniki Street 13, 69513 Tel Aviv (IL).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: SHPIRO, Zeev [IL/IL]; Beit Zuri Street 8, 69122 Tel Aviv (IL).

(74) Agents: COLB, Sanford, T. et al.; Sanford T. Colb & Co., P.O. Box 2273, 76122 Rehovot (IL).

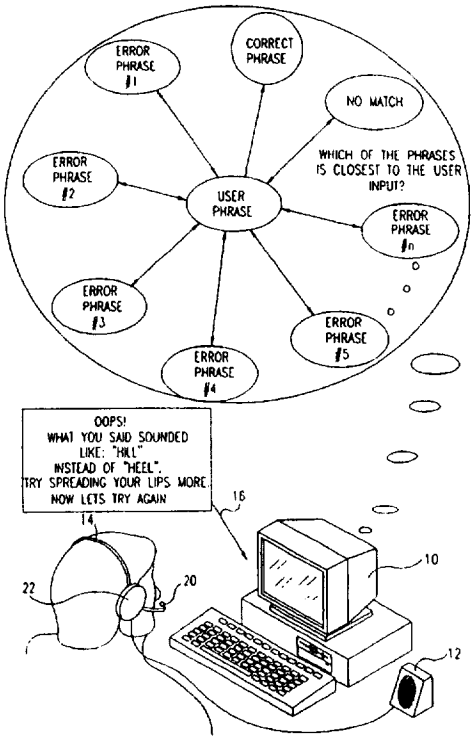
(81) Designated States: AL, AM, AT, AT (Utility model), AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, CZ (Utility model), DE, DE (Utility model), DK, DK (Utility model), EE, EE (Utility model), ES, FI, FI (Utility model), GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (Utility model), TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: APPARATUS FOR INTERACTIVE LANGUAGE TRAINING

(57) Abstract

This invention is an apparatus for interactive language training including a trigger generator for eliciting expected audio responses by a user; an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of the first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors; an audio response scorer which indicates the relationship between the expected audio response provided by the user and the reference expected responses; and a user feedback interface (12, 14, 16), which indicates to the user the pronunciation errors in the expected audio responses provided by the user. The present invention also discloses speech recognition apparatus including at least one data base containing speech elements of at least first and second languages, a receiver receiving spoken speech to be recognized, and a comparator comparing features of said spoken speech with a combination of features of said speech elements of at least first and second languages. It is appreciated that in certain cases a combination of the speech elements may include a single speech element. A method for speech recognition is also disclosed.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

APPARATUS FOR INTERACTIVE LANGUAGE TRAINING

FIELD OF THE INVENTION

The present invention relates to speech recognition systems having application inter alia in educational systems and more particularly to computerized systems providing phoneme based speech recognition and for teaching language.

BACKGROUND OF THE INVENTION

Computerized systems for teaching language are known. There is described in U.S. Patent 5,487,671, one inventor of which is the inventor of the present invention, a computerized system for teaching language which inter alia provides an indication of the relationship between a user's language to a reference.

A product having substantially the same feature is commercially available from The Learning Company under the trade name "Learn to Speak English".

Other commercially available products in this field are available from HyperGlot, Berlitz, Syracuse Language Systems Mindscape Global Language and Rosetta Stone Language Library.

Computerized systems for phoneme based speech recognition are known and commercially available. Examples of such systems include:

"IBM VoiceType, Simply Speaking for students, home users and small businesses", marketed by IBM;

"IBM VoiceType for professional and business use", marketed by IBM;

"Talk To Me", marketed by Dragon Systems, of Newton, Massachusetts, U.S.A.;

"ASR-1500", marketed by Lernout & Hauspie Speech Products N.V. of Leper, Belgium.

SUMMARY OF THE INVENTION

The present invention seeks to provide a further improved computerized system for teaching language which provides an indication to the user of the type of pronunciation error or errors that the user is making.

There is thus provided in accordance with a preferred embodiment of the present invention apparatus for interactive language training comprising:

a trigger generator for eliciting expected audio responses by a user;

an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of the first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors;

an audio response scorer which indicates the relationship between the expected audio response provided by the user and the reference expected responses; and

a user feedback interface which indicates to the user the pronunciation errors in the expected audio responses provided by the user.

Preferably, the user feedback interface also provides instruction to the user how to overcome the pronunciation errors.

In accordance with a preferred embodiment of the present invention, the user feedback interface indicates to the user each pronunciation error immediately following each expected audio response.

Preferably, the feedback interface provides audio and visual indications of the pronunciation errors.

In accordance with a preferred embodiment of the present invention, the audio specimen generator is operative such that the expected audio response is a repetition of the audio specimen.

Alternatively, the audio specimen generator is operative such that the expected audio response is other than a repetition of the audio specimen.

As a further alternative, the audio specimen generator is operative such that the expected audio response is an audio specimen which may be chosen from among more than one possible expected audio responses.

Preferably, the trigger generator comprises an audio specimen generator for playing audio specimens to a user.

Alternatively or additionally, the trigger generator comprises an visual trigger generator for providing a visual trigger output to a user.

Preferably, the expected audio response library comprises an expected audio response reference database.

In accordance with a preferred embodiment of the present invention, the expected audio response reference database comprises a multiplicity of templates and is speaker independent.

There is also provided in accordance with a preferred embodiment of the present invention a method for interactive language training comprising:

eliciting expected audio responses by a user;

providing an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of the first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors;

indicating the relationship between the expected audio response provided by the user and the reference expected responses; and

indicating to the user the pronunciation errors in the expected audio responses provided by the user.

Further in accordance with a preferred embodiment of the present invention the method also includes providing instruction to the user how to overcome the pronunciation errors.

Still further in accordance with a preferred embodiment of the present invention the method also includes indicating to the user each pronunciation error immediately following each expected audio response.

Additionally in accordance with a preferred embodiment of the present invention the method also includes providing audio and visual indications of said pronunciation errors to said user.

Further in accordance with a preferred embodiment of the present invention the method also includes the expected audio response is a repetition of said audio specimen.

Alternatively, the method also includes the expected audio response is other than a repetition of said audio specimen.

Additionally in accordance with a preferred embodiment of the present invention the expected audio response is an audio specimen which may be chosen from among more than one possible expected audio responses.

Furthermore in accordance with a preferred embodiment of the present invention the step of eliciting audio responses includes playing audio specimens to a user.

Still further in accordance with a preferred embodiment of the present invention the step of eliciting comprises providing a visual trigger output to a user.

There is also provided in accordance with a preferred embodiment of the present invention a speech recognition apparatus including at least one data base containing speech elements of at least first and second languages, a receiver, receiving spoken speech to be recognized, and a comparator, comparing features of the spoken speech with a combination of features of the speech elements of at least first and second languages. It is appreciated that in certain cases a combination of features of the speech elements may include the features of a single speech element. A feature of the speech element may include the speech element signal.

There is also provided in accordance with a preferred embodiment of the present invention a language teaching system including a trigger generator for eliciting expected audio responses by a user, a speech recognizer receiving the expected audio responses spoken by a user, the speech recognizer including at least one data base containing speech

elements of at least first and second languages, a receiver, receiving spoken speech to be recognized, and a comparator, comparing features of said spoken speech with a combination of features of said speech elements of at least first and second languages, and a user feedback interface which indicates to the user errors in the expected audio responses spoken by the user. It is appreciated that in certain cases a combination of the features of the speech elements may include the features of a single speech element. A feature of the speech element may include the speech element signal.

Further in accordance with a preferred embodiment of the present invention the speech elements include at least one of phonemes, diphones and transitions between phonemes.

Still further in accordance with a preferred embodiment of the present invention the language teaching system also includes a template generator operative to generate phrase templates.

Additionally in accordance with a preferred embodiment of the present invention the language teaching system also includes a feature extractor operative to extract features of spoken speech received by the receiver.

There is also provided in accordance with a preferred embodiment of the present invention a method for speech recognition including providing at least one data base containing speech elements of at least first and second languages, receiving spoken speech to be recognized, and comparing features of the spoken speech with a combination of features of the speech elements of at least first and second languages. It is appreciated that in certain cases a combination of the features of the speech elements may include the features of a single speech element. A feature of the speech element may include the speech element signal.

Further in accordance with a preferred embodiment of the present invention the spoken speech is spoken in a first language by a user who is a native speaker of a second language and wherein the at least one data base contains speech elements of both the first and the second languages.

Still further in accordance with a preferred embodiment of the present invention the at least first and second languages include different national languages.

Additionally in accordance with a preferred embodiment of the present invention the at least first and second languages include different dialects of a single national language.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be more fully understood and appreciated from the following detailed description, taken in conjunction with the drawings in which:

Fig. 1 is a generalized pictorial illustration of an interactive language teaching system constructed and operative in accordance with a preferred embodiment of the present invention;

Fig. 2 is a generalized functional block diagram of the operation of the system of Fig. 1 during language teaching;

Fig. 3 is a generalized functional block diagram of the operation of the system of Fig. 1 during audio reference library generation in accordance with one embodiment of the present invention;

Fig. 4 is a generalized functional block diagram of the operation of the system of Fig. 1 during audio reference library generation in accordance with another embodiment of the present invention;

Figs. 5A and 5B together constitute a generalized flow chart illustrating operation of the system during language teaching in accordance with the generalized functional block diagram of Fig. 2;

Figs. 6A, 6B and 6C together constitute a generalized flow chart illustrating one method of operation of the system during audio reference library generation for language teaching in accordance with the generalized functional block diagram of Fig. 3;

Fig. 7 is a generalized flow chart illustrating operation of the system during audio reference library generation for language teaching in accordance with the generalized functional block diagram of Fig. 4;

Fig. 8 is a simplified illustration of the creation of a phonetic template database of the type employed in Fig. 4;

Fig. 9 is a simplified illustration of a labeled speech waveform;

Fig. 10 illustrates the creation of a multiple language phonetic database in accordance with a preferred embodiment of the present invention;

Fig. 11 is an illustration of speech recognition employing phonemes; and

Fig. 12 is an illustration of speech recognition employing phonemes of various languages.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Reference is now made to Fig. 1, which is a generalized pictorial illustration of an interactive language teaching system constructed and operative in accordance with a preferred embodiment of the present invention and to Fig. 2, which is a generalized functional block diagram of the operation of the system of Fig. 1 during language teaching.

It is to be appreciated that the system of Fig. 1 has many similarities to the Computerized System for Teaching Speech described in U.S. Patent 5,487,671, the disclosure of which is hereby incorporated by reference.

As will be described in detail hereinbelow, the system of the present invention differs from that of U.S. Patent 5,487,671 in that it operates with reference expected responses each having different pronunciation errors and includes an audio response scorer which indicates the relationship between the expected audio response provided by the user and the reference expected responses having pronunciation errors.

The system of Figs. 1 and 2 incorporates speech recognition functionality in accordance with a preferred embodiment of the present invention.

The system of Figs. 1 and 2 is preferably based on a conventional personal computer 10, such as an IBM PC or compatible, using an Intel 80486 CPU running at 33 MHZ or higher, with at least 8MB of memory and running a DOS rev. 6.0 or above operating system. The personal computer 10 is preferably equipped with an auxiliary audio module 12. For example, a suitable audio module 12 is the Digispeech Plus audio adapter (DS311) manufactured by Digispeech, Inc. and distributed in the USA by DSP SOLUTIONS Inc., Mountain View, CA. A headset 14 is preferably associated with audio module 12.

Generally, the personal computer 10 and audio module 12 are supplied with suitable software so as to provide the following functionalities:

a trigger generator for eliciting expected audio responses by the user. The trigger generator preferably comprises an audio specimen generator for playing audio specimens to a user but may additionally or alternatively comprise a visual trigger generator for providing a visual trigger output to a user;

an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of the first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors. The second plurality of reference expected responses may include responses constructed from phonemes of various languages and may have application in speech recognition generally;

an audio response scorer which indicates the relationship between the expected audio response provided by the user and the reference expected responses; and

a user feedback interface which indicates to the user the pronunciation errors, if any, in the expected audio responses provided by the user.

The user feedback interface preferably provides audio feedback via the audio module 12 and headset 14. Additionally, as seen in Figs. 1 and 2, a display 16 is preferably provided to indicate pronunciation errors to the user in a visible manner, as illustrated, for example in Fig. 1.

In accordance with a preferred embodiment of the present invention, a total of six different databases are employed. For convenience and ease of understanding of the invention, the six databases are briefly described hereinbelow in the order in which they are created and used in the invention:

A. Interim Audio Specimens Database - This database is generated by recording a plurality of native speakers including a distribution of various geographical origins, various ages and both genders. The plurality of native speakers may include speakers speaking various different languages. Each speaker pronounces a plurality of predetermined phrases. For each of the plurality of predetermined phrases, each speaker pronounces the phrase correctly and also repeats the phrase incorrectly a few times, each time with a different one of a plurality of predetermined pronunciation errors. Preferably, this database includes plural recordings of each of the above pronounced phrases for each speaker, so as to provide an enhanced statistical base.

B. Expected Audio Response Reference Database - This is a database containing templates rather than recorded speech.

Various types of templates may be provided. One type of template, useful in word-based speech recognition, may be derived from Database A in a manner described hereinbelow. Another type of template, useful in phoneme-based speech recognition, comprises various combinations of features of speech elements which together represent a phrase.

Templates useful in word-based speech recognition may be derived from the Interim Audio Specimens Database A, by extracting the speech parameters of each of the pronounced phrases and combining them statistically so as to represent the pronunciation of the plurality of native speakers referred to hereinabove.

Thus each template represents a statistical combination of the pronunciations of a group of native speakers.

A single template may be produced to cover all of the native speakers whose pronunciation is recorded in the Interim Audio Specimens Database A, or plural templates may be used, when a single template does not accurately represent the entire range of native speakers. For example, one template may represent males and the other females. Alternatively or additionally separate templates may each contain phonemes of a different language.

In accordance with a preferred embodiment of the invention, the Expected Audio Response Reference Database B constitutes the expected audio response reference library, referred to above. This is a speaker independent database.

Various types of templates may be provided. One type of template useful in word-based speech recognition may be derived from Database A in a manner described hereinabove. Another type of template, useful in phoneme-based speech recognition, comprises various combinations of features of speech elements which together represent a phrase.

C. Phonetic Database - This is a commercially available database of speech parameters of phonemes for a given language. Such databases are available, for example, from AT & T,

Speech Systems Incorporated of Boulder, Colorado, U.S.A. and Lernout & Hauspie Speech Products N.V. of Leper, Belgium. Multiple phonetic databases, each containing speech parameters of phonemes of a different language may be provided and are collectively referred to as a Phonetic Database.

D. User Followup Database - This is a collection of recorded user responses.

E. Expected Audio Specimens Database - This is a collection of recordings of each of a single trained speaker pronouncing each of the plurality of phrases correctly.

F. Reference Audio Specimens Database - This is a collection of recordings of a single trained speaker pronouncing each of the plurality of phrases incorrectly a few times, each with a different one of a plurality of predetermined pronunciation errors.

Reference is now made to Fig. 2, which is a generalized functional block diagram of the operation of the system of Fig. 1 during language teaching.

Audio specimens stored in Expected Audio Specimen Database E are played to the user via the audio module 14 (Fig. 1) in order to elicit expected audio responses by the user. A microphone 20, normally part of headset 14, is employed to record the user's audio responses, which are stored in User Followup Database D. The audio specimens typically include spoken phrases. The phrases may include one or more words. Alternatively or additionally there may be provided a visual trigger generator for providing a visual trigger output to a user for eliciting expected audio responses by the user.

Spoken phrase parameters are extracted from the user's audio responses and are compared with reference phrase parameters to measure the likelihood of a match between the spoken phrase parameters of the user's audio response and the reference phrase parameters of a corresponding correct or incorrect phrase stored in the Expected Audio Response Reference Database B.

It is appreciated that the reference phrase parameters need not necessarily comprise words or combinations of words. Instead the reference phrase parameters may comprise various combinations of features of speech elements, particularly when phoneme based speech recognition is being carried out.

The result of the likelihood measurement is selection of a phrase which is closest to the user's audio response or an indication of failure to make any match. An audio and preferably also visible feedback indication is provided to the user, identifying the matched phrase and indicating whether it is correct or incorrect. Preferably, the user response may include a word, several words, a sentence or a number of sentences out of which one only or several phrases are matched during the teaching process. Additional teaching information as how to overcome indicated errors is preferably also provided in an audio-visual manner. Headphones 22, preferably forming part of headset 14 (Fig. 1) and display 16 are preferably employed for this purpose.

Reference is now made to Fig. 3, which is a generalized functional block diagram of the operation of the system of Fig. 1 during generation of the Expected Audio Response Reference Database B in accordance with one embodiment of the present invention. Here, a microphone 30, is used to record phrases spoken by a plurality of native speakers, including a distribution of various geographical origins, various ages and both genders.

Each speaker pronounces a plurality of predetermined phrases. For each of the plurality of predetermined phrases, each speaker pronounces the phrase correctly and also repeats the phrase incorrectly a few times, each time with a different one of a plurality of predetermined pronunciation errors. The recordings are retained in the Interim Audio Specimens database A. Preferably, this database includes plural recordings of each of the above pronounced phrases for each speaker, so as to provide an enhanced statistical base.

When word-based speech recognition is provided, spoken phrase parameters are extracted and merged with phrase parameters already stored in the Expected Audio Response Reference Database B to build up the Expected Audio Response Reference Database B. This database contains a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of the first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors.

It may be appreciated that each phrase is recorded correctly N times by each of a plurality of M speakers. It is additionally recorded N times by each of M speakers in L different forms each containing a different pronunciation error.

Reference is now made to Fig. 4, which is a generalized functional block diagram of the operation of the system of Fig. 1 during audio reference library generation in accordance with another embodiment of the present invention. Here, the Expected Audio Response Reference Database B is computer generated by generating text and phonetic language files which are employed to produce phonetic language records. The phonetic language record is employed together with Phonetic Database C to generate phrase templates which together constitute the Expected Audio Response Reference Database B.

In the embodiment of Fig. 4, the phrase templates are typically not words or combinations of words but rather combinations of features of speech elements, such as phonemes, diphones and transitions between phonemes. In phoneme-based speech recognition, features of speech to be recognized are compared with these combinations in order to find a best match.

Reference is now made to Figs. 5A and 5B, which together constitute a generalized flow chart illustrating operation of the system during language teaching in accordance with the generalized functional block diagram of Fig. 2. Once the indicated initial preparations indicated in the flowchart are complete, and preferably after the voice type to be heard from Database E is selected, a lesson is selected and the user is provided an explanation of how to pronounce a selected sound. For each selected sound, a reference audio specimen taken from Reference Audio Specimens Database E is played for the user in order to elicit an expected audio response by the user.

The user's response is recorded and compared with reference expected responses contained in the Expected Audio Response Reference Database B, by the Student Response Specimen Recorder as described in US Patent No. 5,487,671, the disclosure of which is hereby incorporated by reference.

If the best match is to the correct response, positive feedback is provided to the user and the lesson progresses to the next audio specimen.

If the best match is to a reference expected response having a pronunciation error then appropriate feedback is provided to the user. This feedback preferably includes an explanation of the error and how to correct it as well as a playback of the reference expected response. In accordance with a preferred embodiment of the present invention, the mispronounced phrase is played to the user from the Reference Audio Specimens Database F.

A User Followup database D may be employed to play back the latest or earlier user responses for indicating user progress, to be included in the system feedback, or other purposes.

Reference is now made to Figs. 6A, 6B and 6C, which together constitute a generalized flow chart illustrating operation of the system during audio reference library generation for language teaching in accordance with the generalized functional block diagram of Fig. 3.

Once the initial preparations indicated in the flowchart are complete the trained speaker speaks the correct phrase and a plurality of incorrect phrases, whose pronunciation is similar to the correct phrase but for one or more errors in pronunciation to provide reference expected responses each having different pronunciation errors. Each such set of correct and incorrect phrases is recorded. In accordance with a preferred embodiment of the invention, the Interim Audio Specimens Database A contains the various recordings. Database A is employed, as described above with reference to Fig. 3, to produce the Expected Audio Response Reference Database B, Fig. 6C for use in word-based speech recognition.

Reference is now made to Fig. 7, which is a generalized flow chart illustrating operation of the system during audio reference library generation for language teaching in accordance with the generalized functional block diagram of Fig. 4. Here a computer is employed to enter plain text and a phonetic language and to convert the text to the indicated phonetic language. Using a Phonetic Database C of the type described above, a phrase template is generated. The phrase template is then stored in the Expected Audio Response reference Database B. This described process is carried out for each phrase template being

used by the system. It is appreciated that that the phrase templates are typically not words or combinations of words but rather combinations of features of speech elements, such as phonemes, diphones and transitions between phonemes. In phoneme-based speech recognition, features of speech to be recognized are compared with these combinations in order to find a best match.

Reference is now made to Figs. 8 and 9, which illustrate the creation of Phonetic Database C of the type employed in Figs. 4 and 7 in accordance with a preferred embodiment of the present invention. A database 50 of labeled speech, typically of the type illustrated, for example in Fig. 9, can be obtained from TIMIT Acoustic-Phonetic Continuous Speech Corpora, available from the Linguistic Data Consortium, the University of Pennsylvania, at e-mail address online-service@ldc.upenn.edu. A template builder 52, typically embodied in commercially available software, such as HTK (Hidden Markov Model Toolkit) available from Entropic Cambridge Research Laboratories, Ltd., e-mail address sales@entropic.com, operates on the database 50 and provides the Phonetic Database C. The technique of Fig. 8 is applicable to various languages.

Where the phonetic database 58 comprises phonemes from various languages, the Phonetic Database C is realized by combining plural phonetic databases 54, 56, as illustrated in Fig. 10. It is a particular feature of the invention that phonetic databases 54 and 56, including phonemes of a language being learned or spoken as well as the native language of the user, may thus be combined to provide enhanced speech recognition.

Reference is now made to Fig. 11, which is an illustration of speech recognition employing phonemes. In the illustrated example the expected word is "tomato". A net of expected alternative pronunciations is created. Here, the speaker can pronounce the first "o" as "O", "OW" or "U", the "O" pronunciation being considered to be correct.

Similarly the user may pronounce the "a" as "A" or "EY", the "EY" pronunciation being considered to be correct.

Fig. 11 is characterized that all of the phonemes being used for speech recognition belong to a single language.

Reference is now made to Fig. 12, which is an illustration of speech recognition employing phonemes of various languages. The present example is designed for recognizing English spoken by a native Japanese speaker. Here the expected word is "Los" as in "Los Angeles". It is seen that here, the speaker can pronounce the "L" as "L" (circled "L"), an English "R" (circled "R") or a Japanese "R" (boxed "R").

Fig. 12 is characterized that not all of the phonemes being used for speech recognition belong to a single language. In the example of Fig. 12, some of the phonemes are English language phonemes (circled letters) and some of the phonemes are Japanese language phonemes (boxed letters).

It may thus be appreciated that when using the speech recognition technique of Fig. 12 for language teaching, the native Japanese characteristic mispronunciations are recognized by the system and the necessary teaching feedback will be provided to the user. When the speech recognition technique of Fig. 12 is used for other speech recognition applications, it enables English spoken by native Japanese speakers whose English pronunciation is not perfect to be recognized.

It will be appreciated by persons skilled in the art that the present invention is not limited by what has been particularly shown and described hereinabove. Rather the scope of the present invention includes both combinations and subcombinations of the various features and elements described hereinabove as well as obvious variations and extensions thereof.

CLAIMS

I claim:

1. Apparatus for interactive language training comprising:

a trigger generator for eliciting expected audio responses by a user;

an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of said first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors;

an audio response scorer which indicates the relationship between the expected audio response provided by the user and the reference expected responses; and

a user feedback interface which indicates to the user the pronunciation errors in the expected audio responses provided by the user.

2. Apparatus according to claim 1 and wherein said user feedback interface also provides instruction to the user how to overcome the pronunciation errors.

3. Apparatus according to claim 1 and wherein said user feedback interface indicates to the user each pronunciation error immediately following each expected audio response.

4. Apparatus according to claim 1 and wherein said feedback interface provides audio and visual indications of said pronunciation errors.

5. Apparatus according to claim 1 and wherein said audio specimen generator is operative such that the expected audio response is a repetition of said audio specimen.

6. Apparatus according to claim 1 and wherein said audio specimen generator is operative such that the expected audio response is other than a repetition of said audio specimen.

7. Apparatus according to claim 1 and wherein said audio specimen generator is operative such that the expected audio response is an audio specimen which may be chosen from among more than one possible expected audio responses.

8. Apparatus according to claim 1 and wherein said trigger generator comprises an audio specimen generator for playing audio specimens to a user.

9. Apparatus according to claim 1 and wherein said trigger generator comprises an visual trigger generator for providing a visual trigger output to a user.

10. Apparatus according to claim 1 and wherein said expected audio response library comprises an expected audio response reference database.

11. Apparatus according to claim 10 and wherein said expected audio response reference database comprises a multiplicity of templates.

12. Apparatus according to claim 10 and wherein said expected audio response reference database is speaker independent.

13. Apparatus according to claim 11 and wherein said expected audio response reference database is speaker independent.

14. A method for interactive language training comprising:

eliciting expected audio responses by a user;

providing an expected audio response reference library containing a multiplicity of reference expected responses, the multiplicity of reference expected responses including a first plurality of reference expected responses having acceptable pronunciation and for each of said first plurality of reference expected responses having acceptable pronunciation, a second plurality of reference expected responses each having different pronunciation errors;

indicating the relationship between the expected audio response provided by the user and the reference expected responses; and

indicating to the user the pronunciation errors in the expected audio responses provided by the user.

15. A method according to claim 14 and also comprising providing instruction to the user how to overcome the pronunciation errors.

16. A method according to claim 14 and also comprising indicating to the user each pronunciation error immediately following each expected audio response.

17. A method according to claim 14 and also comprising providing audio and visual indications of said pronunciation errors to said user.

18. A method according to claim 14 and wherein said expected audio response is a repetition of said audio specimen.

19. A method according to claim 14 and wherein said expected audio response is other than a repetition of said audio specimen.

20. A method according to claim 14 and wherein said expected audio response is an audio specimen which may be chosen from among more than one possible expected audio responses.

21. A method according to claim 14 and wherein said step of eliciting audio responses includes playing audio specimens to a user.

22. A method according to claim 14 and wherein said step of eliciting comprises providing a visual trigger output to a user.

23. Speech recognition apparatus comprising:

at least one data base containing speech elements of at least first and second languages;

a receiver, receiving spoken speech to be recognized; and

a comparator, comparing features of said spoken speech with a combination of features of said speech elements of at least first and second languages.

24. A language teaching system including:

a trigger generator for eliciting expected audio responses by a user;

a speech recognizer receiving the expected audio responses spoken by a user, the speech recognizer including:

at least one data base containing speech elements of at least first and second languages;

a receiver, receiving spoken speech to be recognized; and

a comparator, comparing features of said spoken speech with a combination of features of said speech elements of at least first and second languages; and

a user feedback interface which indicates to the user errors in the expected audio responses spoken by the user.

25. A language teaching system according to claim 23 and wherein said speech elements comprise at least one of phonemes, diphones and transitions between phonemes.

26. A language teaching system according to claim 23 and also comprising a template generator operative to generate phrase templates.

27. A language teaching system according to claim 23 and also comprising a feature extractor operative to extract features of spoken speech received by said receiver.

28. A method for speech recognition comprising:

providing at least one data base containing speech elements of at least first and second languages;

receiving spoken speech to be recognized; and

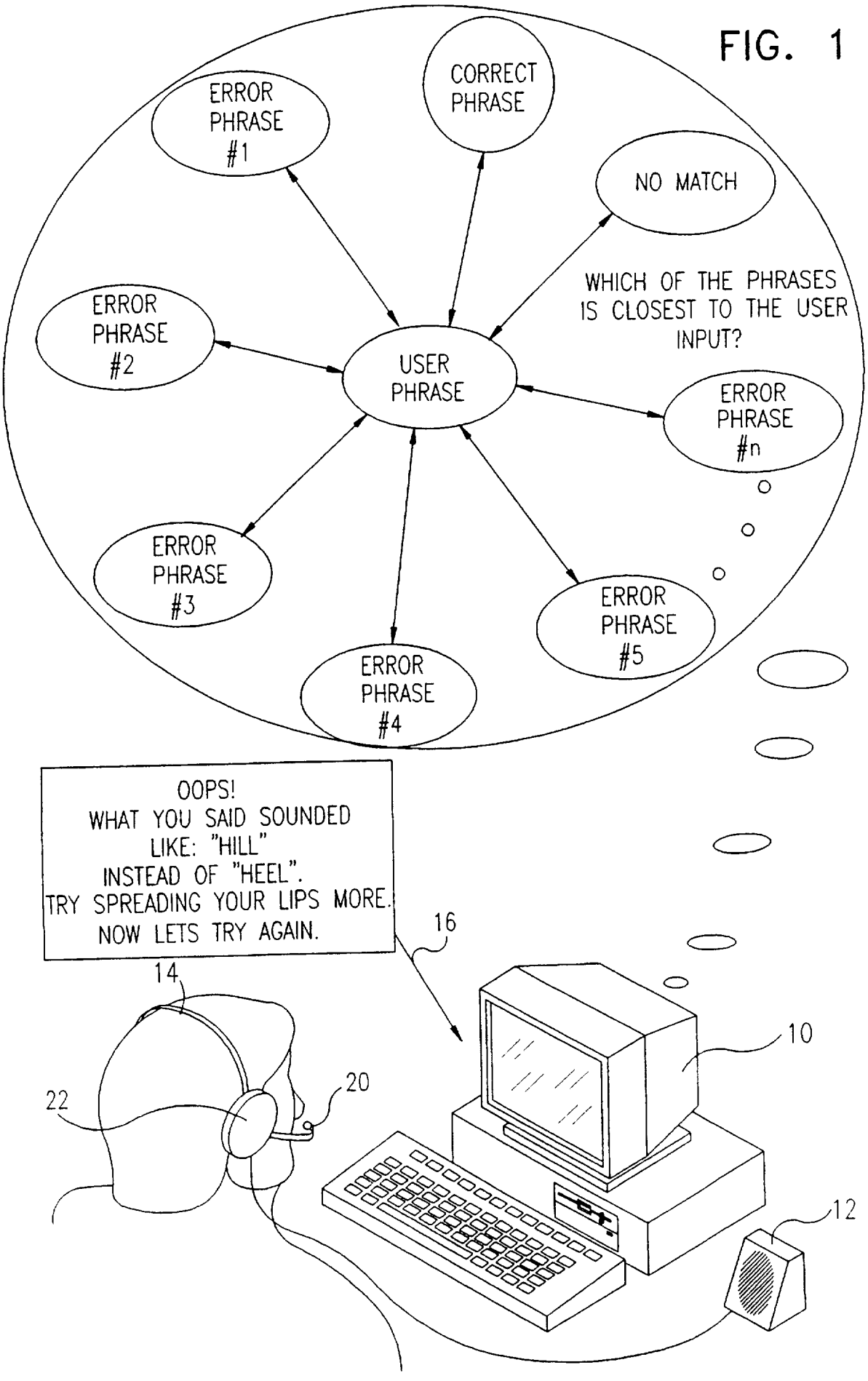
comparing features of said spoken speech with a combination of features of said speech elements of at least first and second languages.

29. A method for speech recognition according to claim 28 and wherein the spoken speech is spoken in a first language by a user who is a native speaker of a second language and wherein the at least one data base contains speech elements of both the first and the second languages.

30. A method according to claim 28 and wherein said at least first and second languages comprise different national languages.

31. A method according to claim 28 and wherein said at least first and second languages comprise different dialects of a single national language.

FIG. 1


ERROR PHRASE #1
CORRECT PHRASE
NO MATCH
WHICH OF THE PHRASES IS CLOSEST TO THE USER INPUT?
ERROR PHRASE #2
USER PHRASE
ERROR PHRASE #n
ERROR PHRASE #3
ERROR PHRASE #5
ERROR PHRASE #4
OOPS!
WHAT YOU SAID SOUNDED LIKE: "HILL"
INSTEAD OF "HEEL".
TRY SPREADING YOUR LIPS MORE.
NOW LETS TRY AGAIN.
16
14
10
22
20
12

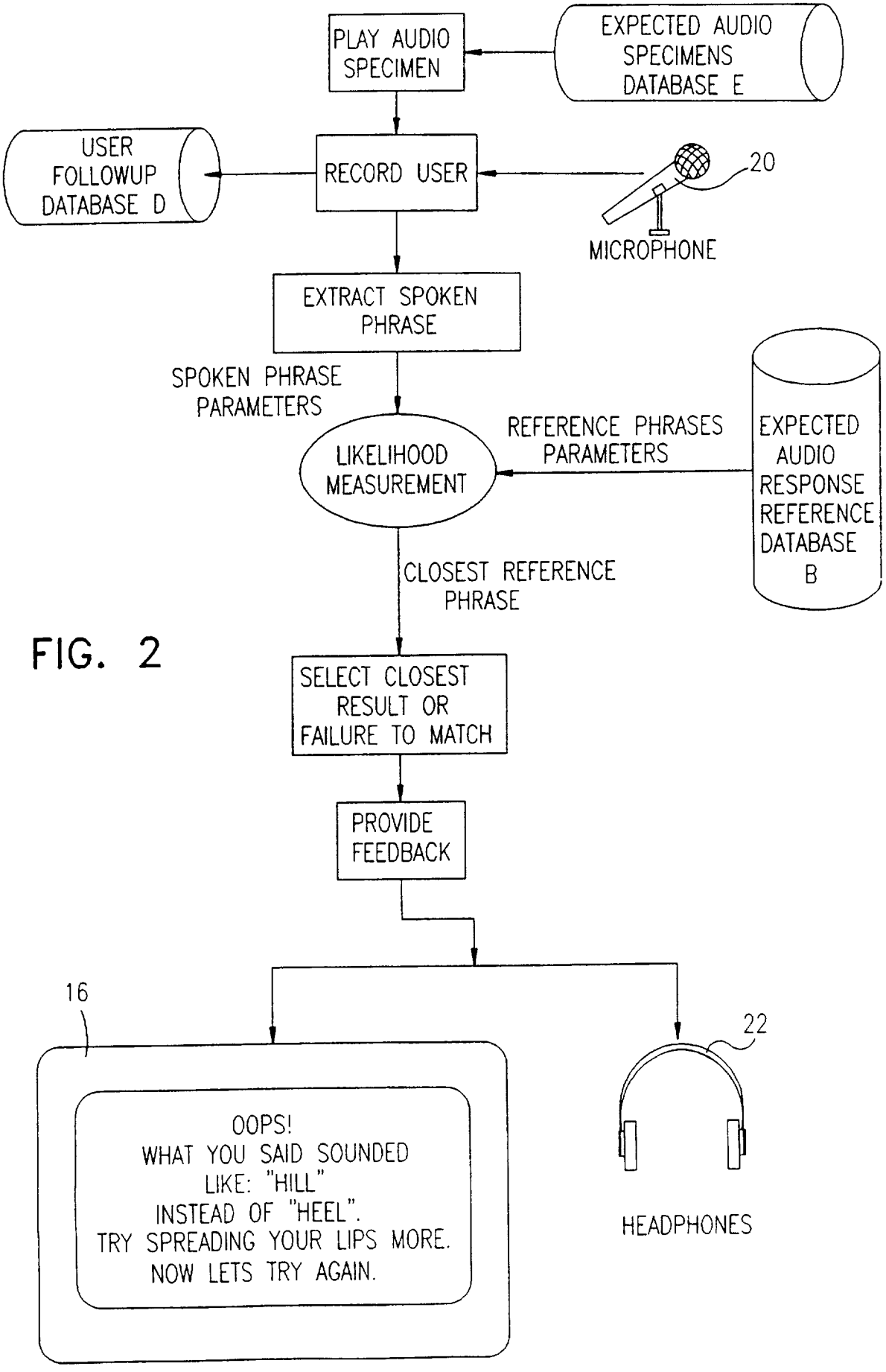

PLAY AUDIO SPECIMEN
EXPECTED AUDIO SPECIMENS DATABASE E
USER FOLLOWUP DATABASE D
RECORD USER
20
MICROPHONE
EXTRACT SPOKEN PHRASE
SPOKEN PHRASE PARAMETERS
LIKELIHOOD MEASUREMENT
REFERENCE PHRASES PARAMETERS
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
CLOSEST REFERENCE PHRASE
SELECT CLOSEST RESULT OR FAILURE TO MATCH
PROVIDE FEEDBACK
16
22
OOPS!
WHAT YOU SAID SOUNDED
LIKE: "HILL"
INSTEAD OF "HEEL".
TRY SPREADING YOUR LIPS MORE.
NOW LETS TRY AGAIN.
HEADPHONES

FIG. 2

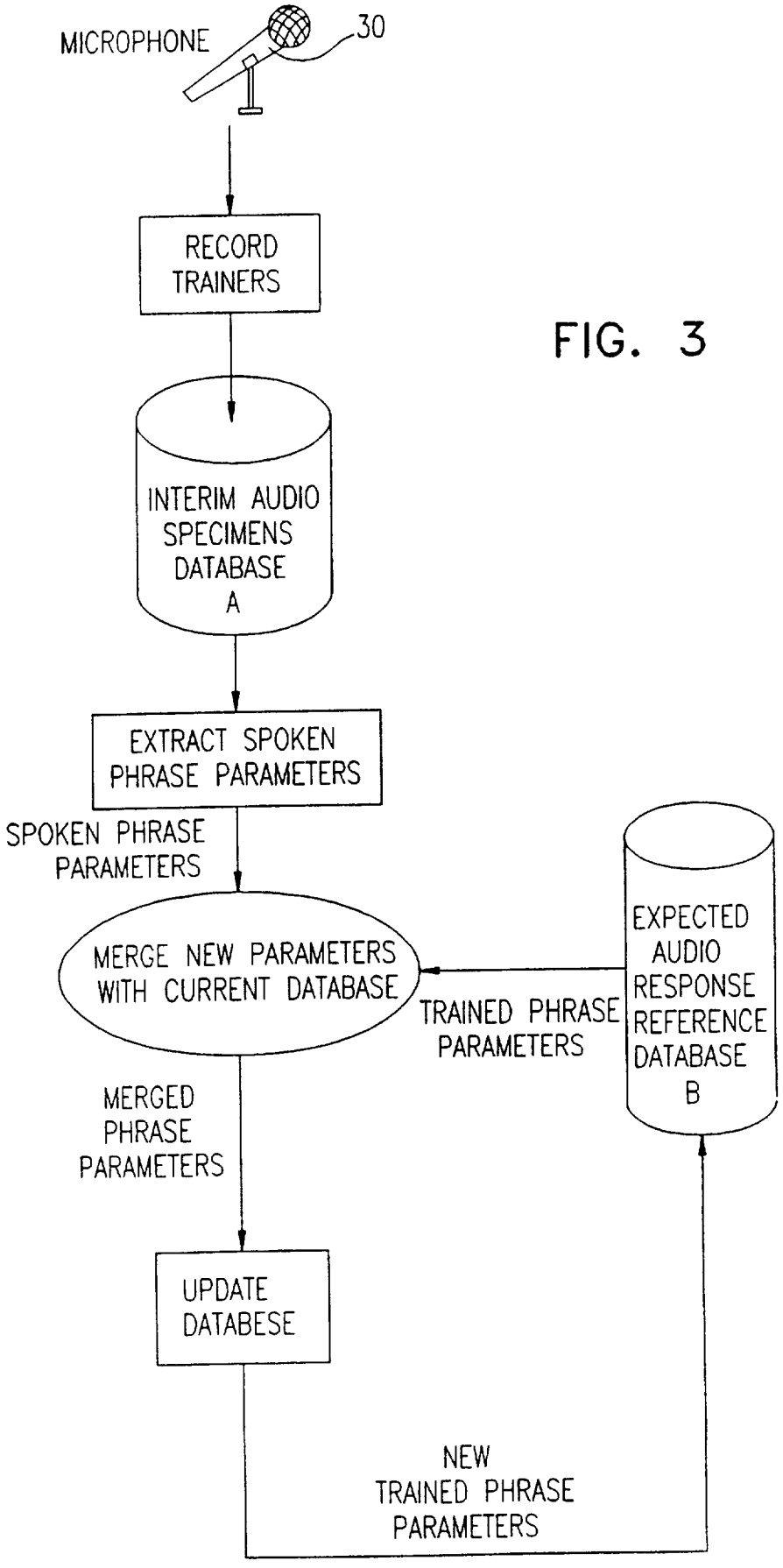

MICROPHONE
30
RECORD TRAINERS
INTERIM AUDIO SPECIMENS DATABASE A
EXTRACT SPOKEN PHRASE PARAMETERS
SPOKEN PHRASE PARAMETERS
MERGE NEW PARAMETERS WITH CURRENT DATABASE
TRAINED PHRASE PARAMETERS
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
MERGED PHRASE PARAMETERS
UPDATE DATABESE
NEW TRAINED PHRASE PARAMETERS

FIG. 3

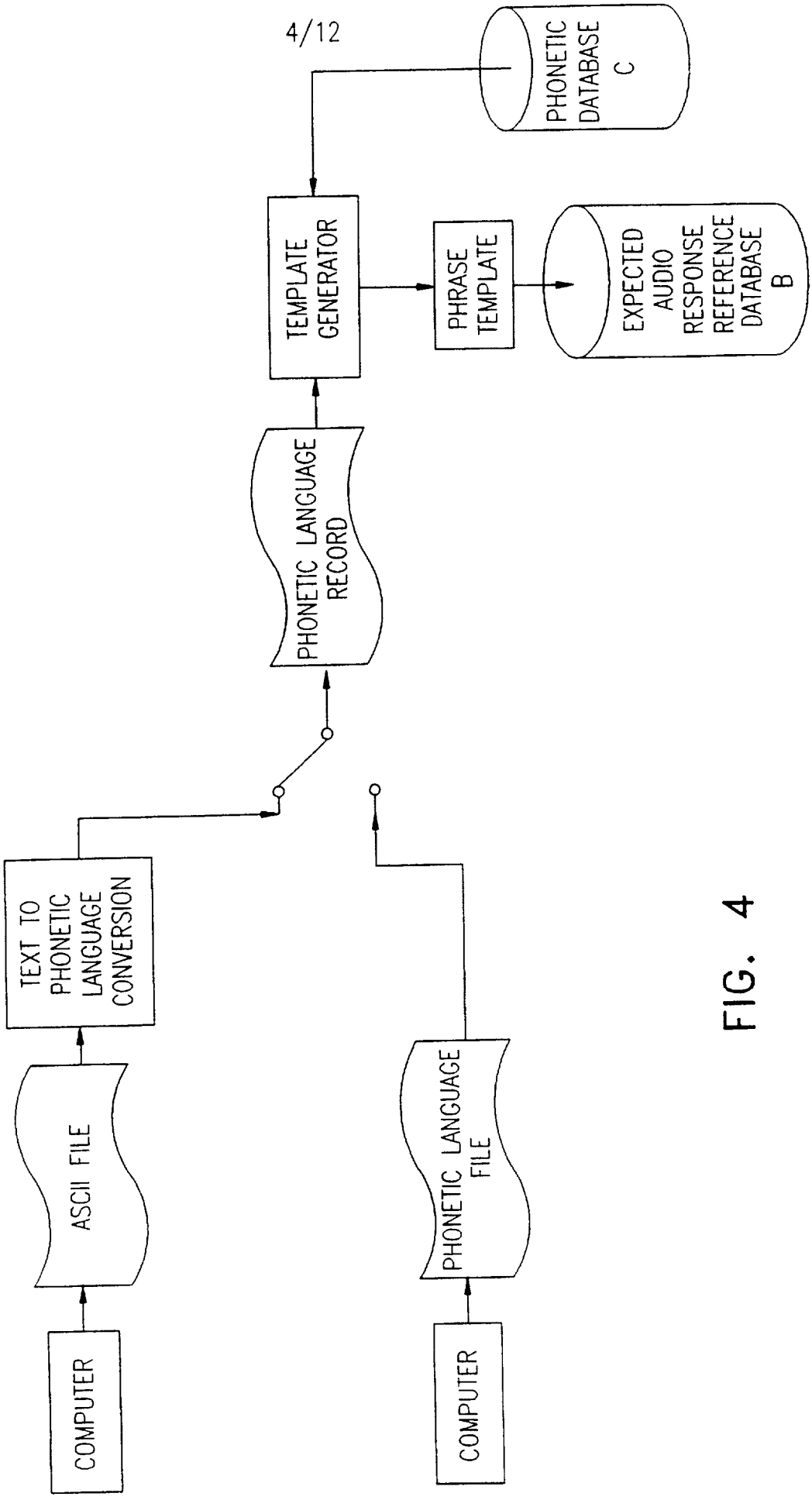

PHONETIC DATABASE C
TEMPLATE GENERATOR
PHRASE TEMPLATE
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
PHONETIC LANGUAGE RECORD
TEXT TO PHONETIC LANGUAGE CONVERSION
ASCII FILE
COMPUTER
PHONETIC LANGUAGE FILE
COMPUTER

FIG. 4

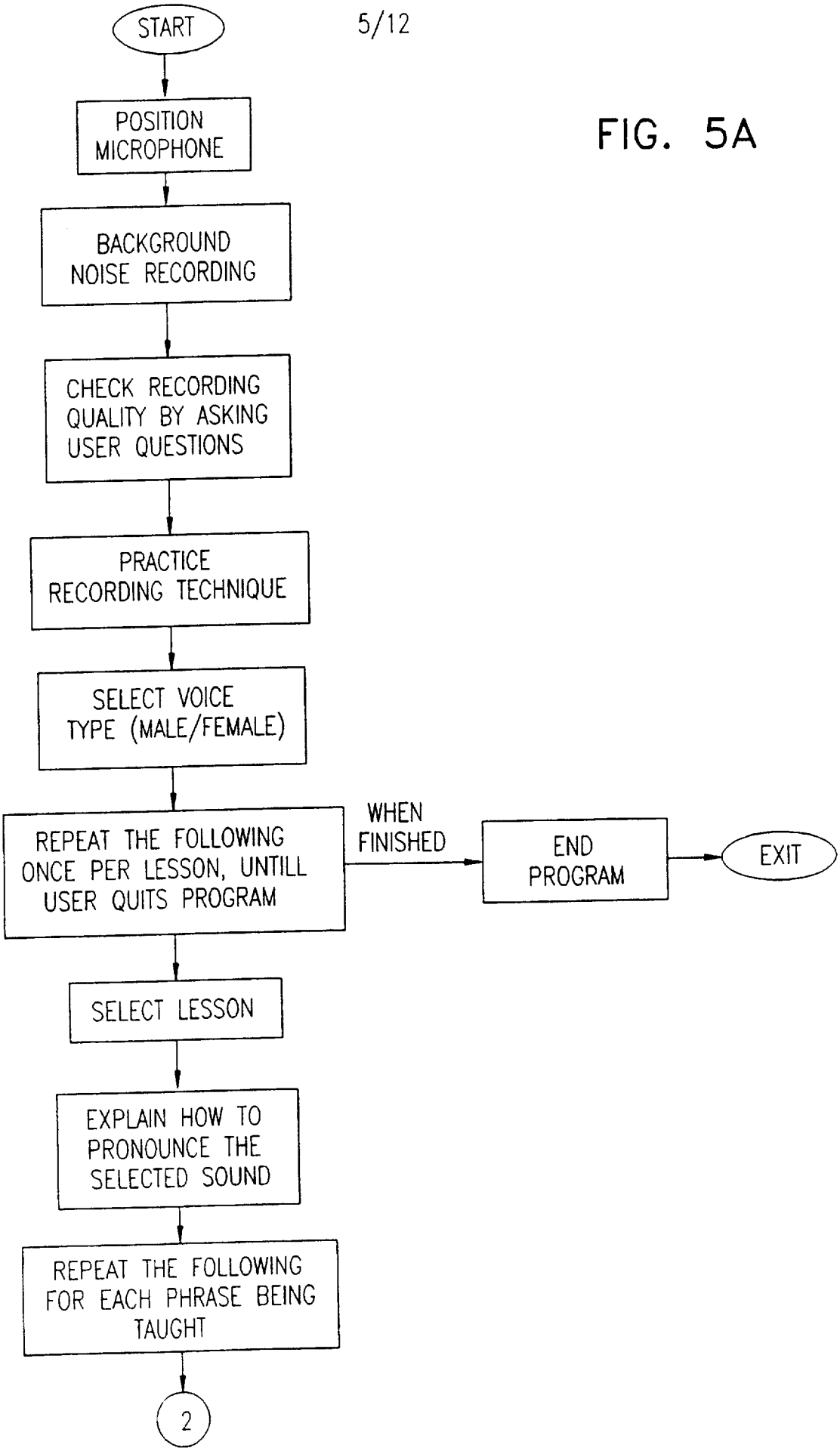

START
POSITION MICROPHONE
BACKGROUND NOISE RECORDING
CHECK RECORDING QUALITY BY ASKING USER QUESTIONS
PRACTICE RECORDING TECHNIQUE
SELECT VOICE TYPE (MALE/FEMALE)
REPEAT THE FOLLOWING ONCE PER LESSON, UNTILL USER QUITS PROGRAM
WHEN FINISHED
END PROGRAM
EXIT
SELECT LESSON
EXPLAIN HOW TO PRONOUNCE THE SELECTED SOUND
REPEAT THE FOLLOWING FOR EACH PHRASE BEING TAUGHT
2

FIG. 5A

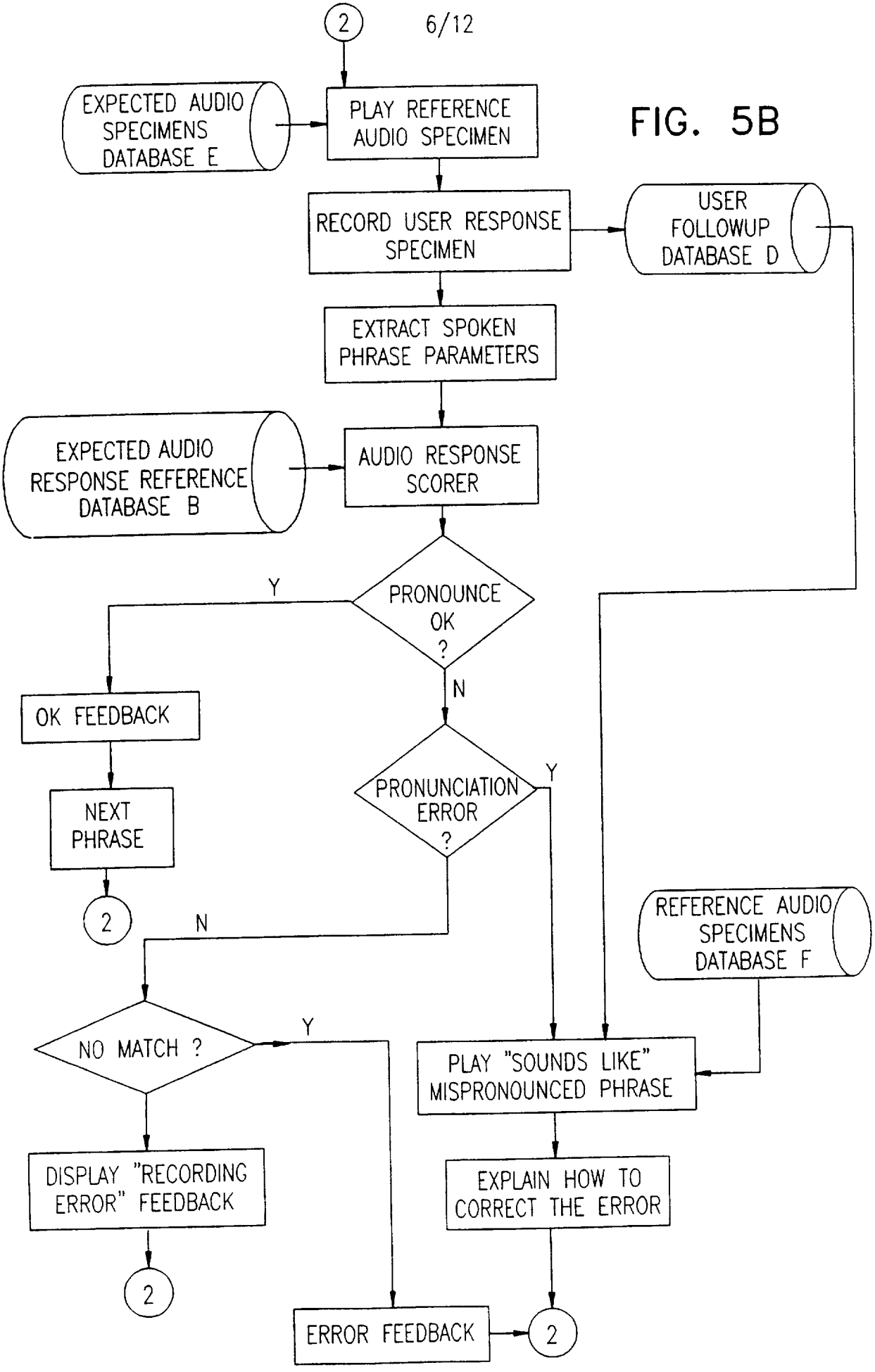
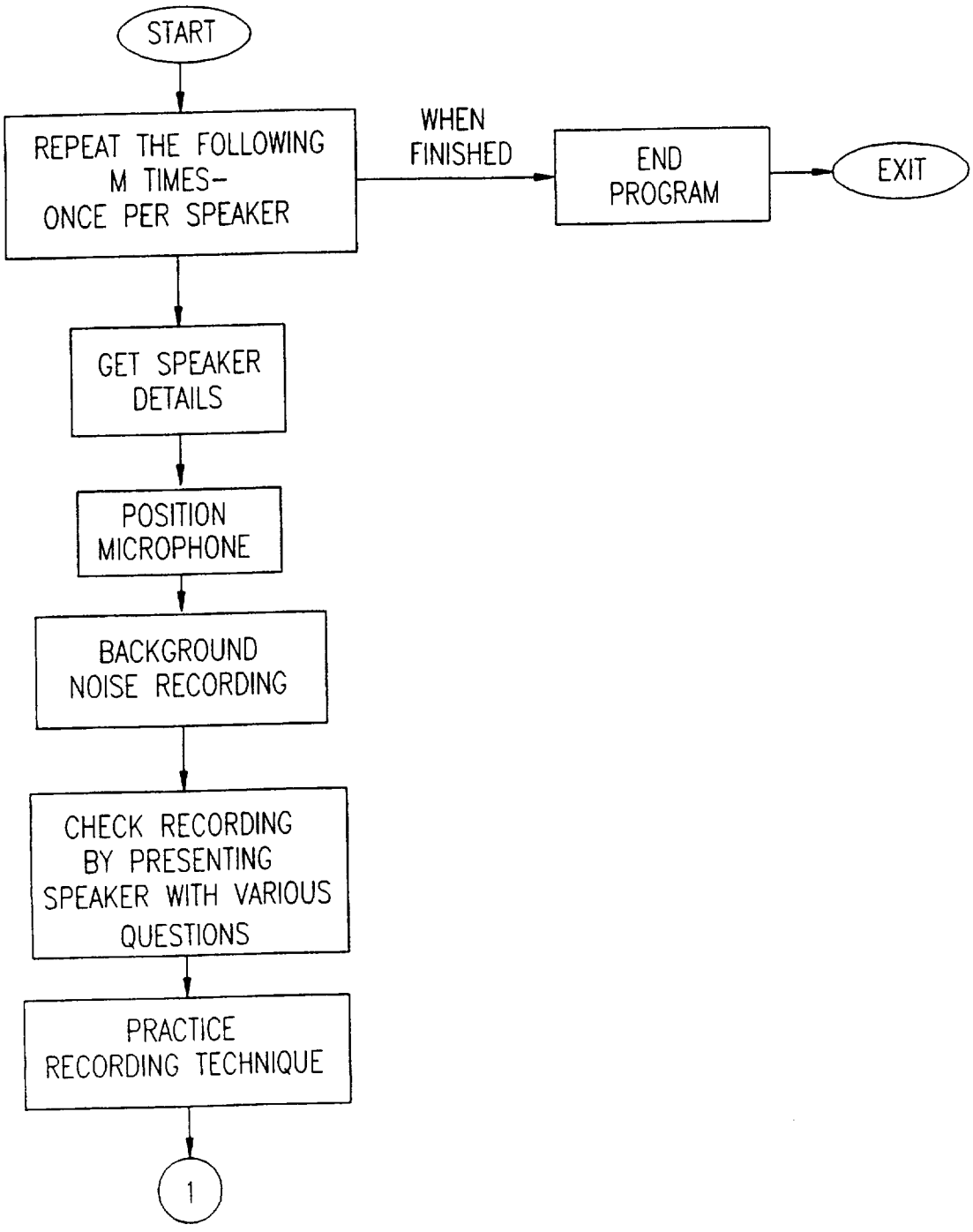

2
EXPECTED AUDIO SPECIMENS DATABASE E
PLAY REFERENCE AUDIO SPECIMEN
RECORD USER RESPONSE SPECIMEN
USER FOLLOWUP DATABASE D
EXTRACT SPOKEN PHRASE PARAMETERS
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
AUDIO RESPONSE SCORER
PRONOUNCE OK ?
Y
N
OK FEEDBACK
NEXT PHRASE
2
PRONUNCIATION ERROR ?
Y
N
REFERENCE AUDIO SPECIMENS DATABASE F
NO MATCH ?
Y
PLAY "SOUNDS LIKE" MISPRONOUNCED PHRASE
DISPLAY "RECORDING ERROR" FEEDBACK
EXPLAIN HOW TO CORRECT THE ERROR
2
ERROR FEEDBACK
2

FIG. 5B

FIG. 6A


START
REPEAT THE FOLLOWING M TIMES- ONCE PER SPEAKER
WHEN FINISHED
END PROGRAM
EXIT
GET SPEAKER DETAILS
POSITION MICROPHONE
BACKGROUND NOISE RECORDING
CHECK RECORDING BY PRESENTING SPEAKER WITH VARIOUS QUESTIONS
PRACTICE RECORDING TECHNIQUE
1

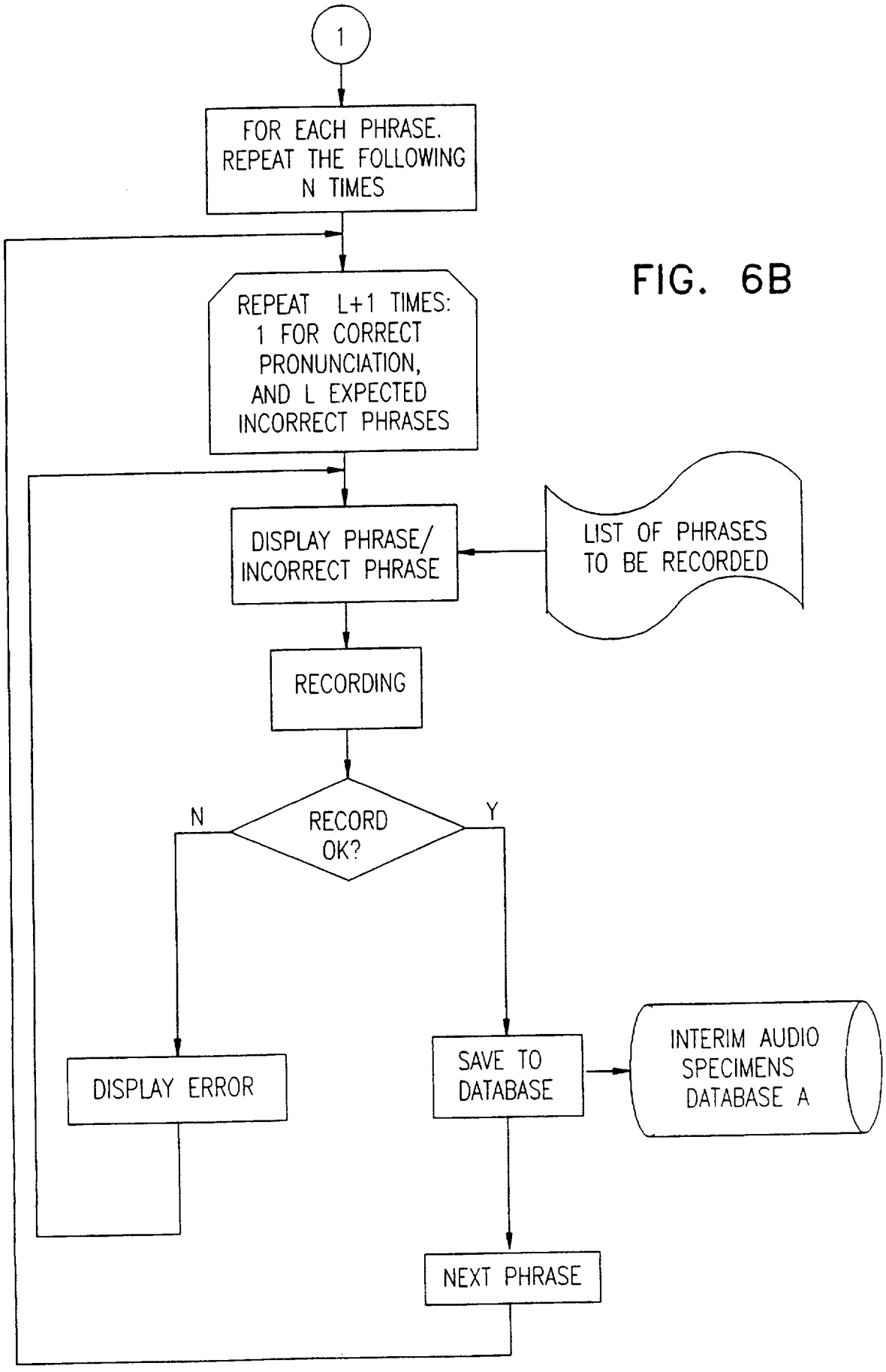

1
FOR EACH PHRASE. REPEAT THE FOLLOWING N TIMES
REPEAT L+1 TIMES: 1 FOR CORRECT PRONUNCIATION, AND L EXPECTED INCORRECT PHRASES
DISPLAY PHRASE/ INCORRECT PHRASE
LIST OF PHRASES TO BE RECORDED
RECORDING
RECORD OK?
N
Y
DISPLAY ERROR
SAVE TO DATABASE
INTERIM AUDIO SPECIMENS DATABASE A
NEXT PHRASE

FIG. 6B

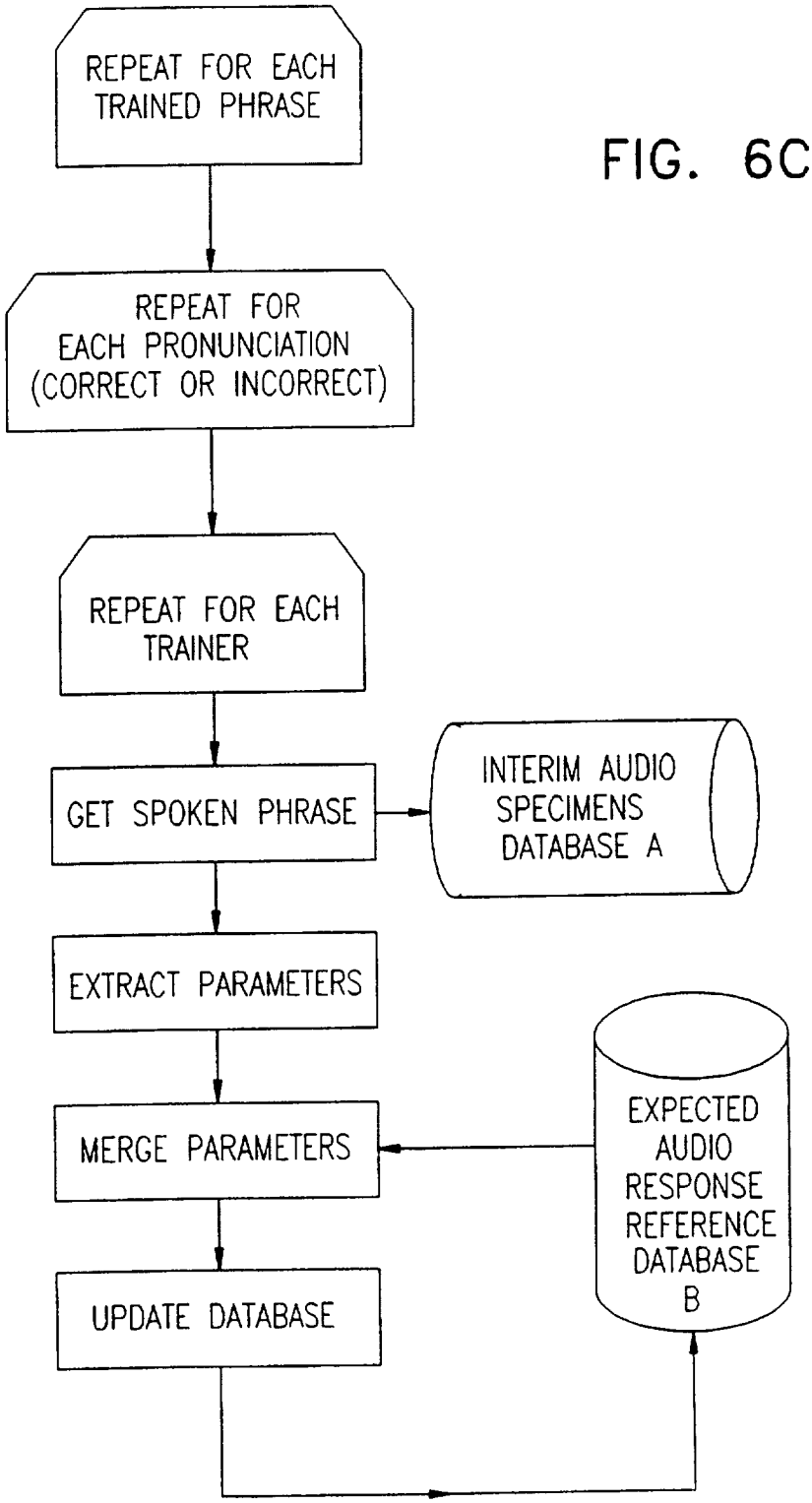

REPEAT FOR EACH TRAINED PHRASE
REPEAT FOR EACH PRONUNCIATION (CORRECT OR INCORRECT)
REPEAT FOR EACH TRAINER
GET SPOKEN PHRASE
INTERIM AUDIO SPECIMENS DATABASE A
EXTRACT PARAMETERS
MERGE PARAMETERS
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
UPDATE DATABASE

FIG. 6C

FIG. 7

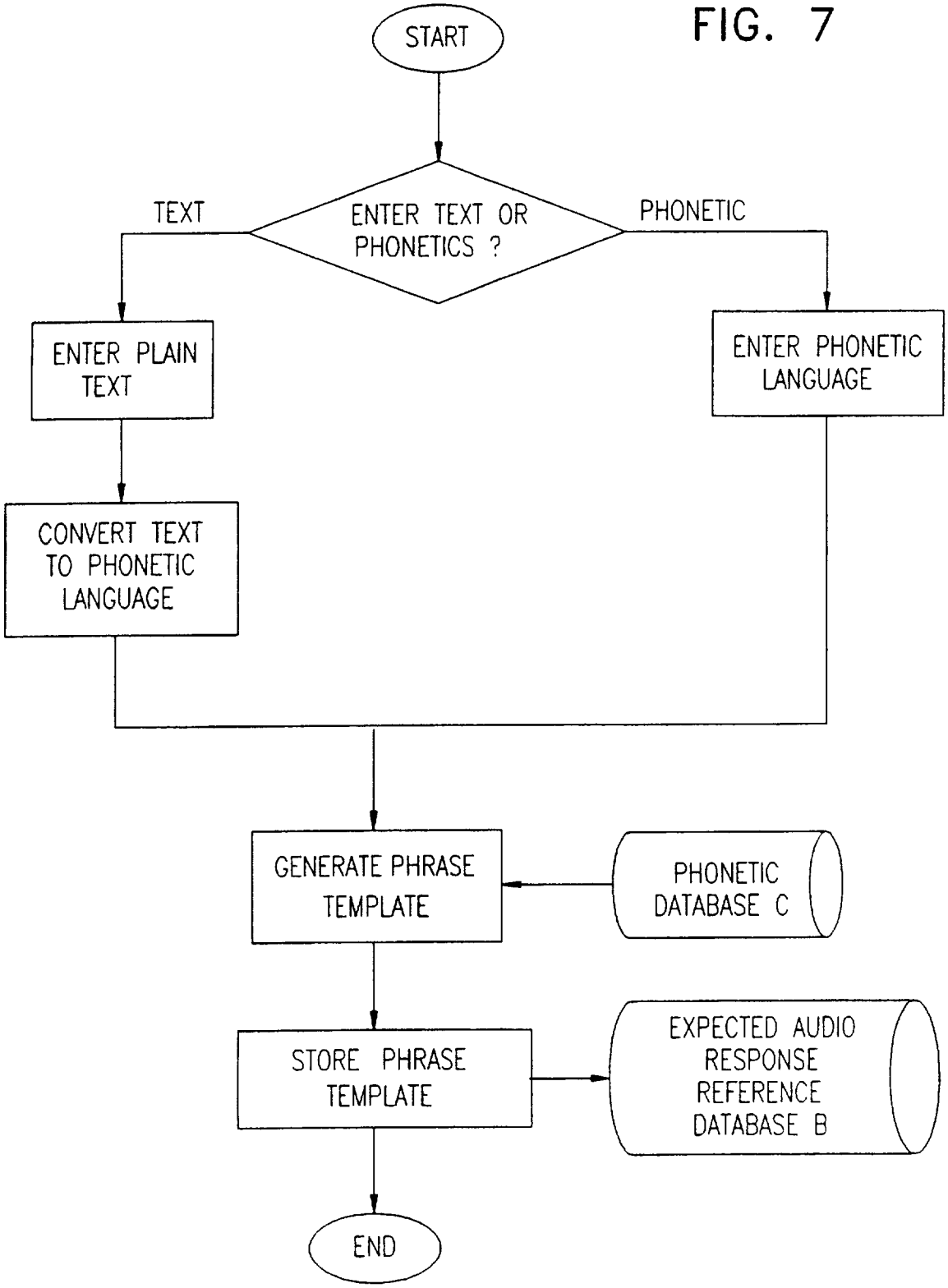
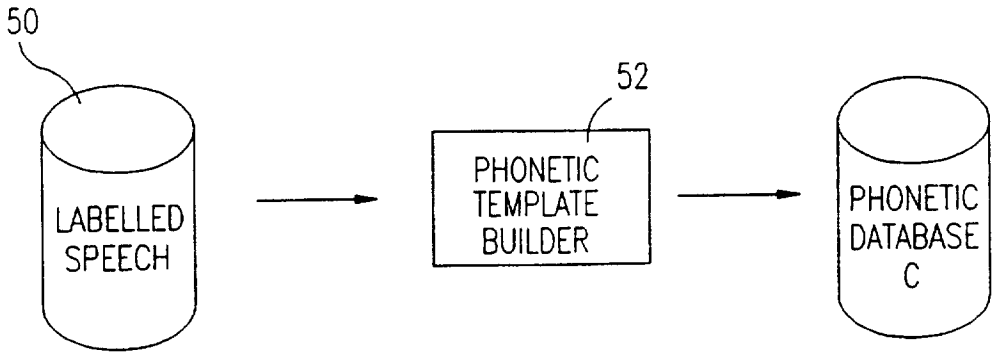

START
ENTER TEXT OR PHONETICS ?
TEXT
PHONETIC
ENTER PLAIN TEXT
ENTER PHONETIC LANGUAGE
CONVERT TEXT TO PHONETIC LANGUAGE
GENERATE PHRASE TEMPLATE
PHONETIC DATABASE C
STORE PHRASE TEMPLATE
EXPECTED AUDIO RESPONSE REFERENCE DATABASE B
END

FIG. 8



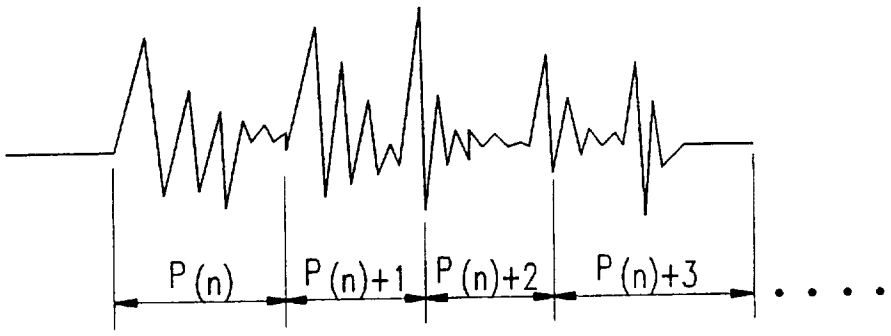


FIG. 9

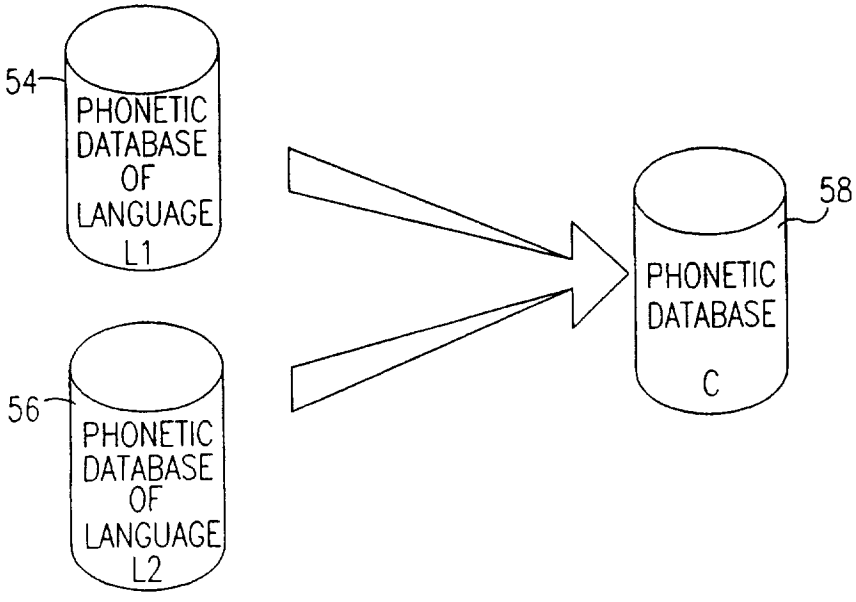


FIG. 10

FIG. 11

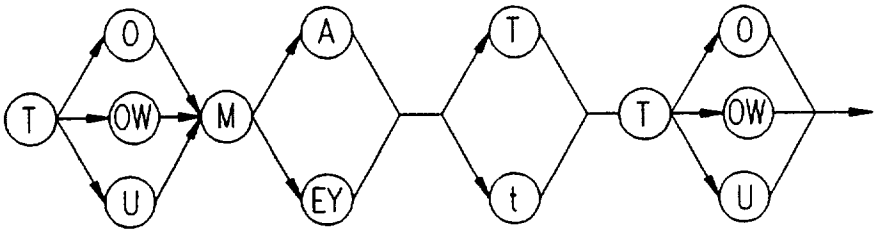
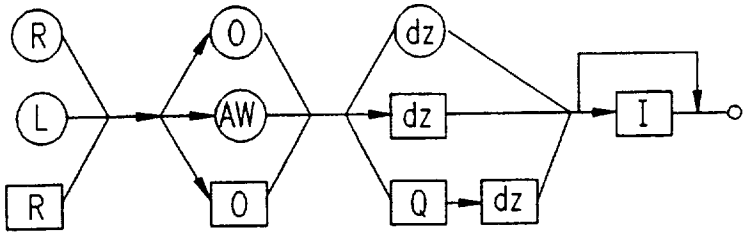


FIG. 12



INTERNATIONAL SEARCH REPORT

International application No.
PCT/IL97/00143

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G09B 1/00, 5/00, 19/00, 04, 06
US CL :434/156, 157, 167, 169, 185
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)

U.S. : 434/156, 157, 167, 169, 185

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 practicable, search terms used)

GPI Web Client
Search Terms: foreign, voice, recognition, words, correct, incorrect, database

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,487,671 A (SHPIRO et al) 30 January 1996, whole document	23-31
&	US 5,503,560 A (STENTIFORD) 2 April 1996.	1-31
&	US 5,393,236 A (BLACKMER et al) 28 February 1995.	1-31

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"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
"E"	earlier document published on or after the international filing date	"Y"	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 documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
24 OCTOBER 1997

Date of mailing of the international search report
12 NOV 1997

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
JOHN ROVNAK

Telephone No. (703) 308-3087

Form PCT/ISA/210 (second sheet)(July 1992)*