



(11) (21) (C) **2,151,399**  
(86) 1994/10/12  
(87) 1995/04/20  
(45) 2001/02/27

(72) Hirschberg, Julia, US

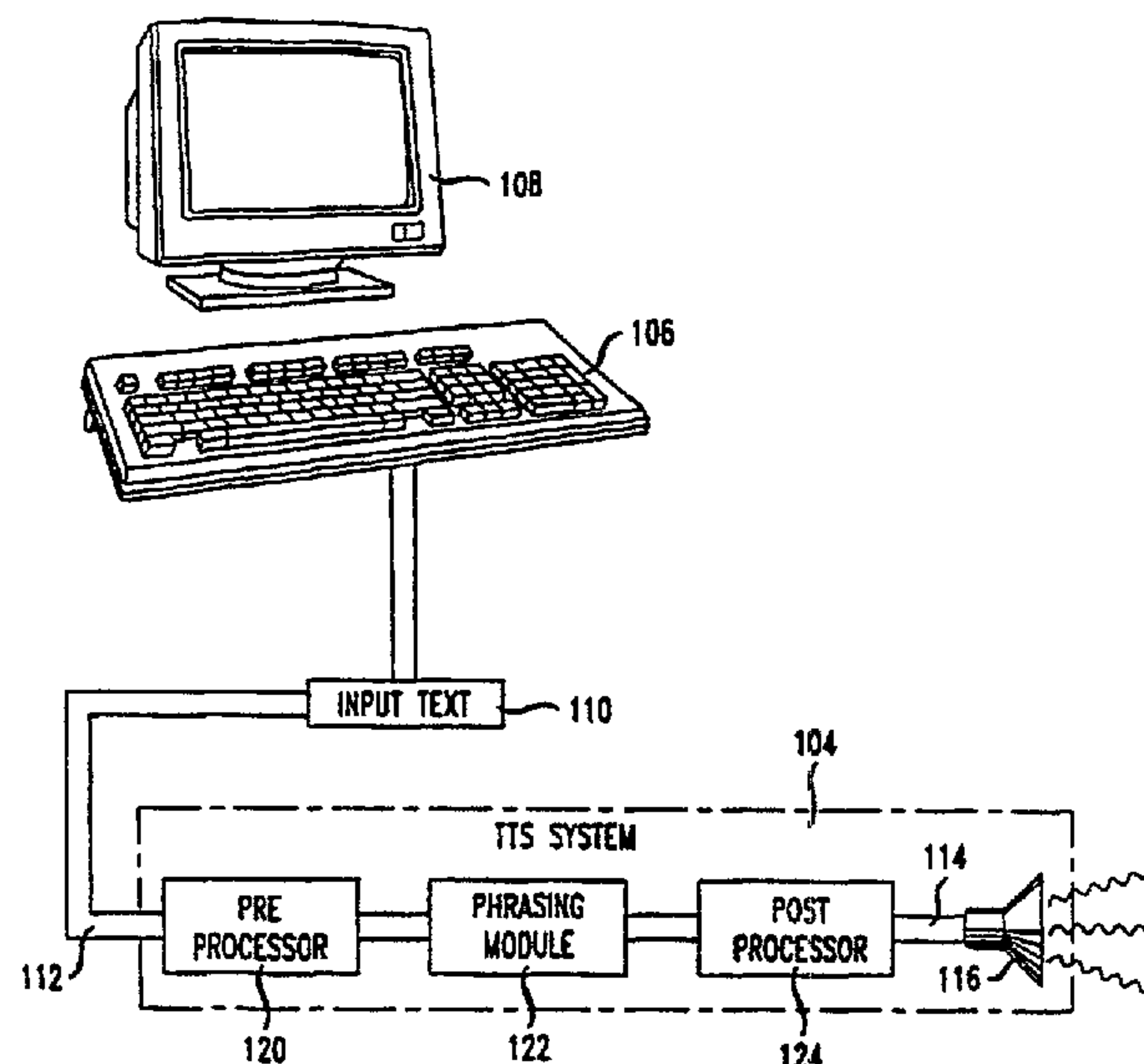
(73) AT&T CORP., US

(51) Int.Cl.<sup>6</sup> G10L 5/02, G10L 9/02

(30) 1993/10/15 (138,577) US

(54) **METHODE D'ENTRAINEMENT D'UN SYSTEME DE  
CONVERSION TEXTE-PAROLE, DISPOSITIF  
D'APPLICATION ET METHODE D'UTILISATION ASSOCIEE**

(54) **A METHOD FOR TRAINING A TEXT TO SPEECH SYSTEM,  
THE RESULTING APPARATUS, AND METHOD OF USE  
THEREOF**



(57) L'invention concerne un procédé permettant d'apprendre à un système de conversion de textes en langage parlé (TTS) (104) à attribuer à un texte (110) des caractéristiques d'intonation telles que les limites des syntagmes. Ce procédé d'apprentissage consiste à prendre un ensemble texte prédéterminé (110), qu'un utilisateur accompagne d'annotations relatives aux caractéristiques d'intonation. Ce texte passe ensuite dans le préprocesseur (120) et dans le module de mise en syntagmes (122), dans lequel un ensemble de noeuds de décision est généré par des informations d'analyse statistique basées sur la structure du texte prédéterminé. La représentation statistique peut être ensuite mémorisée et utilisée de manière répétée pour synthétiser la parole par l'intermédiaire du postprocesseur (124) à partir de nouveaux ensembles texte d'entrée sans apprentissage supplémentaire.

(57) A method of training a TTS (104) to assign intonational features, such as intonational phrase boundaries, to input text (110). The method of training involves taking a set of predetermined text (110) and having a human annotate it with intonational feature annotations. The text is passed through the preprocessor (120) and the phrasing module (122) wherein a set of decision nodes is generated by statistically analyzing information based upon the structure of the predetermined text. The statistical representation may then be stored and repeatedly used to generate synthesized speech, through the post processor (124), from new sets of input text without further training.



Abstract

The present invention relates to a method used in conjunction with a speech related system. The method is comprised of the steps of annotating a set of predetermined text with intonational feature annotations to generate annotated text, the set of predetermined text being unrelated to speech and generating a set of information regarding the predetermined text. The method also includes the step of generating a statistical representation that is a function of the set of information and the intonational feature annotations. The apparatus of the present invention is comprised of a stored statistical representation that is a function of a set of predetermined text and intonational feature annotations therefor. The set of predetermined text is unrelated to speech. A unit is provided for applying a set of input text to the stored statistical representation to generate an output representative to the set of input text.

A METHOD FOR TRAINING A TEXT TO SPEECH SYSTEM, THE  
RESULTING APPARATUS, AND METHOD OF USE THEREOF

Field of the Invention

The present invention relates to methods and systems for converting  
5 text-to-speech ("TTS"). The present invention also relates to the training of TTS  
systems.

Background of the Invention

In using a typical TTS system, a person inputs text, for example, via a  
computer system. The text is transmitted to the TTS system. Next, the TTS system  
10 analyzes the text and generates a synthesized speech signal that is transmitted to an  
acoustic output device. The acoustic output device outputs the synthesized speech  
signal.

The creation of the generated speech of TTS systems has focused on two  
characteristics, namely intelligibility and naturalness. Intelligibility relates to  
15 whether a listener can understand the speech produced (i.e., does "dog" really sound  
like "dog" when it is generated or does it sound like "dock"). However, just as  
important as intelligibility is the human-like quality, or naturalness, of the generated  
speech. In fact, it has been demonstrated that unnaturalness can affect intelligibility.

Previously, many have attempted to generate natural sounding speech  
20 with TTS systems. These attempts to generate natural sounding speech addressed a  
variety of issues.

One of these issues is the need to assign appropriate intonation to the  
speech. Intonation includes such intonational features, or "variations," as  
intonational prominence, pitch range, intonational contour, and intonational  
25 phrasing. Intonational phrasing, in particular, is "chunking" of words in a sentence  
into meaningful units separated by pauses, the latter being referred to as intonational  
phrase boundaries. Assigning intonational phrase boundaries to the text involves  
determining, for each pair of adjacent words, whether one should insert an  
intonational phrase boundary between them. Depending upon where intonational  
30 phrase boundaries are inserted into the candidate areas, the speech generated by a  
TTS system may sound very natural or very unnatural.

Known methods of assigning intonational phrase boundaries are  
disadvantageous for several reasons. Developing a model is very time consuming.  
Further, after investing much time to generate a model, the methods that use the  
35 model simply are not accurate enough (i.e., they insert a pause where one should not

2151399

- 2 -

be present and/or they do not insert a pause where one should be present) to generate natural sounding synthesized speech.

The pauses and other intonational variations in human speech often have great bearing on the meaning of the speech and are, thus, quite important. For example, with respect to intonational phrasing, the sentence "The child isn't screaming because he is sick" spoken as a single intonational phrase may lead the listener to infer that the child is, in fact, screaming, but not because he is sick. However, if the same sentence is spoken as two intonational phrases with an intonational phrase boundary between "screaming" and "because," (i.e., "The child isn't screaming, because he is sick") the listener is likely to infer that the child is not screaming, and the reason is that he is sick.

Assigning intonational phrasing has previously been carried out using one of at least five methods. The first four methods have an accuracy of about 65 to 75 percent when tested against human performance (e.g., where a speaker would have paused/not paused). The fifth method has a higher degree of accuracy than the first four methods (about 90 percent) but takes a long time to carry out the analysis.

A first method is to assign intonational phrase boundaries in all places where the input text contains punctuation internal to a sentence (i.e., a comma, colon, or semi-colon, but not a period). This method has many shortcomings. For example, not every punctuation internal to the sentence should be assigned an intonational phrase boundary. Thus, there should not be an intonational phrase boundary between "Rock" and "Arkansas" in the phrase "Little Rock, Arkansas." Another shortcoming is that when speech is read by a person, the person typically assigns intonational phrase boundaries to places other than internal punctuation marks in the speech.

A second method is to assign intonational phrase boundaries before or after certain key words such as "and," "today," "now," "when," "that," or "but." For example, if the word "and" is used to join two independent clauses (e.g. "I like apples and I like oranges"), assignment of an intonational phrase boundary (e.g., between "apples" and "and") is often appropriate. However, if the word "and" is used to join two nouns (e.g., "I like apples and oranges"), assignment of an intonational phrase boundary (e.g., between "apples" and "and") is often inappropriate. Further, in a sentence like "I take the 'nuts and bolts' approach," the assignment of an intonational phrase boundary between "nuts" and "and" would clearly be inappropriate.

A third method combines the first two methods. The shortcomings of these types of methods are apparent from the examples cited above.

A fourth method has been used primarily for the assignment of intonational phrase boundaries for TTS systems whose input is restricted by its application or domain (e.g., names and addresses, stock market quotes, etc...). This method has generally involved using a sentence or syntactic parser, the goal of which is to break up a sentence into subjects, verbs, objects, complements, etc.... Syntactic parsers have shortcomings for use in the assignment of intonational phrase boundaries in that the relationship between intonational phrase boundaries and syntactic structure has yet to be clearly established. Therefore, this method often assigns phrase boundaries incorrectly. Another shortcoming of syntactic parsers is their speed (or lack thereof), or inability to run in real time. A further shortcoming is the amount of memory needed for their use. Syntactic parsers have yet to be successfully used in unrestricted TTS systems because of the above shortcomings. Further, in restricted-domain TTS systems, syntactic parsers fail particularly on unfamiliar input and are difficult to extend to new input and new domains.

A fifth method that could be used to assign intonational phrase boundaries would increase the accuracy of appropriately assigning intonational phrase boundaries to about 90 percent. This is described in Wang and Hirschberg, "Automatic classification of intonational phrase boundaries," Computer Speech and Language, vol. 6, pages 175 - 196 (1992). The method involves having a speaker read a body of text into a microphone and recording it. The recorded speech is then prosodically labelled. Prosodically labeling speech entails identifying the intonational features of speech that one desires to model in the generated speech produced by the TTS system.

This method also has significant drawbacks. It is expensive because it usually entails the hiring of a professional speaker. A great amount of time is necessary to prosodically label recorded speech, usually about one minute for each second of recorded speech and even then only if the labelers are very experienced. Moreover, since the process is time-consuming and expensive, it is difficult to adapt this process to different languages, different applications, different speaking styles.

More specifically, a particular implementation of the last- mentioned method used about 45 to 60 minutes of natural speech that was then prosodically labeled. Sixty minutes of speech takes about 60 hours (e.g., 3600 minutes) just for prosodic labeling the speech. Additionally, there is much time required to record the speech and process the data for analysis (e.g., dividing the recorded data into

sentences, filtering the sentences, etc...). This usually takes about 40 to 50 hours. Also, the above assumes that the prosodic labeler has been trained; training often takes weeks, or even months.

### **Summary of the Invention**

5           We have discovered a method of training a TTS or other system to assign intonational features, such as intonational phrase boundaries, to input text that overcomes the shortcomings of the known methods. The method of training involves taking a set of predetermined text (not speech or a signal representative of speech) and having a human annotate it with intonational feature annotations (e.g., intonational  
10 phrase boundaries). This results in annotated text. Next, the structure of the set of predetermined text is analyzed – illustratively, by answering a set of text-oriented queries – to generate information which is used, along with the intonational feature annotations, to generate a statistical representation. The statistical representation may then be repeatedly used to generate synthesized speech from new sets of input text  
15 without training the TTS system further.

Advantageously, the invention improves the speed in which one can train a system that assigns intonational features, thereby also serving to increase the adaptability of the invention to different languages, dialects, applications, etc.

Also advantageously, the trained system achieves about 95 percent  
20 accuracy in assigning one type of intonational feature, namely intonational phrase boundaries, when measured against human performance.

In accordance with one aspect of the present invention there is provided a method of training a system for converting between text and speech, the method comprising the steps of: (a) annotating a set of predetermined text with intonational  
25 feature annotations to generate annotated text, the set of predetermined text being unrelated to speech, said annotating being performed by a human operator; (b) generating a set of structural information regarding the predetermined text; (c) generating a statistical representation of intonational feature information based on the set of structural information and the intonational feature annotations; and (d) storing  
30 said statistical representation in said system for use in the system in converting between text and speech.

- 4a -

In accordance with another aspect of the present invention there is provided an apparatus for performing text-to-speech conversion on a set of input text, the apparatus comprising: (a) a stored statistical representation of intonational feature information, the stored statistical representation based on a set of predetermined text and intonational feature annotations therefor, the set of predetermined text being unrelated to speech, the intonational feature annotations having been provided by a human annotator; and (b) a processor and a phrasing module for applying the set of input text to the stored statistical representation to generate an output representative of the set of input text, the output comprising intonational feature information associated with the set of input text.

In accordance with yet another aspect of the present invention there is provided a method for performing text-to-speech conversion on a set of input text, the method comprising the steps of: (a) accessing a stored statistical representation of intonational feature information, the stored statistical representation based on a set of predetermined text and intonational feature annotations therefor, the set of predetermined text being unrelated to speech, the intonational feature annotations having been provided by a human annotator; and (b) with processor means and a phrasing module means, applying the set of input text to the stored statistical representation to generate an output representative of the set of input text, the output comprising intonational feature information associated with the set of input text.

In accordance with still yet another aspect of the present invention there is provided an apparatus for converting text to speech, said apparatus comprising: (a) an input for receiving a set of input text having a physically tangible readable form; and (b) a phrasing module adapted to receive the set of input text from said input, said phrasing module including a stored statistical representation, the stored statistical representation being a function of a set of predetermined text and intonational feature annotations therefor, said phrasing module applying the set of input text to the stored statistical representation to generate an output representative of the set of input text.

In accordance with still yet another aspect of the present invention there is provided a machine implemented method of converting text to speech said method comprising: (a) accessing a stored statistical representation from a phrasing module, the stored statistical representation being a function of a set of predetermined text and

- 4b -

intonational feature annotations therefor; and (b) applying a set of input text having a physically tangible readable form to the stored statistical representation to generate an output representative of the set of input text.

### **Brief Description of the Drawings**

5                   Figure 1 shows a TTS system;  
                    Figure 2 shows a more detailed view of the TTS system; and  
                    Figure 3 shows a set of predetermined text having intonational feature annotations inserted therein.

### **Detailed Description**

10                   Figure 1 shows a TTS system 104. A person inputs, for example via a keyboard 106 of a computer 108, input text 110. The input text 110 is transmitted to the TTS system 104 via communications line 112. The TTS system 104 analyzes the input text 110 and generates a synthesized speech signal 114 that is transmitted to a loudspeaker 116. The loudspeaker 116 outputs a speech signal 118.

15                   Figure 2 shows, in more detail, the TTS system 104. The TTS system is comprised of four blocks, namely a pre-processor 120, a phrasing module 122, a post-processor 124, and an acoustic output device 126 (e.g., telephone, loudspeaker,

headphones, etc...). The pre-processor 120 receives as its input from communications line 112 the input text 110. The pre-processor takes the input text 110 and outputs a linked list of record structures 128 corresponding to the input text. The linked list of record structures 128 (hereinafter "records 128") comprises representations of words  
5 in the input text 110 and data regarding those words ascertained from text analysis. The records 128 are simply a set of ordered data structures. Except for the phrasing module 122, which implements the present invention, the other components of the system are of conventional design.

#### The pre-processor

10 Again referring to Figure 2, the pre-processor 120, which is of conventional design, is comprised of four sub-blocks, namely, a text normalization module 132, a morphological analyzer 134, an intonational prominence assignment module 136, and a dictionary look-up module 138. These sub-blocks are referred to as "TNM," "MA," "IPAM," and "DLUM," respectively, in Figure 2. These sub-  
15 blocks, which are arranged in a pipeline configuration (as opposed to in parallel), take the input text 110 and generate the records 128 corresponding to the input text 110 and data regarding the input text 110. The last sub-block in the pipeline (dictionary look-up module 138) outputs the records 128 to the phrasing module 122.

The text normalization module 132 of Figure 2 has as its input the input  
20 text 110 from the communications line 112. The output of the text normalization module 132 is a first intermediate set of records 140 which represents the input text 110 and includes additional data regarding the same. For example, the first intermediate set of records 140 includes, but is not limited to, data regarding:

- 25 (1) identification of words, punctuation marks, and explicit commands to the TTS system 104 such as an escape sequence;
- (2) interpretation for abbreviations, numbers, etc...; and
- (3) part of speech tagging based upon the words identified in "(1)" above (i.e., the identification of nouns, verbs, etc...).

The morphological analyzer 134 of Figure 2 has as its input the first  
30 intermediate set of records 140. The output of the morphological analyzer 134 is a second intermediate set of records 142, containing, for example, additional data regarding the lemmas or roots of words (e.g., "child" is the lemma of "children", "go" is the lemma of "went", "cat" is the lemma of "cats", etc...).

2151399

- 6 -

The intonational prominence assignment module 136 of Figure 2 has as its input the second intermediate set of records 142. The output of the intonational prominence assignment module 136 is a third intermediate set of records 144, containing, for example, additional data regarding whether each real word (as opposed to punctuation, etc...) identified by the text normalization module 132 should be made intonationally prominent when eventually generated.

The dictionary look-up module 138 of Figure 2 has as its input the third intermediate set of records 144. The output of the dictionary look-up module 138 is the records 128. The dictionary look-up module 138 adds to the third intermediate set of records 144 additional data regarding, for example, how each real word identified by the text normalization module 132 should be pronounced (e.g., how do you pronounce the word "bass") and what its component parts are (e.g., phonemes and syllables).

#### The phrasing module

The phrasing module 122 of Figure 2 embodying the invention, has as its input the records 128. The phrasing module 122 outputs a new linked list of record structures 146 containing additional data including but not limited to a new record for each intonational boundary assigned by the phrasing module 122. The phrasing module determines, for each potential intonational phrase boundary site (i.e., positions between two real words), whether or not to assign an intonational phrase boundary at that site. This determination is based upon a vector 148 associated with each individual site. Each site's vector 148 comprises a set of variable values 150. For example, for each potential intonational phrase boundary site  $\langle w_i, w_j \rangle$  (wherein  $w_i$  and  $w_j$  represent real words to the left and right, respectively, of the potential intonational phrase boundary site) one may ask the following set of text-oriented queries to generate the site's vector 148:

(1) is  $w_i$  intonationally prominent and if not, is it further reduced (i.e., cliticized)?;

(2) is  $w_j$  intonationally prominent and if not, is it further reduced (i.e., cliticized)?;

(3) what is the part of speech of  $w_i$  ?;

(4) what is the part of speech of  $w_{i-1}$  ?;

(5) what is the part of speech of  $w_j$  ?;

- 7 -

- (6) what is the part of speech of  $w_{j+1}$ ?
- (7) how many words are in the current sentence?
- (8) what is the distance, in real words, from  $w_j$  to the beginning of the sentence?
- 5 (9) what is the distance, in real words, from  $w_j$  to the end of the sentence?
- (10) what is the location (e.g., immediately before, immediately after, within, between two noun phrases, or none of the above) of the potential intonational boundary site with respect to the nearest noun phrase?
- 10 (11) if the potential intonational phrase boundary site is within a noun phrase, how far is it from the beginning of the noun phrase (in real words)?
- (12) what is the size, in real words, of the current noun phrase (defaults to zero if  $w_j$  is not within a noun phrase)?
- (13) how far into the noun phrase is  $w_j$  (i.e., if  $w_j$  is within a noun phrase, divide "(11)" above by "(12)" above, otherwise this defaults to zero)?
- 15 (14) how many syllables precede the potential intonational boundary site in the current sentence?
- (15) how many strong (lexically stressed) syllables precede the potential intonational boundary site in the current sentence?
- (16) what is the total number of strong syllables in the current sentence?
- 20 (17) what is the stress level (i.e., primary, secondary, or unstressed) of the syllable immediately preceding the potential intonational boundary site?
- (18) what is the result when one divides the distance from  $w_j$  to the last intonational boundary assigned, by the total length of the last intonational phrase?
- 25 (19) is there punctuation (e.g., comma, dash, etc...) at the potential intonational boundary site?; and
- (20) how many primary or secondary stressed syllables exist between the potential intonational boundary site and the beginning of the current sentence.

**SUBSTITUTE SHEET (RULE 26)**

2151399

- 8 -

The variable values corresponding to the answers to the above 20 questions are encoded into the site's vector 148 in a vector generator 151 (referred to as "VG" in Figure 2). An vector 148 is formed for each site. The vectors 148 are sent, in serial fashion, to a set of decision nodes 152. Ultimately, the set of decision nodes 152 provide an indication of whether or not each potential intonational phrase boundary site should or should not be assigned as an intonational phrase boundary. The set of above twenty questions are asked because the set of decision nodes 152 was generated by applying the same set of 20 text-oriented queries to a set of annotated text in accordance with the invention. Preferably, the set of decision nodes 152 comprises a decision tree 154. Preferably, the decision tree has been generated using classification and regression tree ("CART") techniques that are known as explained in Brieman, Olshen, and Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, California (1984).

It should be noted that the above set of queries comprises text-oriented queries and is currently the preferred set of queries to ask. However, those skilled in the art will realize that subsets of the above set of queries, different queries, and/or additional queries may be asked that obtain satisfactory results. For example, instead of asking queries relating to part-of-speech of words in the sentence (as in (3) through (6) above), queries relating to the syntactic constituent structure of the input text or co-occurrence statistics regarding adjacent words in the input text may be asked to obtain similar results. The queries relating syntactic constituent structure focus upon the relationship of the potential intonational phrase boundary to the syntactic constituents of the current sentence (e.g., does the potential intonational phrase boundary occur between a noun phrase and a verb phrase?). The queries relating co-occurrence focus upon the likelihood of two words within the input text appearing close to each other or next to each other (e.g., how frequently does the word "cat" co-occur with the word "walk").

#### The post-processor

Again referring to Figure 2, post-processor 124, which is of conventional design, has as its input the new linked list of records 146. The output of the post-processor is a synthesized speech signal 114. The post-processor has seven sub-blocks, namely, a phrasal phonology module 162, a duration module 164, an intonation module 166, an amplitude module 168, a dyad selection module 170, a dyad concatenation module 172, and a synthesizer module 173. These sub-blocks are referred to as "PPM," "DM," "IM," "AM," "DSM," "DCM," and "SM,"

respectively, in Figure 2. The above seven modules address, in a serial fashion, how to realize the new linked list of records 146 in speech.

The phrasal phonology module 162 takes the new linked list of records 146. The phrasal phonology module outputs a fourth intermediate set of records 174  
5 containing, for example, what tones to use for phrase accents, pitch accents, and boundary tones and what prominences to associate with each of these tones. The above terms are described in Pierrehumbert, The Phonology and Phonetics of English Intonation, (1980) M.I.T. Ph.D. Thesis.

The duration module 164 takes the fourth intermediate set of records  
10 174 as its input. This module outputs a fifth set of intermediate records 176 containing, for example, the duration of each phoneme that will be used to realize the input text 110 (e.g., in the sentence "The cat is happy" this determines how long the phoneme "/p/" will be in "happy").

The intonation module 166 takes the fifth set of records 176 as its input.  
15 This module outputs a sixth set of intermediate records 178 containing, for example, the fundamental frequency contour (pitch contour) for the current sentence (e.g., whether the sentence "The cat is happy" will be generated with falling or rising intonation).

The amplitude module 168 takes the sixth set of records 178 as its input.  
20 This module outputs a seventh set of intermediate records 180 containing, for example, the amplitude contour for the current sentence (i.e., how loud each portion of the current sentence will be).

The dyad selection module 170 takes the seventh set of records 180 as its input. This module outputs a eighth set of intermediate records 182 containing,  
25 for example, a list of which concatenative units (i.e., transitions from one phoneme to the next phoneme) should be used to realize the speech.

The dyad concatenation module 172 takes the eighth set of records 182 as its input. This module outputs a set of linear predictive coding reflection coefficients 184 representative of the desired synthetic speech signal.

30 The synthesizer module 173 takes the set of linear predictive coding reflection coefficients 184 as its input. This module outputs the synthetic speech signal to the acoustic output device 126.

2151399

- 10 -

Training the system

The training of TTS system 104 will now be described in accordance with the principles of the present invention.

5 The training method involves annotating a set of predetermined text 105 with intonational feature annotations to generate annotated text. Next, based upon structure of the set of predetermined text 105, information is generated. Finally, a statistical representation is generated that is a function of the information and the intonational feature annotations.

10 Referring to Figure 3, an example of the set of predetermined text 105 is shown separately and then is shown as "annotated text." The symbols ' ', designated by reference numerals 190, are used to denote 'predicted intonational boundary.' In practice, much more text than the amount shown in Figure 3 will likely be required to train a TTS system 104. Next, the set of predetermined text 105 is passed through the pre-processor 120 and the phrasing module 122, the latter module being the  
15 module wherein, for example, a set of decision nodes 152 is generated by statistically analyzing information. More specifically, the information (e.g., information set) that is statistically analyzed is based upon the structure of the set of predetermined text 105. Next, a statistical analysis may be done by using CART techniques, as described above. This results in the statistical representation (e.g., the  
20 set of decision nodes 152). The set of decision nodes 152 takes the form of a decision tree. However, those skilled in the art will realize that the set of decision nodes could be replaced with a number of statistical analyses including, but not limited to, hidden Markov models and neural networks.

The statistical representation (e.g., the set of decision nodes 152) may  
25 then be repeatedly used to generate synthesized speech from new sets of text without training the TTS system further. More specifically, the set of decision nodes 152 has a plurality of paths therethrough. Each path in the plurality of paths terminates in an intonational feature assignment predictor that instructs the TTS system to either insert or not insert an intonational feature at the current potential intonational feature  
30 boundary site. The synthesized speech contains intonational features inserted by the TTS system. These intonational features enhance the naturalness of the sound that emanates from the acoustic output device, the input of which is the synthesized speech.

The training mode can be entered into by simply setting a "flag" within  
35 the system. If the system is in the training mode, the phrasing module 122 is run in its "training" mode as opposed to its "synthesis" mode as described above with

- 11 -

reference to Figures 1 and 2. In the training mode, the set of decision nodes 152 is never accessed by the phrasing module 122. Indeed, the object of the training mode is to, in fact, generate the set of decision nodes 152.

5 It will be appreciated by those skilled in the art that given different sets annotated text will result in different sets of decision nodes. For example, fictional text might be annotated in quite a different manner by the human annotator than scientific, poetic, or other types of text.

The invention has been described with respect to a TTS system. However, those skilled in the art will realize that the invention, which is defined in 10 the claims below, may be applied in a variety of manners. For example, the invention, as applied to a TTS system, could be one for either restricted or unrestricted input. Also, the invention, as applied to a TTS system, could differentiate between major and minor phrase boundaries or other levels of phrasing. Further, the invention may be applied to a speech recognition system. Additionally, 15 the invention may be applied to other intonational variations in both TTS and speech recognition systems. Finally, those skilled in the art will realize that the sub-blocks of both the pre-processor and post-processor are merely important in that they gather and produce data and that the order in which this data is gathered and produced is not tantamount to the present invention (e.g., one could switch the order 20 of sub-blocks, combine sub-blocks, break the sub-blocks into sub-sub-blocks, etc...). Although the system described herein is a TTS system, those skilled in the art will realize that the phrasing module of the present invention may be used in other systems such as speech recognition systems. Further, the the above description focuses on an evaluation of whether to insert an intonational phrase boundary in each 25 potential intonational phrase boundary site. However, those skilled in the art will realize that the invention may be used with other types of potential intonational feature sites.

**Claims:**

1. A method of training a system for converting between text and speech, the method comprising the steps of:
  - (a) annotating a set of predetermined text with intonational feature annotations to generate annotated text, the set of predetermined text being unrelated to speech, said annotating being performed by a human operator;
  - (b) generating a set of structural information regarding the predetermined text;
  - (c) generating a statistical representation of intonational feature information based on the set of structural information and the intonational feature annotations; and
  - (d) storing said statistical representation in said system for use in the system in converting between text and speech.
2. The method of claim 1 wherein the step of annotating comprises prosodically annotating the set of predetermined text with expected intonational features.
3. The method of claim 1 wherein the system is a text-to-speech system.
4. The method of claim 3 wherein the intonational features comprise intonational phrase boundaries.
5. The method of claim 1 wherein generating a statistical representation comprises generating a set of decision nodes.
6. The method of claim 5 wherein generating the set of decision nodes comprises generating a hidden Markov model.
7. The method of claim 5 wherein generating the set of decision nodes comprises generating a neural network.
8. The method of claim 5 wherein generating the set of decision nodes comprises performing classification and regression tree techniques.
9. An apparatus for performing text-to-speech conversion on a set of input text, the apparatus comprising:

- 13 -

(a) a stored statistical representation of intonational feature information, the stored statistical representation based on a set of predetermined text and intonational feature annotations therefor, the set of predetermined text being unrelated to speech, the intonational feature annotations having been provided by a human  
5 annotator; and

(b) a processor and a phrasing module for applying the set of input text to the stored statistical representation to generate an output representative of the set of input text, the output comprising intonational feature information associated with the set of input text.

10 10. The apparatus of claim 9 further comprising:

(a) means for post-processing the output to generate a synthesized speech signal; and

(b) means for applying the synthesized speech signal to an acoustic output device.

15 11. The apparatus of claim 9 wherein the stored statistical representation comprises a decision tree.

12. The apparatus of claim 9 wherein the stored statistical representation comprises a hidden Markov model.

20 13. The apparatus of claim 9 wherein the stored statistical representation comprises a neural network.

25 14. The apparatus of claim 9 wherein the phrasing module comprises means for answering a set of stored queries regarding the set of input text, the set of input text comprising a current sentence, the current sentence comprising a beginning, an end, and a plurality of words, each word in the plurality of words being a part of at least one set of words,  $w_i$  and  $w_j$ , wherein  $w_i$  and  $w_j$  each comprise at least one syllable and each have a part of speech associated therewith and each have a potential noun phrase associated therewith, the potential noun phrase having a beginning and an end, and further wherein  $w_i$  and  $w_j$  represent real words to the left and right, respectively, of a potential intonational phrase boundary site,  $\langle w_i \text{ and } w_j \rangle$ , and  
30 wherein  $w_{i-1}$  and  $w_{j+1}$  represent real words to the left and right, respectively of  $w_i$  and  $w_j$ , the set of stored queries comprising at least one query selected from the group consisting of:

- 14 -

- (a) is  $w_i$  intonationally prominent and if not, is it further reduced?;
- (b) is  $w_j$  intonationally prominent and if not, is it further reduced?;
- (c) what is the part of speech of  $w_i$ ?;
- (d) what is the part of speech of  $w_{i-1}$ ?;
- 5 (e) what is the part of speech of  $w_j$ ?;
- (f) what is the part of speech of  $w_{j+1}$ ?;
- (g) how many words are in the current sentence?;
- (h) what is the distance, in real words, from  $w_j$  to the beginning of the sentence?;
- 10 (i) what is the distance, in real words, from  $w_j$  to the end of the sentence?;
- (j) what is the location of the potential intonational boundary site with respect to the nearest noun phrase?;
- (k) if the potential intonational boundary site is within a noun phrase, how far is it from the beginning of the noun phrase?;
- 15 (l) what is the size, in real words, of the current noun phrase?;
- (m) how far into the noun phrase is  $w_i$ ?;
- (n) how many syllables precede the potential intonational boundary site in the current sentence?;
- (o) how many lexically stressed syllables precede the potential
- 20 intonational boundary site in the current sentence?;
- (p) what is the total number of strong syllables in the current sentence?;
- (q) what is the stress level of the syllable immediately preceding the potential intonational boundary site?;
- (r) what is the result when one divides the distance from  $w_j$  to the last
- 25 intonational boundary assigned by the total length of the last intonational phrase?;
- (s) is there punctuation at the potential intonational boundary site?; and
- (t) how many primary or secondary stressed syllables exist between the potential intonational boundary site and the beginning of the current sentence.

15. A method for performing text-to-speech conversion on a set of input  
30 text, the method comprising the steps of:

- (a) accessing a stored statistical representation of intonational feature information, the stored statistical representation based on a set of predetermined text and intonational feature annotations therefor, the set of predetermined text being unrelated to speech, the intonational feature annotations having been provided by a
- 35 human annotator; and

- 15 -

(b) with processor means and a phrasing module means, applying the set of input text to the stored statistical representation to generate an output representative of the set of input text, the output comprising intonational feature information associated with the set of input text.

5           16. The method of claim 15 further comprising the steps of:

(a) post-processing the output to generate a synthesized speech signal;  
and

(b) applying the synthesized speech signal to an acoustic output device.

10           17. The method of claim 15 wherein the stored statistical representation comprises a decision tree.

18. The method of claim 15 wherein the stored statistical representation comprises a hidden Markov model.

19. The method of claim 15 wherein the stored statistical representation comprises a neural network.

15           20. The method of claim 15 wherein the step of applying comprises answering the set of stored queries regarding the set of input text, the set of input text comprising a current sentence, the current sentence comprising a beginning, an end, and a plurality of words, each word in the plurality of words being a part of at least one set of words,  $w_i$  and  $w_j$ , wherein  $w_i$  and  $w_j$  each comprise at least one syllable and  
20 each have a part of speech associated therewith and each have a potential noun phrase associated therewith, the potential noun phrase having a beginning and an end, and further wherein  $w_i$  and  $w_j$  represent real words to the left and right, respectively, of a potential intonational phrase boundary site,  $\langle w_i \text{ and } w_j \rangle$ , the set of stored queries comprising at least one query selected from the group consisting of:

25           (a) is  $w_i$  intonationally prominent and if not, is  $w_i$  further reduced?;

(b) is  $w_j$  intonationally prominent and if not, is  $w_j$  further reduced?;

(c) what is  $w_i$ 's part of speech?;

(d) what is  $w_{i-1}$ 's part of speech?;

(e) what is  $w_j$ 's part of speech?;

30           (f) what is  $w_{j+1}$ 's part of speech?;

(g) how many words are in the current sentence?;

(h) how far, in real words, is  $w_j$  from the beginning of the sentence?;

- 16 -

- (i) how far, in real words, is  $w_j$  from the end of the sentence?;
- (j) where is the potential intonational phrase boundary site with respect to the potential noun phrase?;
- (k) if  $\langle w_i \text{ and } w_j \rangle$  is within the potential noun phrase, how far is  $\langle w_i \text{ and } w_j \rangle$  from the beginning of the potential noun phrase?;
- (l) how many words are in the potential noun phrase?;
- (m) how far into the potential noun phrase is  $w_i$ ?;
- (n) how many syllables precede the potential intonational phrase boundary site in the current sentence?;
- (o) how many lexically stressed syllables precede the potential intonational phrase boundary site in the current sentence?;
- (p) how many strong syllables are there in the current sentence?;
- (q) what is a stress level of a syllable in  $w_i$  immediately preceding the potential intonational boundary site?;
- (r) what is a result of dividing a distance from  $w_j$  to a last intonational boundary assigned by a total length of the last intonational phrase?;
- (s) is there punctuation at the potential intonational phrase boundary site?; and
- (t) how many primary and secondary stressed syllables exist between the potential intonational phrase boundary site and the beginning of the current sentence?

21. An apparatus for converting text to speech, said apparatus comprising:

- (a) an input for receiving a set of input text having a physically tangible readable form; and
- (b) a phrasing module adapted to receive the set of input text from said input, said phrasing module including a stored statistical representation, the stored statistical representation being a function of a set of predetermined text and intonational feature annotations therefor, said phrasing module applying the set of input text to the stored statistical representation to generate an output representative of the set of input text.

22. A machine implemented method of converting text to speech said method comprising:

- 17 -

(a) accessing a stored statistical representation from a phrasing module, the stored statistical representation being a function of a set of predetermined text and intonational feature annotations therefor; and

5 (b) applying a set of input text having a physically tangible readable form to the stored statistical representation to generate an output representative of the set of input text.

23. A machine implemented method of converting text to speech, said method comprising:

10 (a) accessing a stored statistical representation from a phrasing module, the stored statistical representation being a function of a set of predetermined text and intonational feature annotations therefor;

(b) applying a set of input text having a physically tangible readable form to the stored statistical representation to generate an output representative of the set of input text; and

15 (c) post-processing the output to generate a synthesized speech signal.

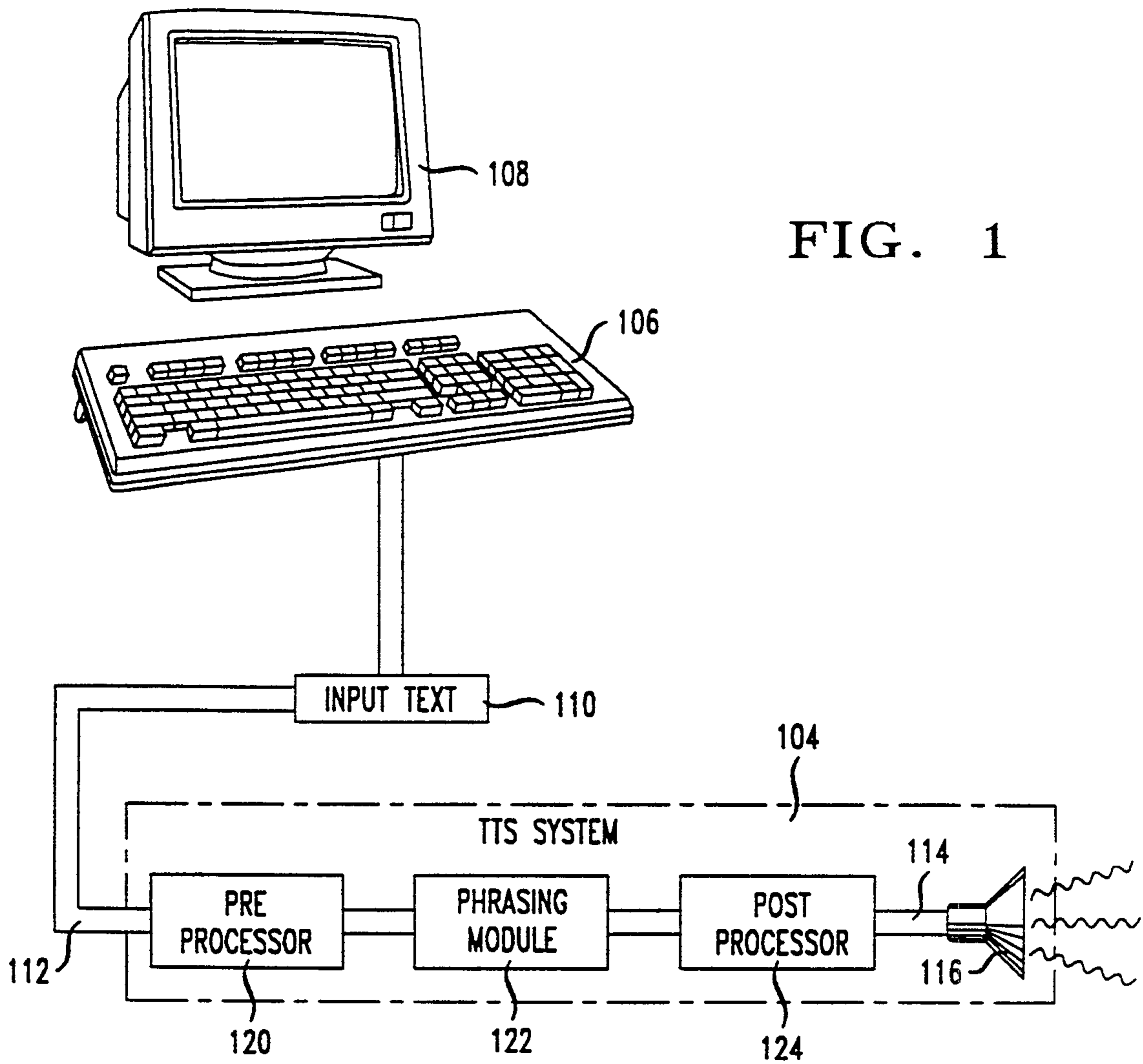


FIG. 1

FIG. 3

105

Picco, who was instrumental in freeing the last American and British hostages in Lebanon last year, has in the past met with kidnapers and their emissaries in Eastern Lebanon close to the Syrian border.

190

Picco, who was instrumental in freeing the last American and British hostages in Lebanon last year, has in the past met with kidnapers and their emissaries in Eastern Lebanon close to the Syrian border.

190 190

FIG. 2

