(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 606 520 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **93106692.2**

(22) Date of filing: **24.04.93**

(51) Int. Cl.[5]: **G10L 5/02**, G10L 7/02, G10L 7/10

(30) Priority: **15.01.93 IT MI930044**

(43) Date of publication of application:
20.07.94 Bulletin 94/29

(84) Designated Contracting States:
AT BE CH DE ES FR GB IT LI NL SE

(71) Applicant: **ALCATEL ITALIA S.p.A.**
Via Monte Rosa 15
I-20149 Milano(IT)
(84) **IT**

(71) Applicant: **ALCATEL N.V.**
Strawinskylaan 341,
(World Trade Center)
NL-1077 XX Amsterdam(NL)
(84) **BE CH DE ES FR GB LI NL SE AT**

(72) Inventor: **Abbattista, Giuseppe**
Via Mereu, 30
I-70033 Corato (BA)(IT)
Inventor: **Tambone, Gabriella**
Via Mereu, 30
I-70033 Corato (BA)(IT)

(74) Representative: **Pohl, Herbert, Dipl.-Ing et al**
Alcatel SEL AG
Patent- und Lizenzwesen
Postfach 30 09 29
D-70449 Stuttgart (DE)

(54) Method of implementing intonation curves for vocal messages, and speech synthesis method and system using the same.

(57) The present invention refers to a method of implementing intonation curves for vocal messages, and to a speech synthesis method and system using the same.

Such methods characterize in that the generation of the pronunciation and of the intonation are carried out independently and separately using a coding technique that permits the instantaneous control of intonation, i.e. of pitch.

EP 0 606 520 A2

The present invention refers to a method of implementing intonation curves for vocal messages, and to a speech synthesis method and system using the same.

Nowadays the use of vocal messages is very common in many technical applications, for example for telephone services, such as the "right time of day" and the "news", and at the airports and stations. Till some years ago such vocal messages were pronounced by physical persons; afterwards it became popular to pre-record whole vocal messages, for example on magnetic media, and to emit them by selection. Today tendency is to use automatic speech synthesis systems which, using a vocabulary of pre-recorded words (or syllables or groups of letters), synthesize the vocal message concatenating such elementary entities according to the correspondent text messages.

The quality of such automatic speech synthesis systems is evaluated on the base of the naturalness of the emitted vocal message as can be appreciated by a listener; the naturalness depends on, in addition to other elements, the emission intonation and the length of the single words and, as a consequence, of the whole vocal message.

The fact that such a concatenation does not result in a natural vocal message depends on how the vocabulary is generated: a professional speaker pronounces an enormous number of sentences, the sentences are recorded, and the various words are extracted from such recording; each pre-recorded word has a length and an intonation that depends on the sentence from which it is extracted (particularly, but not only, on the position of the word inside the sentence) and therefore can not be suitable for every sentence to be synthesized.

One way of achieving a reasonable degree of naturalness is proposed in the European patent EP-B1 0 093 022: here it is disclosed a signal generator, for use in providing a spoken message, in which data defining predetermined words and/or phrases are stored in a data store in a plurality of forms corresponding to different intonations, and in which control means are arranged to select the appropriate form in dependence upon the position of such words and/or phrases in the message. As can be clearly seen from Fig.2 this leads to a disadvantageous increase in the dimension of the database; to limit the dimension of the data store it is also proposed, in the same European patent, that the data stored in such data store, instead of being in PCM format, to be derived from parametrisation of analogue speech waveforms using, for example, a method known as LPC (Linear Predictive Coding).

It is the general task of the present invention to provide a speech synthesis system having good intonation performances and suitable for both text-to-speech applications and message applications.

It is a first object of the present invention to provide a method of implementing intonation curves which is independent from the vocal messages and easy to be implemented.

It is a second object of the present invention to provide a speech synthesis method which requires a small amount of additional memory for the generation of the suitable intonation.

It is a third object of the present invention to provide a speech synthesis method which does not require the linguistic analysis of the vocal message to be emitted in order to generate the suitable intonation.

These and other objects, which will be apparent from the following description, are reached through the method of implementing intonation curves for vocal messages as set out in claim 1, the speech synthesis method as set out in claim 7 and the speech synthesis system as set out in claim 13; further advantageous aspects of the present invention are set out in the dependent claims.

The methods of the present invention are based on the idea of handling independently and separately the pronunciation and the intonation of vocal messages.

More particularly, the methods of the present invention make use of a small database including word intonation curves of words, extract from such database the appropriate (from the intonation point of view) word intonation curves corresponding to the words of the vocal message to be synthesized, and concatenate them to form the message intonation curve.

Further advantageously the word intonation database includes word intonation curves only of a limited number of words, and the methods extract from such database the word intonation curves corresponding, only from the intonation point of view, to the words of the vocal message to be synthesized.

An important advantage of the present invention is that it is applicable to many different languages with minor changes which are very simple and straightforward for the person skilled in the art of speech processing.

The present invention will become more apparent from the following description.

The methods of the present invention may be used only if the vocal messages are coded through a technique that permits the instantaneous control of intonation, i.e. of pitch, for example the LPC technique. The majority of these techniques divides the vocal signal into a sequence of frames and codifies each frame through a set of parameters, one of them being or being directly related to the pitch.

The methods of implementing a message intonation curve for a vocal message, which will be described, are embodiments of the idea of generating the intonation independently and separately from the generation of the pronunciation of the vocal message.

A first method, according to the present invention, of implementing a message intonation curve for a vocal message corresponding to a text message consisting of the concatenation of words, makes use of a word intonation database including word intonation curves and includes the steps of:

a) determining, for each word of the text message, the value of at least one parameter corresponding its position inside the text message, and

b) extracting, for each word of the text message, from the word intonation database, a word intonation curve corresponding to each word and to such value.

The message intonation curve is the concatenation of the word intonation curves so obtained.

The word intonation curves could be sequences of values corresponding to instantaneous intonation or pitch.

The message intonation curve depends, in addition, on the general intonation of the text message, i.e. interrogative, affirmative, exclamatory, dubitative.

In this case, in step a) the value of a second parameter is determined corresponding to the general intonation of the text message, and in step b) the word intonation curve corresponds to the word and to the values of two parameters.

From the above it is clear that, for each word of the vocabulary necessary for a specific application, a number of different intonation curves should be included into the word intonation database.

Before prosecuting further in the description of more advantageous features of the present invention, it is important to evidence that the concept of "syllable" is slightly different from language to language but still can be found in every language at least as a part of a word.

If a particular language is considered, a number may be determined which represents the maximum number of syllables of a word in that particular language; naturally a lower number may be chosen (predetermined limit) and as a consequence only a certain percentage of the words of that particular language fall within such limit. In an embodiment of the present invention for the Italian language it turned out that choosing this predetermined limit equal to six will lead to satisfactory results.

The words of a language may be divided in a number of classes each corresponding to a different number of syllables. Each class may be further divided according to the position of the accented syllable (primary accent). In the same embodiment named above the number of classes so obtained is twenty-two : two classes for the words composed of one syllable, two classes for two syllables, three classes for three syllables, ..., six classes for six syllables. The fact that two classes are necessary for one syllable depends from the fact that the Italian words "il" and "la" (one-syllable words) have two different intonations as the vowel in the first case is followed by a consonant and not in the second case.

In addition to the characteristics of the word itself, its correct intonation in a sentence is related to its position inside it; at least three positions should be considered corresponding to: the beginning, the middle, and the end of the sentence.

Once these preliminary considerations have been set out, another method according to the present invention will be described.

A second method, according to the present invention, of implementing a message intonation curve for a vocal message corresponding to a text message consisting of the concatenation of words, makes use of a word intonation database including word intonation curves and includes the steps of:

a) determining, for each word of the text message, the values of at least three parameters respectively corresponding to the number of syllables, the position of the accented syllable, and the position of each word inside the text message, and

b) extracting, for each word of the text message, from the word intonation database, a word intonation curve corresponding to such values.

The message intonation curve is the concatenation of the word intonation curves so obtained.

The word intonation database must include, according to the above-mentioned embodiment, at least sixty-six word intonation curves corresponding to twenty-two classes times three positions; the amount of memory required for storing such database is clearly small, at least if compared with the amount of memory necessary for storing the vocabulary.

If better results are desired the number of predetermined positions may be increased; it is surely useful to consider as a possible position the last but one in the sentence: in fact the last but one word of a sentence, from the intonation point of view, is the preparation to end of the sentence; in addition another possible position is the second in the sentence, particularly useful when the first word in the sentence is short. Naturally the number of word intonation curves, and consequently of memory required, increases

respectively to eighty-eight and one hundred and ten.

The message intonation curve depends, in addition, on the general intonation of the text message, i.e. interrogative, affirmative, exclamatory, dubitative, etc.. If the implementation of the message intonation curve shall take this element into account the number of curves in the database shall increase further: if, for example, three different general intonations are considered, the number of curves could be 198 [3 X 66].

In this case, in step a) the value of a fourth parameter is determined corresponding to the general intonation of the text message, and in step b) the word intonation curve corresponds to the values of four parameters.

It is now time to explain how can be solved the problem of words exceeding the predetermined limit of number of syllables.

A first word intonation curve corresponding to a first word having a number of syllables bigger than the predetermined limit is obtained through the concatenation of at least a second and a third word intonation curves respectively corresponding to at least a second and a third words; the global number of syllables of said second and said third words being equal to the number of syllables of said first word.

In the already mentioned embodiment the following table was used:

number of syllables	combination to be used
7	3 + 4
8	3 + 5
9	4 + 5
10	4 + 6
etc.	

obtaining good results.

In this case two extractions from the database take place; the values of the other two parameters, the position of the accented syllable and the position of the word, should be determined as if the word would be actually splitted into two words: if the position of the first word (9 syllables) is "begin of sentence", the position of second word (4 syllables) is "begin of sentence" and of the third word (5 syllables) is "middle of sentence"; if the accented syllable of the first word (8 syllables) is the seventh one, the accented syllable of the second word (4 syllables) is chosen by chance and the accented syllable of the third word (5 syllables) is the third.

In the case of the accent an unavoidable error is made; this is extremely limited when the secondary accent coincide with such a chance choice. If anyway the predetermined limit is suitably chosen this error is very rare.

The method, according to the present invention, of synthetizing an intonated vocal message corresponding to a text message consisting of the concatenation of words, includes at first the generation of the pronunciation of the vocal message, at second the generation of the intonation of the vocal message independent and separate from the generation of its pronunciation, and at last the association of the pronunciation and the intonation into the intonated vocal message.

The intonation of the vocal message may be generated according to a message intonation curve obtained through one of the above-described methods; the pronunciation of the vocal message may be generated either algorithmically or by using a word vocal database.

In a method of synthetizing an "intonated" (having a correct intonation) vocal message corresponding to a text message consisting of the concatenation of words, using a word vocal database including for each word a sequence of frames, and using a word intonation database including word intonation curves, each of said word intonation curves being a sequence of pitch values corresponding to a sequence of frames of a word of said word vocal database, and being identified by the value of at least one parameter corresponding the position of such word inside a sentence,

the vocal message is obtained through the extraction of the sequences of frames corresponding to the words of said text message from said word vocal database and the concatenation of said sequences,

the message intonation curve of said vocal message is obtained through the above-described first method, and

the intonated vocal message is obtained associating to each frame of said vocal message a corresponding pitch value of said message intonation curve.

In another method of synthesizing an "intonated" (having a correct intonation) vocal message corresponding to a text message consisting of the concatenation of words, using a word vocal database including for each word a sequence of frames and at least two data fields respectively corresponding to the

number of syllables and the position of the accented syllable, and using a word intonation database including word intonation curves of sample words, each of the word intonation curves being a sequence of pitch values corresponding to the sequence of frames of the correspondent sample word, and being identified by the values of at least three parameters respectively corresponding to the number of syllables of the correspondent sample word, the position of the accented syllable of the correspondent sample word, and the position of the correspondent sample word inside a sentence,

the vocal message is obtained through the extraction of the sequences of frames corresponding to the words of the text message from the word vocal database and the concatenation of the sequences,

the message intonation curve of the vocal message is obtained through the above-described second method, and

the intonated vocal message is obtained associating to each frame of the vocal message a corresponding pitch value of the message intonation curve.

Sometimes it is useful to store in the word vocal database not only sequences of frames corresponding to single words but also sequences of frames corresponding to short groups of words, such as: "The flight for", "is cancelled", "is delayed", etc..

In this case the word intonation curve shall be implemented splitting the group of words into the single words; for example, if the group "The flight for" has to be synthesized, at first the correspondent sequence of frames has to be extracted from the word vocal database, at second three suitable intonation curves corresponding to the words "The", "flight", "for" has to be extracted from the word intonation database, and at last such three curves has to be concatenated to form a "long" word intonation curve; this requires that the data fields associated to each sequence of frames provide the following information: number of words, numbers of syllables of the different words, and positions of the accented syllables inside the different words.

The lengths of the various sequences of frames of the word vocal database are quite different and depends at least on the number of syllables of the correspondent various words; in addition such lengths are related to the words themselves.

It is possible that, during the construction of such database, such lengths be normalized to predetermined lengths: for example for 1 syllable words a sequence of 15 frames of 20 mS each, for 2 syllables words 30 frames, for 3 syllables words 45 frames, etc..

The lengths of the various sequences of pitch values are important because they contribute in determining the naturalness of the emitted vocal message. Therefore, during the construction of the word intonation database, they must be carefully considered and, during the synthesis, they must be used as a reference.

If a correct one-to-one association has to be obtained, the length of the sequence of frames of a word and the length of the relevant sequence of pitch values should be the same.

This may be accomplished through the adjustment of the vocal message, obtained by simple extraction and concatenation, in such a way that the length of each of the sequences of frames, corresponding to the words of the text message, be equal to the length of the relevant word intonation curve extracted.

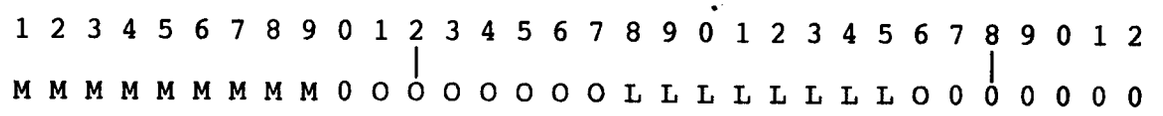
If the word vocal database includes for each word a further data field corresponding to stationary vowel frames, and if the word intonation database includes for each word intonation curve a further data field corresponding to stationary vowel pitch values, the adjusting may consist in duplicating or deleting frames and is carried out in such a way as to obtain a uniform expansion or reduction and that the stationary vowel frames of the adjusted sequence of frames respectively coincide to the stationary vowel pitch values.

By "stationary vowel frames" it is meant those frames where the energy and the formants remain stable.

Let us suppose that the text message includes the Italian word "volo" in second position, that in the word vocal database it corresponds to the following sequence of 26 frames:

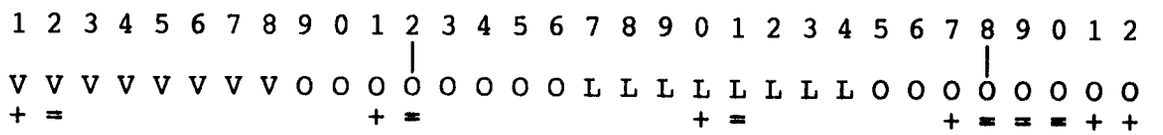
```
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
                  |                       |
V V V V V V V O O O O O O O L L L L L L L 0 0 0 0 0 0
=                 =             =             = = =
```

and that the stationary vowel frames are in position number 10 and 24; let us also suppose that in the word intonation database the intonation curve for a two-syllable word in second position and with the first accented syllable is derived from the pronunciation of the Italian word "molo", corresponding to the following sequence of 32 frames:



having the stationary vowels frames in positions 12 and 28, and is therefore composed of a sequence of 32 pitch values.

In order to adjust the sequence of frames of the word "volo" to the correct length of the word intonation curve the frames marked by the "=" sign are duplicated, thus obtaining the following sequence of 32 frames:



The stationary vowel frames and the length of such adjusted sequence of (32) frames (position 12 and 28) now correspond to those of the word intonation curve.

From the above description of the methods it is clear that the construction of both a suitable word intonation database and a suitable word vocal database is of key importance. Some remarks may be useful to help the person skilled in the art in such construction.

The construction of the word intonation database may comprise the recording of a number of sentences equal to the number of word intonation curves of the database, the extraction of a word from each sentence, and the computation of the pitch value in each frame of the word using a standard technique. If the number of classes considered is 22 and the position considered is 3 the number of sentences is 66.

In case of Italian, preferably the word to be extracted should not be a verb or an adjective but a substantive, should be preceded by a word terminating with a consonant and be followed by a word starting with a plosive consonant. This means that the sentence to be chosen should permit the easy extraction of the intonation curve.

In addition the stationary vowel frames should be identified (one every syllable) and consequently the corresponding stationary vowel pitch values.

The construction of the word vocal database may comprise the recording of a number of sentences equal to the number of words of the necessary vocabulary, the extraction of a word from each sentence, the classification according to the values of the already mentioned three parameters, the generation of the correspondent sequence of frames, the identification of the stationary vowels frames.

The speech synthesis system according to the present invention shall synthesize intonated vocal messages using the speech synthesis method described above.

It comprises the normal HW elements of a standard speech synthesis system but is programmed in such a way as to carry out such method; therefore it may contain, in a program memory, a suitable program and, in a data memory, at least a word intonation database of the kind described above. In addition, if the generation of the pronunciation is not carried out algorithmically, it may comprise further at least a word vocal database of the kind described above.

Claims

1. Method of implementing a message intonation curve for a vocal message corresponding to a text message consisting of the concatenation of words, characterized in that it is carried out independently and separately from the generation of the pronunciation of the vocal message.

2. Method according to claim 1, using a word intonation database including word intonation curves, including the steps of:

   a) determining, for each word of said text message, the value of at least one parameter corresponding to the position of said each word inside said text message, and

   b) extracting, for each word of said text message, from said word intonation database, a word intonation curve corresponding to said each word and to said value;

   whereby said message intonation curve is the concatenation of the word intonation curves so obtained.

3. Method according to claim 2 characterized in that in said step a) the value of a second parameter is determined corresponding to the general intonation of said text message, and in step b) said word intonation curve corresponds to said each word and to the values of two parameters.

4. Method according to claim 1, using a word intonation database including word intonation curves, including the steps of:

a) determining, for each word of said text message, the values of at least three parameters respectively corresponding to the number of syllables, the position of the accented syllable, and the position of said each word inside said text message, and

b) extracting, for each word of said text message, from said word intonation database, a word intonation curve corresponding to said values;

whereby said message intonation curve is the concatenation of the word intonation curves so obtained.

5. Method according to claim 4 characterized in that in said step a) the value of a fourth parameter is determined corresponding to the general intonation of said text message, and in said step b) said word intonation curve corresponds to the values of four parameters.

6. Method according to claim 4 characterized in that a first word intonation curve corresponding to a first word having a number of syllables bigger than a predetermined limit is obtained through the concatenation of at least a second and a third word intonation curves respectively corresponding to at least a second and a third words; the global number of syllables of said second and said third words being equal to the number of syllables of said first word.

7. Method of synthesizing an intonated vocal message corresponding to a text message consisting of the concatenation of words, in which at first the pronunciation of the vocal message is generated, in which, secondly, the intonation of the vocal message is generated independently and separately from its pronunciation, and

in which at last the intonated vocal message is obtained associating said pronunciation and said intonation.

8. Method according to claim 7, in which the intonation of the vocal message is generated according to a message intonation curve obtained through the method according to any of claims 2 to 6.

9. Method according to claim 7, using a word vocal database including for each word a sequence of frames, and using a word intonation database including word intonation curves, each of said word intonation curves being a sequence of pitch values corresponding to a sequence of frames of a word of said word vocal database, and being identified by the value of at least one parameter corresponding to the position of such word inside a sentence,

in which the vocal message is obtained through the extraction of the sequences of frames corresponding to the words of said text message from said word vocal database and the concatenation of said sequences, in which the message intonation curve of said vocal message is obtained through the method according to claims 2 or 3, and

in which the intonated vocal message is obtained associating to each frame of said vocal message a corresponding pitch value of said message intonation curve.

10. Method according to claim 7, using a word vocal database including for each word a sequence of frames and at least two data fields respectively corresponding to the number of syllables and the position of the accented syllable, and using a word intonation database including word intonation curves of sample words, each of said word intonation curves being a sequence of pitch values corresponding to the sequence of frames of the correspondent sample word, and being identified by the values of at least three parameters respectively corresponding to the number of syllables of the correspondent sample word, the position of the accented syllable of the correspondent sample word, and the position of the correspondent sample word inside a sentence,

in which the vocal message is obtained through the extraction of the sequences of frames corresponding to the words of said text message from said word vocal database and the concatenation of said sequences, in which the message intonation curve of said vocal message is obtained through the method according to any of claims 4 to 6, and

in which the intonated vocal message is obtained associating to each frame of said vocal message a

corresponding pitch value of said message intonation curve.

11. Method according to claims 9 or 10, characterized in that said vocal message so obtained is adjusted in such a way that the length of each of said sequences of frames, corresponding to the words of said text message, be equal to the length of the relevant word intonation curve extracted.

12. Method according to claim 11 in which said word vocal database includes for each word a further data field corresponding to stationary vowel frames, in which said word intonation database includes for each word intonation curve a further data field corresponding to stationary vowel pitch values, and in which such adjusting consists in duplicating or deleting frames and is carried out in such a way as to obtain a uniform expansion or reduction, and that said stationary vowel frames of the adjusted sequence of frames respectively coincide with said stationary vowel pitch values.

13. Speech synthesis system for synthesizing an intonated vocal message characterized in that it carries out the method according to any of claims 7 to 12.