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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 February 2004 (05.02.2004)

PCT

(10) International Publication Number
WO 2004/012183 A2

(51) International Patent Classification[7]: G10L 13/00

(21) International Application Number:
PCT/IB2003/002965

(22) International Filing Date: 24 July 2003 (24.07.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
02127007.4 25 July 2002 (25.07.2002) CN

(71) Applicant *(for all designated States except US)*: **MOTOROLA INC** [US/US]; 1303 E.Algonquin Road, Schaumburg, IL 60196 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HUANG, Jian, Cheng** [US/CN]; 168 An Fu Road, 25C, Shanghai 200031 (CN). **CHEN, Fang** [CN/AU]; Unit 3, 59 Hudson St, Hurtsville, New South Wales 2000 (AU).

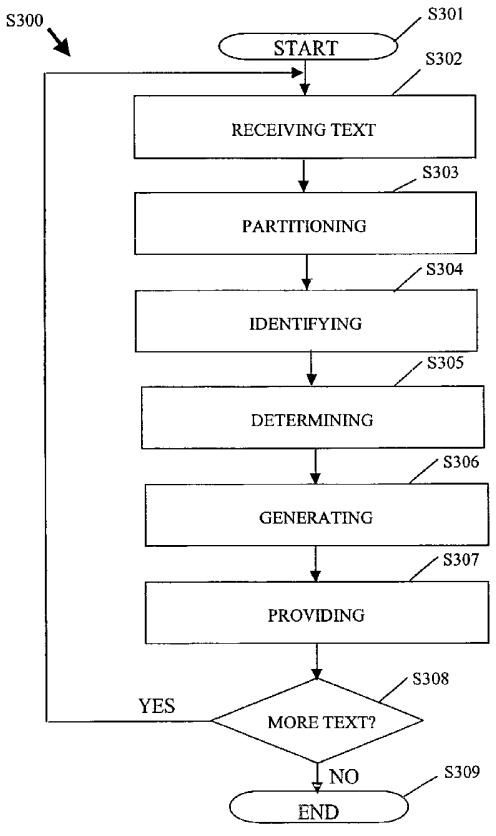
(74) Common Representative: **MOTOROLA INC**; Midpoint, Alencon Link, Basingstoke, Hampshire RG21 7PL (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: CONCATENATIVE TEXT-TO-SPEECH CONVERSION



(57) Abstract: The present invention provides a method for text to speech conversion (S300) including partitioning (S303) text into segmented phonetic units and then identifying (S304) a suitable acoustic unit for each of the phonetic units. Each acoustic unit AU being representative of acoustic segments forming a phonetic cluster determined by their acoustic similarity. The method (S300) then performs determining (S305) variances between prosodic parameters of an acoustic unit AU and each of the phonetic units. A step of generating (S306) acoustic parameters from the prosodic parameters of the acoustic unit and associated variances is then performed and thereafter a step of providing (S307) an output speech signal based on the acoustic parameters is effected. The invention may provide improved synthesized speech quality, system performance and reduced memory overheads suitable for portable devices.

WO 2004/012183 A2

ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

CONCATENATIVE TEXT-TO-SPEECH CONVERSION

FIELD OF THE INVENTION

The present invention relates to a concatenative text-to-speech (TTS) conversion. The invention is particularly useful for, but not necessarily limited to, concatenative TTS synthesis with prosodic control.

BACKGROUND OF THE INVENTION

Reading large volumes of text documents stored in the computers, mobile telephones, or personal data assistants (PDA) may easily cause vision tiredness. And sometimes, reading the data on the electronic screen in a moving vehicle is not convenient. Therefore, it is desired to transform the text documents into speech being played for the reader to listen so as to solve those problems.

At present, almost all high quality text-to-speech TTS synthesis technologies are based on utterance waveform concatenation of each corresponding character, word, or phrase. The desired utterance waveforms are usually derived from an utterance waveform corpus, where the utterance waveform corpus stores various sentences, phrases, and their corresponding utterance waveforms. The quality of desired synthesized utterance depends on the size of such as corpus.

Figure 1 shows an existing typical concatenative TTS system. The system includes three portions, that is, a text processing portion, acoustic segment base, and a speech synthesizer. The system first breaks up sentences and words into word segments, and then it assigns the corresponding characters with phonetic symbols with

assistance of a Lexicon. Then, the sequence of segmented phonetic symbols will be matched with acoustic segments from the utterance or phrase waveform corpus, whereby obtaining the most matched acoustic segments. Finally, the selected acoustic segments will be concatenated with insertion of proper breaks to obtain the output speech.

Such an existing TTS system normally stores the utterance waveform directly. However, in order to obtain the speech effect that is very close to a person's utterance, it would require storing large volumes of utterance waveforms in all kinds of speech environments to cover the speech characteristics of most of situations. The storage of the huge amount of utterance waveform requires lots of memory space. A high quality text-to-speech system requires normally a memory capacity of hundreds of mega bytes. For a hand-held device, such as a mobile telephone or PDA, the memory capacity is usually only few mega bytes due to the limitation of hardware and cost. Therefore, on those portable devices, it is hard to have high quality text-speech. This limits the use of text-to-speech conversion in these technical fields.

SUMMARY OF THE INVENTION

The present invention provides a method for text to speech conversion, the method including: partitioning text into a segmented phonetic units; identifying a suitable acoustic unit for each of the phonetic units, each acoustic unit being representative of acoustic segments forming a phonetic cluster determined by their acoustic similarity; determining variances between prosodic parameters of an acoustic unit and each of the phonetic units; generating acoustic parameters from the prosodic parameters of the acoustic unit and

associated variances to select an acoustic segment ;and

providing an output speech signal based on the acoustic segment.

Suitably, the prosodic parameters includes pitch, duration or energy.

Preferably, the determining is based on position of the acoustic unit in a phrase or a sentence, co-articulation, phrase length or adjacent characters of the acoustic unit.

The partitioning may be characterized by partitioning sentences of text into syllables. Suitably, the phonetic units are syllables. The phonetic units may be assigned a phonetic symbol. Suitably, the phonetic symbol is a pinyin representation.

In another form, there is provided a text-to-speech converting system. The system comprises a text processor for forming a sequence of phonetic symbols after the word segmentation on the basis of input text. The text-to-speech converting system further comprises an acoustic and prosodic controller that includes at least an utterance annotation corpus, and acoustic unit index (AU index) and prosodic vectors (PV) selection device. The utterance annotation corpus includes at least acoustic unit (AU) indices and prosodic vectors (PV). The acoustic unit index (AU index) and prosodic vector (PV) selection device receives the sequence of phonetic symbols after the word segmentation, and generates a series of control data including the acoustic unit (AU) indices and prosodic vectors (PV). The text-to-speech converting system also comprises a synthesizer that includes at least an acoustic parameter base, and the synthesizer responds to the control data from the acoustic/prosodic controller, thereby synthesizing the speech.

The present invention also provides a method of converting a text entry into a corresponding synthetic speech through a concatenative text-to-speech system. The method comprises the steps of processing and converting a text input to generate a sequence of segmented phonetic symbols; searching an utterance annotation corpus including at least acoustic unit (AU) indices to find a maximum match to fetch a matched annotation context; substituting the matched portions of the sequence of segmented phonetic symbols with AU indices and prosodic vector; generating a sequence of control data having at least AU indices and prosodic vectors; and generating a synthetic speech in response to the control data.

The present invention further provides a method of forming a symbolic corpus. The method comprises the steps of slicing utterances into acoustic segments (AS); grouping said AS into clusters in consideration of phonetic classification and acoustic similarity; selecting an acoustic unit (AU) in representation of all acoustic segments in a cluster; converting the acoustic units into respective sequences of parameters frame-by-frame; vector-quantifying the frame parameters of each AU into a sequence of vector indices; forming an AU parameter base containing frame-based scalar parameters and vector indices; finding matched AU for all AS and determining the respective prosodic vectors between AU and AS; and substituting the acoustic segments with the phonetic symbols, AU indices, and prosodic vectors to form an utterance annotation corpus in place of an original AS waveform corpus. In this way, on the basis of the collection of real persons' utterance for the corpus, the present invention groups the utterance or acoustic segments, saves only an acoustic unit as representative of all acoustic segments in a cluster and the difference between the acoustic

segments and the acoustic unit, and uses parameters in representation of the original utterance waveforms, thereby reducing efficiently the amount of data stored in the utterance annotation corpus.

According to the present invention, the phonetic symbols are used to replace any acoustic segments of each cluster, thereby reducing efficiently the number of desired dada of memory and saving the memory space. Besides, the present invention converts each acoustic unit waveform into a series of parameters to form an acoustic unit parameter base, using such parameters in place of the acoustic unit waveform, thereby further reducing the memory space required for storing the acoustic units. The present invention represents the acoustic segments by using the difference between the acoustic units and acoustic segments, and replaces the waveform of the acoustic segments with the phonetic symbols of each acoustic segment and its corresponding acoustic unit parameters and the difference therebetween. This can express utterance information of a syllable corresponding to each acoustic segment, thereby reducing the distortion.

The present invention provides a high efficient text-to-speech converting method and apparatus, and provides the high quality synthetic speech. The required system performance and memory space make it suitable not only for normal computers, but also for small portable devices.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 is an illustration of a prior art text-to-speech conversion system;

Figure 2 is an illustration of the text-to-speech conversion

system in accordance with the present invention; and

Figure 3 is a flow diagram illustrating a method for text to speech conversion in accordance with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT OF THE INVENTION

Referring to Figure 1 there is illustrated a prior art TTS conversion system. The system includes three main portions: a text processor 100, an acoustic segment base 200, and a synthesizer 300. The primary function of the text processor 100 is to have input text normalized and segmented, and then to assign characters of the text with corresponding phonetic symbols. The system uses the obtained phonetic symbol sequence to match with the phonetic symbol sequence stored in the acoustic segment base 200, and then replaces the phonetic symbols corresponding acoustic segments of corresponding utterance or phrases. Finally, the synthesizer 300 concatenates these acoustic segments according to the text with insertion of proper breaks, thereby obtaining the desired speech output. The acoustic segment base 200 stores a huge amount of text content information and utterances of the text content. The more the utterance information, the closer the synthetic speech is to an actual person's utterance. If a sentence of the input text were matched completely and directly with a sentence stored in the acoustic segment base, the waveform of this stored sentence could be used directly for speech output, that is, the actual utterance recorded. However, in most of situations, the system cannot find such a completely matched sentence. In this case, partial matching of words and phrases of this sentence are required and therefore it is necessary to conduct word segmentation.

Then, corresponding acoustic segments are identified to provide TTS conversion.

In Figure 1, the input text is first normalized using a text normalization unit 110. Then, a word segmentation unit 130, guided by a lexicon 120, carries out sentence partitioning, by punctuation identification and word segmentation procedures. After the word segmentation, a phonetic symbol assignment unit 140 and acoustic segment selection unit 250 utilizes an utterance or phrase corpus 260 to search and select acoustic segments in the acoustic segment base 200. The selected segments are sent to a break generation unit 380 and to the acoustic segment concatenation unit 370. The break generation unit 380 generates break information provided to the acoustic segment concatenation unit 370. The acoustic segment concatenation unit 370 concatenates and adds the proper breaks, and outputs the speech signals to the waveform post-processing device. A waveform post-processing unit 390 then outputs synthesized converted speech signals.

For a concatenative TTS converting method or system, the quality of natural pronunciation is dependent on the size of the utterance waveform corpus and selection of appropriate acoustic segments. To save the memory space, the present invention mainly stores parameters of utterance waveforms, and then utilizes these parameters to synthesize the desired speech, thereby reducing the memory storage overheads.

The present invention provides a method of forming utterance annotation corpus. This method comprises the following steps of forming an utterance waveform corpus. It first records a person's utterances whilst reading various texts, and stores these utterances in a

raw utterance waveform corpus. These utterances were chosen carefully to build the raw utterance waveform corpus with a good phonetic and prosodic balance.

The utterance waveforms are partitioned into a plurality of acoustic segments (AS). Each acoustic segment AS corresponds usually to the utterance of a character in a certain language environment. Each acoustic segment is a detailed representation of a syllable or sub-syllable in a particular text, and has a definite phonetic meaning. Usually, the phonetic symbol of each character in different language environment may correspond to many different acoustic segments. The object of acoustic concatenation is to find out desired proper acoustic segment of each character, word, or phrase in detailed language environment, and then concatenates the acoustic segments together.

According to the phonetic classification and acoustic similarity of the acoustic segments AS, the acoustic segments AS are grouped into clusters CR determined by their acoustic similarity. In each cluster CR, one acoustic segment AS, termed an acoustic unit (AU), is selected as a representation of all acoustic segments AS in that cluster CR. All acoustic units AU form an Acoustic Unit Parameter Base 231. In comparison with the prior art, the present invention uses an acoustic unit AU to represent a cluster CR, all other acoustic segments AS in a cluster CR are stored by offset parameters indicating prosodic variances compared to the acoustic segment of that cluster CR. In this regard, there is a relatively small variance between all acoustic segments AS in a cluster CR. Each acoustic unit AU is therefore converted into a sequence of parameters frame-by-frame and stored in the Acoustic Unit Parameter Base 231. Using a frame vector codebook

232, the "frame parameters" of each acoustic unit will be vector-quantified as a sequence of vector indices and acoustic unit parameters. In this case, the acoustic unit indices are used to replace the actual acoustic unit data, thereby reducing the necessary stored data. During the speech concatenation and synthesis, using acoustic unit indices will lead to the vector indices and acoustic unit parameters, and then the vector indices will lead to the frame parameters of the original utterance waveforms. Then, using the frame parameters the original utterance waveforms of a person can be synthesized.

The frames representing the acoustic units AU, where for example in the Chinese language an AU has an implied tone (1 to 5), are stored in the Acoustic Unit Parameter Base 231 in the following format:

Frame_AU_n_(pitch,duration,energy); where in this embodiment pitch has a range of 180 ~ 330(Hz); duration has a range of 165 ~ 452 ms; and energy has a range of 770 ~ 7406 derived from processed and digitized utterances of varying measured RMS (Root Mean Square) power value.

As will be apparent to a person skilled in the art, pitch, energy and duration are prosodic features represented as prosodic vectors or parameters. Hence, an acoustic unit AU for the phonetic or Pinyin "Yu (2)" may be stored as: Frame_AU_51_(254,251,3142); and "Mao (1)" may be stored as Frame_AU_1001_(280,190,2519).

Each acoustic segment AS of each cluster CR of the utterance waveform corpus is mapped with the corresponding acoustic unit indices of the acoustic unit parameter base. Each acoustic segment can be obtained through the acoustic unit AU representing one of the clusters CR of acoustic segments AS.

The prosodic vector between the acoustic segment and its corresponding acoustic unit can be derived. The prosodic vector indicates the difference of parameters between the acoustic segments of each cluster and the acoustic unit representing the cluster. Such parameter difference is based on their difference of physical instance. Therefore, an acoustic segment can be found through the representative acoustic unit and the certain prosodic vector. The utterance annotation corpus is thereby created by the phonetic symbols of each segment, its corresponding acoustic unit indices and its prosodic vector in place of the acoustic segment waveforms.

Referring to Figure 2, a concatenation synthesis of the text-to-speech is explained. The concatenation of text-to-speech includes three main portions: text processing, acoustic and prosodic control, and the speech synthesis. Through the text processing, the input text is converted into phonetic symbols used for acoustic and prosodic control. Through data-driven control, the acoustic and prosodic control portion uses the utterance annotation corpus to match the phonetic symbols to convert them into acoustic unit indices and prosodic vectors, and then through the rule-driven control, the unmatched phonetic symbols from the acoustic annotation corpus will be converted into the desired acoustic unit indices and prosodic vectors. In the speech synthesizer, the obtained acoustic unit indices and prosodic vectors will be converted into frame parameters of the natural utterance waveform through the acoustic unit parameter base and the frame vector codebook, and then concatenated into a synthetic speech.

First, the text processing is briefly explained. Similar to the existing concatenative text-to-speech conversion, the input text of the

present invention is first processed in a text processor 201. Through the text normalization unit 211, input irregular text is classified and converted into a normalized text format of the system. Then, a word segmentation unit 212 divides the normalized text into series of word segments in accordance with a Lexicon 213 and relevant rule base (not shown). After the segmentation, a phonetic symbol assignment unit 214 converts the characters and words of the input text into a sequence of phonetic symbols. When considering the Chinese language, the phonetic symbols would be represented by a Pinyin representation. Thus if a character "鱼" (the Chinese character for Fish) was received at unit 211 then this would be converted into the Pinyin "Yu (2)" at unit 214 where (2) denotes the second tonal pronunciation of Yu.

An acoustic and prosodic controller 202 of the present invention carries out the analysis and process of the obtained sequence of phonetic symbols. The acoustic and prosodic controller 202 comprises an utterance annotation corpus 221, an acoustic unit index and a prosodic vector selection unit 222, a prosodic rule base 223, and a prosodic refinement unit 224. The present invention uses multiple controls of acoustic and prosodic to generate acoustic and prosodic information. The control includes two stages, that is, a data-driven control and a rule-driven control.

In the prior art, for each phonetic symbol of the input text, it must first search the matching acoustic segment in the utterance waveform corpus as an output. The present invention does not use directly the utterance waveform corpus, but the utterance annotation corpus to search the parameters of the matching acoustic segments.

In the data-driven control stage, for the sequence of phonetic symbols obtained from the word segmentation, the acoustic unit index

and prosodic vector selection unit 222 first finds a match from the utterance annotation corpus 221 by utilizing the text relationship or prosodic relationship. A matching phonetic symbol is replaced by the corresponding acoustic unit index and prosodic vector in the utterance annotation corpus. If the matched portion contains one or more breaks, a special acoustic unit representing the break is inserted accordingly such that the parameters of the acoustic unit include the break information.

For an unmatched phonetic symbol during the data-driven stage, an approximate (the closest) sequence in the utterance annotation corpus is used. Alternatively, the rule-driven control stage of the present invention is used to process the unmatched sequence. During this stage, the phonetic symbols are used as a basis, and the unmatched phonetic symbols are determined through the corresponding acoustic unit indices, prosodic vectors, and break acoustic units in accordance with the rules or tables in a prosodic rule base 223.

An output of the acoustic and prosodic controller 202 includes a series of control data reflecting the utterance characteristics of the acoustic unit, and the prosodic vectors and necessary break symbols. For instance, for the Pinyin "Yu" the output control data includes an acoustic unit index of "Frame_AU_51"

The system also has a speech waveform synthesizer 203 that includes the acoustic unit parameter base 231, the frame vector codebook 232, an acoustic unit parameter array generation unit 233, an acoustic unit parameter array modification unit 234, an acoustic segment array concatenation unit 235, and a waveform synthesizer 236.

The speech waveform synthesis of the present invention converts the obtained acoustic unit indices and prosodic vectors into frame parameters of natural utterance waveforms by utilizing the acoustic unit parameter base 231 and frame vector codebook 232, and then concatenates them into speech. The detail procedure is described hereinafter.

Based on the acoustic and prosodic control data output from the acoustic and prosodic controller 202, the speech waveform synthesizer 203 of the present invention forms speech waveform outputs one acoustic segment AS after another acoustic segment AS. For each acoustic segment AS, the speech waveform synthesizer 203 works primarily from three aspects of acoustic unit indices, prosodic vectors and break symbols.

As stated above, the acoustic unit parameter base 231 of the present invention maps the composition of vector index and the frame parameter with an acoustic unit index. This is achieved by using the acoustic unit indices, the vector indices and corresponding scalar parameters can be obtained from the acoustic unit parameter base 231.

In the frame vector codebook 232, a series of vector indices is mapped with the acoustic unit frame parameters and scalar parameters. Therefore, the vector indices and the frame vector codebook obtained from the acoustic parameter base 231 may be used to acquire the frame parameters of the original utterance waveform. Hence, for the Pinyin "Yu" acoustic unit index of "Frame_AU_51_(254, 251,3142) is accessed.

The acoustic unit parameter array generation unit 233 forms a vector array by using the output of the acoustic unit parameter base 231 and the frame vector codebook 232, that is, the array of acoustic unit

parameters. The components of each group of the vector array are the acoustic unit parameters based on the frame. The size of the array depends on the number of frame of the acoustic units. This array of acoustic unit parameters describes completely all of the acoustic characteristics of the acoustic units.

At this point, the acoustic characteristics representative of the acoustic segments are obtained. The desired array of parameters of acoustic segments can be obtained using the difference between the acoustic segment and the acoustic unit on the basis of acoustic characteristic parameters that are prosodic variances represented by a frame format: Frame_AU_51_(offset1,offset2,offset3). In this regard, for each acustic segment in a cluster CR having Frame_AU_51_(254, 251,3142) as its representative acoustic unit AU, offset1 is a variance indicative of pitch, offset2 is a variance indicative of duration, and, offset2 is a variance indicative of energy.

The acoustic unit parameter modification unit 234 is used to accomplish this operation. During the above stated data-driven and rule-driven stages, it is obtained the prosodic vectors between the acoustic segments and the corresponding acoustic unit. The acoustic unit parameter modification device 234 uses the prosodic vectors to modify the output array of the acoustic unit parameter array generation device, thereby obtaining the acoustic segment parameter array. The acoustic segment parameter array is based on the frame to describe the prosodic characteristics of the acoustic segment, and may extend to include lexical tone, pitch contour, duration, and root mean square of amplitude and phonetic/co-articulator environment identity.

The purpose of synthesizing speech is to reproduce the acoustic segments in the utterance waveform corpus, or to generate acoustic

segments by way of low distortion based on the prosodic rule base 223. The acoustic segment parameter array concatenation device 235 concatenates sequentially the frame vector parameters obtained in the acoustic segment parameter array. And when a break symbol (including break information) is detected, a zero vector is inserted thereto. Finally, the arranged frame vector parameters are outputted to the utterance waveform synthesizer 236. The waveform synthesizer 236 uses each frame vector to generate an acoustic segment waveform of a fixed duration, that is, the frame of the acoustic segment. Concatenation of all frames of acoustic waveforms will obtain the desired speech output.

Data-driven in the prior art permits a TTS system to select acoustic and prosodic information from a set of natural utterance. In order to obtain the natural utterance, the existing TTS system uses waveform corpus, and thus requires lots of memory space.

To acquire the natural utterance effect, the present invention also uses data-driven control. The difference is that the present invention does not use directly the waveform corpus of huge memory space, but uses the utterance annotation corpus to save the memory space. In the utterance annotation corpus, only the description of syllables and the acoustic unit base are stored.

Referring to Figure 3, the present invention is further explained. In Figure 3, there is illustrated a method S300 for text to speech conversion implemented on the system of Figure 2. After a start step S301 the method performs a receiving text step S302 that is normalized by the text normalization unit 110. The method S300 then performs partitioning S303 the received text into segmented phonetic units. This is effected by the word segmentation unit 140 and the phonetic units are assigned a phonetic symbol by the phonetic symbol

assignment unit 140. A segmented phonetic unit of text is typically a single phoneme that in Chinese text is a single character such as "鱼". This phonetic unit is assigned the phonetic symbol or Pinyin "Yu (2)" and at an identifying step S304 a suitable acoustic unit for each of the phonetic units is identified. For instance for "Yu (2)" the acoustic unit AU is "Frame_AU_51". A determining step S305 determines variances between the prosodic parameters of the acoustic unit AU identified by Frame_AU_51 and the required prosodic parameters of the acoustic unit for "Yu(2)". This is effected by the rule base 223 and prosodic refinement unit 224 and is based on position of the acoustic unit character for "Yu(2)" in a phrase or sentence of the received text. The prosodic parameters may also be based on co-articulation, phrase length and adjacent characters. The method determines variances by offset values or indexes typically in the following format "Frame_AU_51_(offset1,offset2,offset3)".

After step S305, a generating step S306 generates acoustic parameters from the prosodic parameters of the acoustic unit AU and associated variances (offset1,offset2,offset3). This is achieved by the addressing the AU parameter base 231 with Frame_AU_51 and the codebook 232. The output from the codebook 232 and AU parameter base 231 are combined to generate a vector matrix at an output of unit 233. The unit 234 determines the appropriate acoustic segment AS based on the variances (offset1,offset2,offset3) and the vector matrix for the acoustic unit Frame_AU_51. The selected acoustic segment AS is identified at the output of the unit 234 a concatenative utterance signal results in a providing step S307 for providing an output speech signal based on the acoustic parameters of selected acoustic segment AS. The concatenative utterance signal is based on the selected

acoustic segment (accessing the required speech waveform in the corpus) and break information provided by unit 224. The method then effects a test step S308 to determine if there is any more text to be proceeds and either terminates at a step S309 or returns to step S302.

Advantageously, the present invention provides for allowing a relatively small number of acoustic units representing clusters. This therefore provides for memory overheads.

The detailed description provides a preferred exemplary embodiment only, and is not intended to limit the scope, applicability, or configuration of the invention. Rather, the detailed description of the preferred exemplary embodiment provides those skilled in the art with an enabling description for implementing preferred exemplary embodiment of the invention. It should be understood that various changes may be made in the function and arrangement of elements without departing from the spirit and scope of the invention as set forth in the appended claims.

WE CLAIM:

1. A method for text to speech conversion, the method including:

partitioning text into a segmented phonetic units;

identifying a suitable acoustic unit for each of the phonetic units, each acoustic unit being representative of acoustic segments forming a phonetic cluster determined by their acoustic similarity;

determining variances between prosodic parameters of an acoustic unit and each of the phonetic units;

generating acoustic parameters from the prosodic parameters of the acoustic unit and associated variances to select an acoustic segment; and

providing an output speech signal based on the acoustic segment.

2. A method for text to speech conversion, as claimed in claim 1, wherein the prosodic parameters includes pitch.

3. A method for text to speech conversion, as claimed in claim 1, wherein the prosodic parameters includes duration.

4. A method for text to speech conversion, as claimed in claim 1, wherein the prosodic parameters includes energy.

5. A method for text to speech conversion, as claimed in

claim 1, wherein the determining is based on position of the acoustic unit in a phrase or sentence.

6. A method for text to speech conversion, as claimed in claim 1, wherein the determining is based on co-articulation.

7. A method for text to speech conversion, as claimed in claim 1, wherein the determining is based on phrase length.

8. A method for text to speech conversion, as claimed in claim 1, wherein the determining is based on adjacent characters of the acoustic unit.

9. A method for text to speech conversion, as claimed in claim 1, wherein the partitioning is characterized by partitioning sentences of text into syllables.

10. A method for text to speech conversion, as claimed in claim 1, wherein the phonetic units are syllables.

11. A method for text to speech conversion, as claimed in claim 1, wherein the phonetic units are assigned a phonetic symbol.

12. A method for text to speech conversion, as claimed in claim 11, wherein the phonetic symbol is a pinyin representation.

13. A method for text to speech conversion, as claimed in claim 1, wherein the output speech signal is a selection of concatenated

selected acoustic segments.

14. A method for text to speech conversion, as claimed in claim 11, wherein the output speech signal is a selection of concatenated selected acoustic segments.

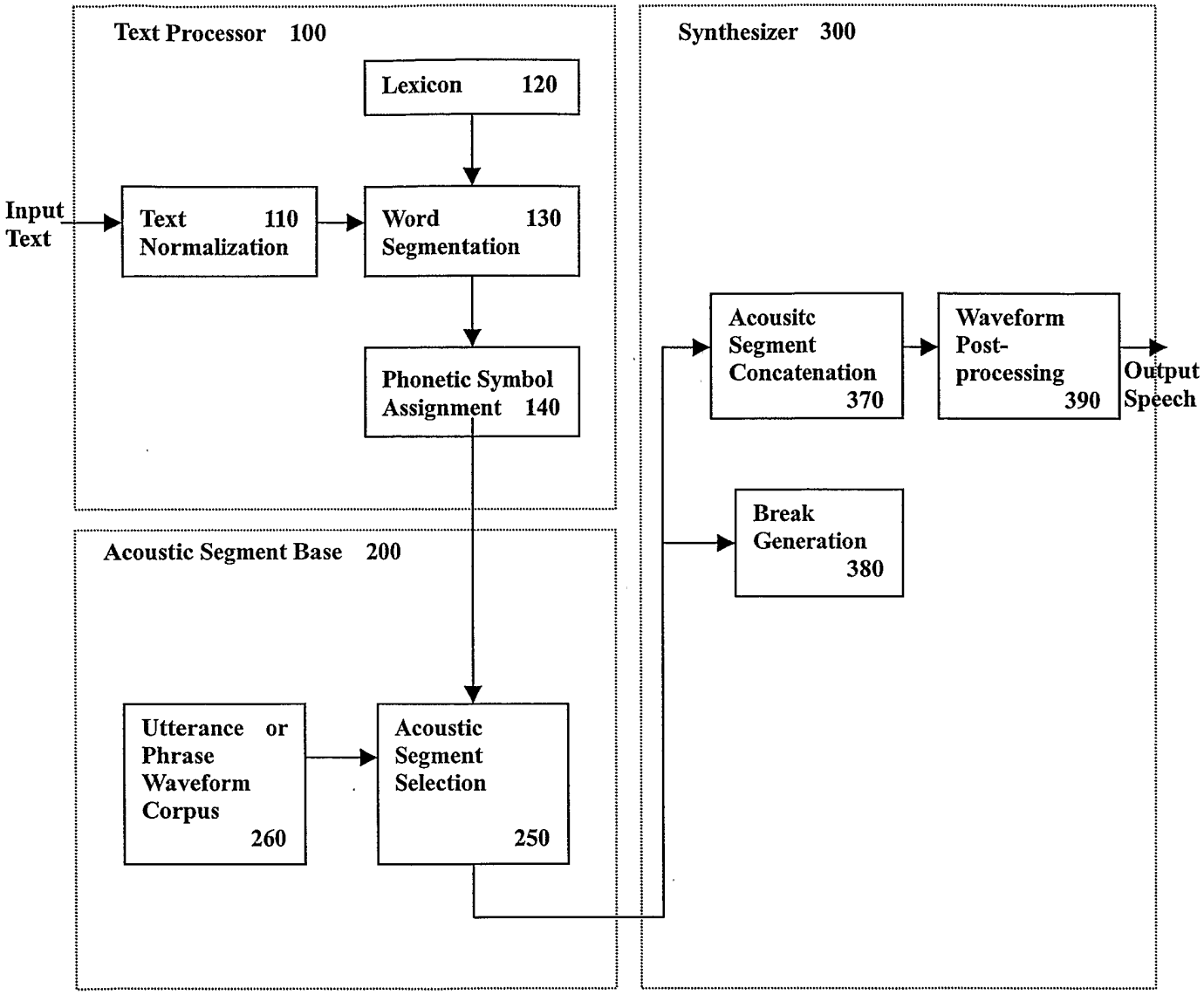


Fig. 1

(PRIOR ART)

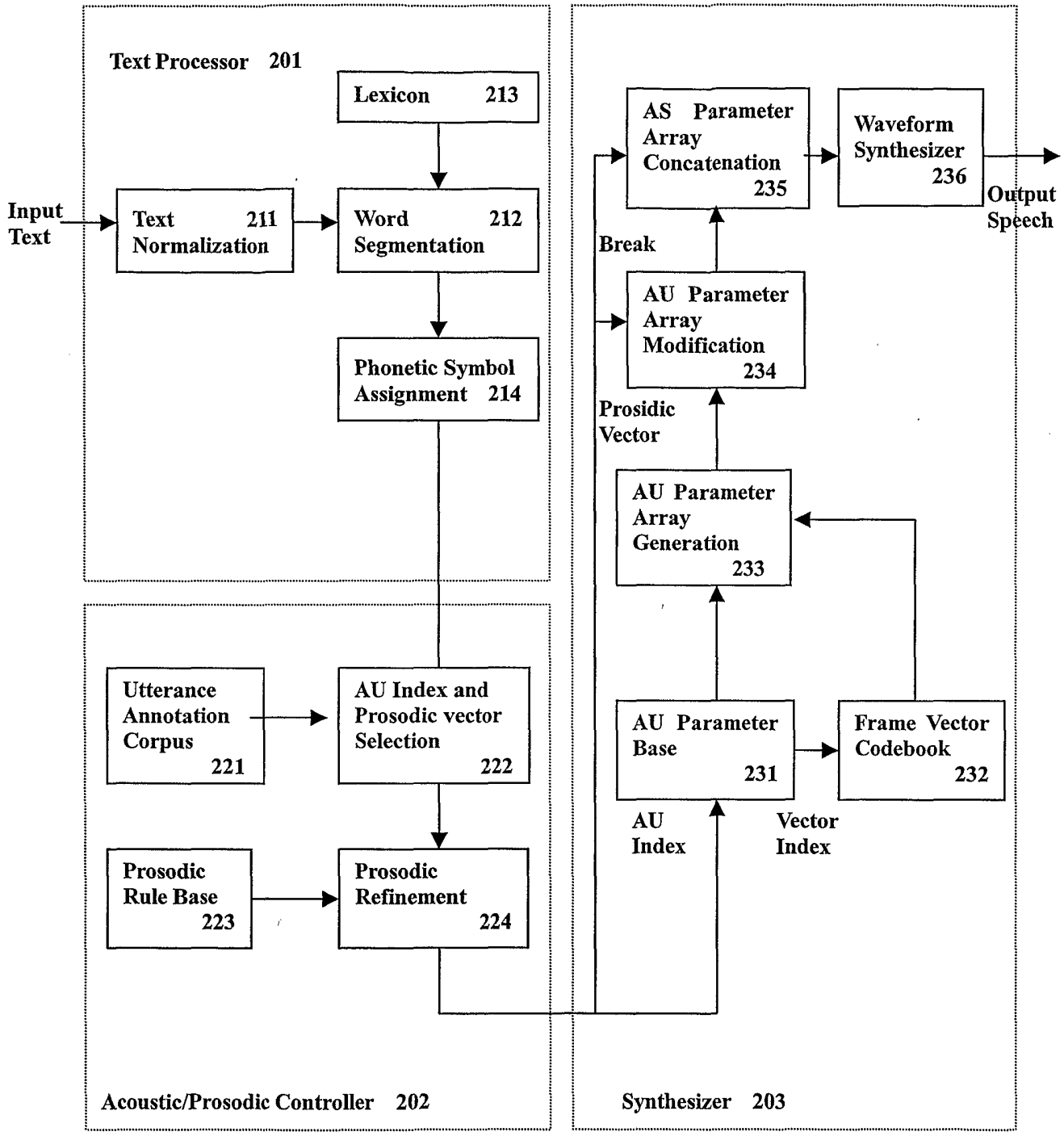

Text Processor 201
Lexicon 213
Input Text
Text Normalization 211
Word Segmentation 212
Phonetic Symbol Assignment 214
Utterance Annotation Corpus 221
AU Index and Prosodic vector Selection 222
Prosodic Rule Base 223
Prosodic Refinement 224
Acoustic/Prosodic Controller 202
AS Parameter Array Concatenation 235
Waveform Synthesizer 236
Output Speech
Break
AU Parameter Array Modification 234
Prosidic Vector
AU Parameter Array Generation 233
AU Parameter Base 231
Frame Vector Codebook 232
AU Index
Vector Index
Synthesizer 203

Fig. 2

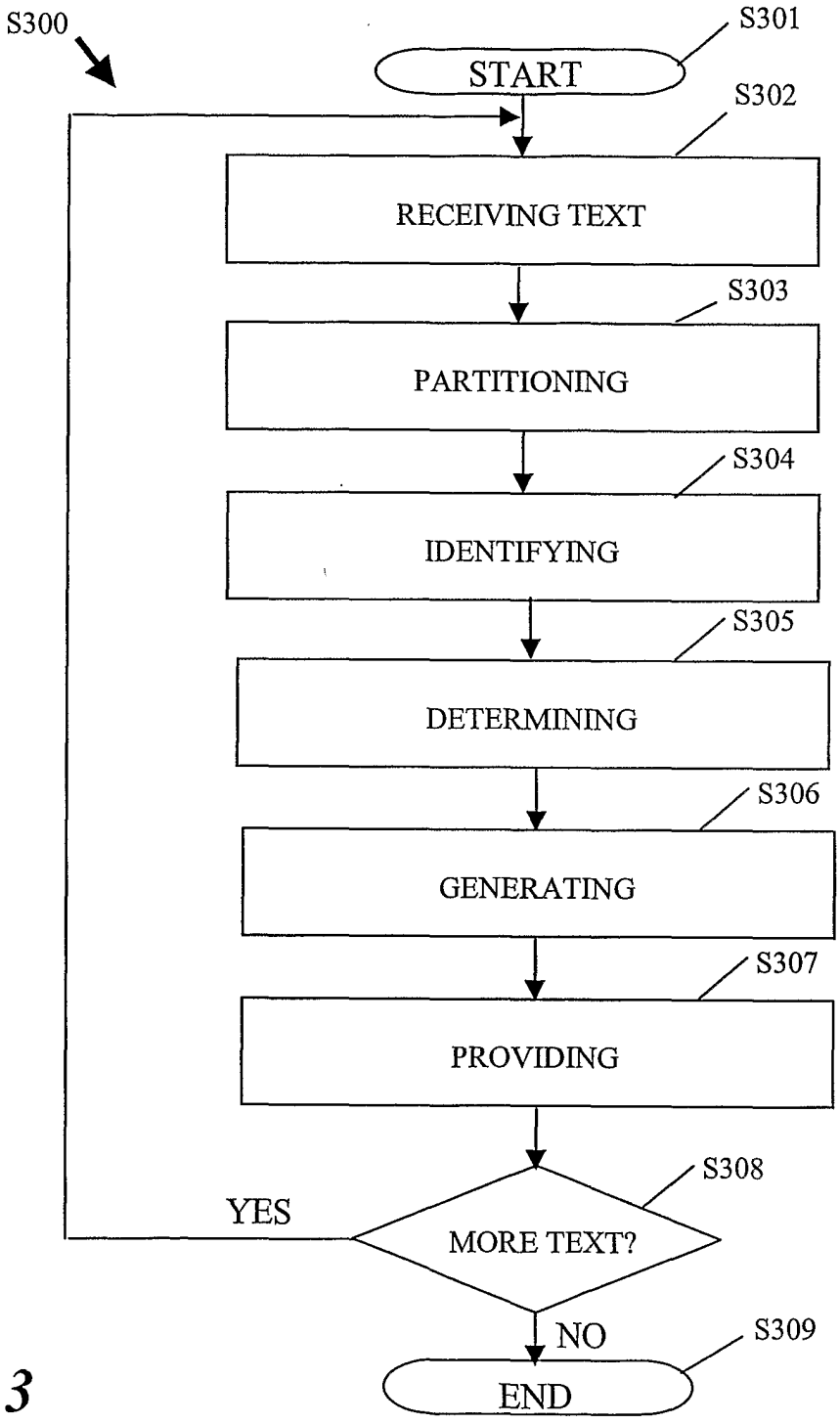

S300
S301
START
S302
RECEIVING TEXT
S303
PARTITIONING
S304
IDENTIFYING
S305
DETERMINING
S306
GENERATING
S307
PROVIDING
S308
YES
MORE TEXT?
NO
S309
END

FIG. 3