

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
4 December 2008 (04.12.2008)

PCT

(10) International Publication Number
WO 2008/147649 A1

(51) International Patent Classification:
G01L 13/06 (2006.01)

(74) Agents: **VAAS, Randall S.**, et al.; 600 North U.S. Highway 45, Libertyville, Illinois 60048 (US).

(21) International Application Number:
PCT/US2008/062822

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(22) International Filing Date: 7 May 2008 (07.05.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
200710104581.3 25 May 2007 (25.05.2007) CN

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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

(72) Inventors; and

(75) Inventors/Applicants (for US only): **ZU, Yi-Qing**, [CN/CN]; 108 Qin Zhou Rd., Room 1406, Building 5, Shanghai 200235 (CN). **CAO, Zhen-Hai**, [CN/CN]; Rm 21-404, LN 160, DongSanLiQiao Rd, Shanghai 200127 (CN).

Published:

- with international search report
- with amended claims

(54) Title: METHOD FOR SYNTHESIZING SPEECH

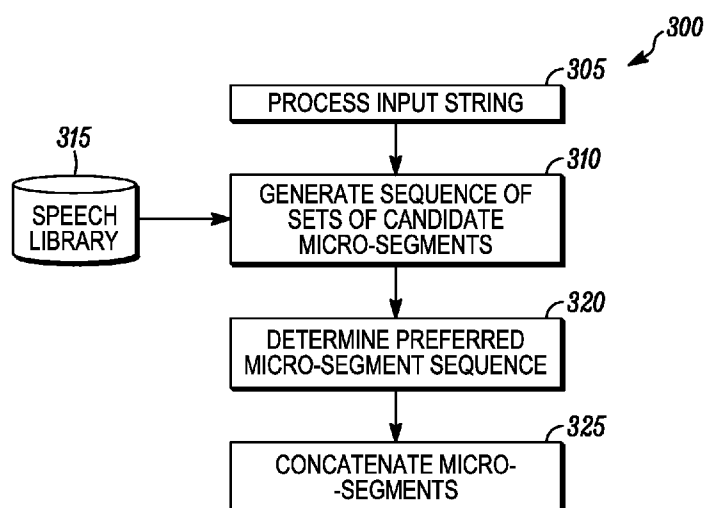


FIG. 3

(57) Abstract: A method for synthesizing speech from an input string enables improved sound quality of Text-To-Speech synthesis. The method includes processing the input string to provide a sequence of acoustic parameters (step 305). For each acoustic parameter in the sequence of acoustic parameters a set of candidate micro-segments is generated from a speech library (step 310). A preferred micro-segment sequence is then determined from the sets of candidate micro-segments for the sequence of acoustic parameters (step 320). Micro-segments in the preferred micro-segment sequence are then concatenated to produce synthesized speech (step 325).

METHOD FOR SYNTHESIZING SPEECH

Field of the Invention

The present invention relates generally to Text-To-Speech (TTS) synthesis
5 and in particular to synthesizing speech from a text string using micro-segments.

Background

Text-To-Speech (TTS) conversion, often referred to as concatenative text
to speech synthesis, enables electronic devices to receive an input text string and
10 provide an audio signal representation of the string in the form of synthesized
speech. For concatenative speech synthesis, basic speech units such as phones or
diphones are concatenated. However, a device that synthesizes speech using
phone-based speech units from a non-deterministic number of received text
strings can have difficulty providing high quality, realistic synthesized speech.
15 That is because the sound of a phone, syllable, or word is often context dependent.

Due to limited memory and processing power in many devices, not all
desired prosodic variations of phones, syllables, or words can be included in a
speech library such as an utterance waveform corpus. For example, although a
phone-based concatenation, such as diphone-to-diphone, might be acceptable for
20 inter-syllable concatenation, the phone-based concatenation of phones within a
syllable may produce unnatural sounds. That is because concatenation points
between voiced-voiced segments often cause unnatural sounding transitions.

A typical diphone speech library for the English language may have
around 1200 diphones, but to reduce the concatenations within voiced-to-voiced
25 phone boundaries, a speech library would require n-phone clusters. Thus a speech
library of all pronunciations of all characters can be prohibitively large. In most
TTS systems there is therefore a need to estimate the appropriate pronunciation of
an input text string based on acoustic analysis using a limited size speech library.

The size of such a speech library can be particularly limited when the speech library is embedded in a handheld electronic device having limited memory capacity.

5

Brief Description of the Figures

In order that the invention may be readily understood and put into practical effect, reference will now be made to exemplary embodiments as illustrated with reference to the accompanying figures, wherein like reference numbers refer to identical or functionally similar elements throughout the separate views. The
10 figures together with a detailed description below, are incorporated in and form part of the specification, and serve to further illustrate the embodiments and explain various principles and advantages, in accordance with the present invention, where:

FIG. 1 is a schematic diagram illustrating an electronic device in the form
15 of a mobile telephone, according to some embodiments of the present invention;

FIG. 2 is a flow diagram illustrating a method for synthesizing speech from an input text string, according to some embodiments of the present invention;

FIG. 3 is a general flow diagram illustrating a method for synthesizing
20 speech from an input string, according to some embodiments of the present invention;

FIG. 4 is a general flow diagram illustrating a method for processing an input string to provide a sequence of acoustic parameters, according to some embodiments of the present invention; and

25 FIG. 5 is a diagram illustrating a pitch model comprising five normalized pitch contour models, according to some embodiments of the present invention.

Skilled artisans will appreciate that elements in the figures are illustrated for simplicity and clarity and have not necessarily been drawn to scale. For

example, the dimensions of some of the elements in the figures may be exaggerated relative to other elements to help to improve understanding of embodiments of the present invention.

5

Detailed Description

Before describing in detail embodiments that are in accordance with the present invention, it should be observed that the embodiments reside primarily in combinations of method steps and apparatus components related to synthesizing speech from an input string. Accordingly, the apparatus components and method
10 steps have been represented where appropriate by conventional symbols in the drawings, showing only those specific details that are pertinent to understanding the embodiments of the present invention so as not to obscure the disclosure with details that will be readily apparent to those of ordinary skill in the art having the benefit of the description herein.

15 In this document, relational terms such as first and second, top and bottom, and the like may be used solely to distinguish one entity or action from another entity or action without necessarily requiring or implying any actual such relationship or order between such entities or actions. The terms “comprises,” “comprising,” or any other variation thereof, are intended to cover a non-
20 exclusive inclusion, such that a process, method, article, or apparatus that comprises a list of elements does not include only those elements but may include other elements not expressly listed or inherent to such process, method, article, or apparatus. An element preceded by “comprises a ...” does not, without more constraints, preclude the existence of additional identical elements in the process,
25 method, article, or apparatus that comprises the element.

Referring to FIG. 1, a schematic diagram illustrates an electronic device in the form of a mobile telephone 100, according to some embodiments of the present invention. The mobile telephone 100 comprises a radio frequency communications unit 102 coupled to be in communication with a common data

and address bus 117 of a processor 103. The telephone 100 also has a keypad 106 and a display screen 105, such as a touch screen coupled to be in communication with the processor 103.

The processor 103 also includes an encoder/decoder 111 with an
5 associated code Read Only Memory (ROM) 112 for storing data for encoding and decoding voice or other signals that may be transmitted or received by the mobile telephone 100. The processor 103 further includes a microprocessor 113 coupled, by the common data and address bus 117, to the encoder/decoder 111, a character Read Only Memory (ROM) 114, a Random Access Memory (RAM) 104,
10 programmable memory 116 and a Subscriber Identity Module (SIM) interface 118. The programmable memory 116 and a SIM operatively coupled to the SIM interface 118 each can store, among other things, a telephone number database (TND) comprising a number field for telephone numbers and a name field for identifiers uniquely associated with the telephone numbers in the number field.

15 The radio frequency communications unit 102 is a combined receiver and transmitter having a common antenna 107. The communications unit 102 has a transceiver 108 coupled to the antenna 107 via a radio frequency amplifier 109. The transceiver 108 is also coupled to a combined modulator/demodulator 110 that is coupled to the encoder/decoder 111.

20 The microprocessor 113 has ports for coupling to the keypad 106 and to the display screen 105. The microprocessor 113 further has ports for coupling to an alert module 115 that typically contains an alert speaker, vibrator motor and associated drivers; to a microphone 120; and to a communications speaker 122. The character ROM 114 stores code for decoding or encoding data such as control
25 channel messages that may be transmitted or received by the communications unit 102. In some embodiments of the present invention, the character ROM 114, the programmable memory 116, or a SIM also can store operating code (OC) for the microprocessor 113 and code for performing functions associated with the mobile telephone 100. For example, the programmable memory 116 can comprise

speech synthesis services program code components 125 configured to cause execution of a method for synthesizing speech from an input string.

Thus some embodiments of the present invention include a method of using the mobile telephone 100 for synthesizing speech from an input string. For example, an input string can be a text message or an email containing a text string received at the mobile telephone 100. The method includes processing the input string to provide a sequence of acoustic parameters. A sequence of sets of candidate micro-segments is then generated from a speech library using the sequence of acoustic parameters. A preferred micro-segment sequence is then determined from the sequence of sets of candidate micro-segments for the sequence of acoustic parameters. Finally, micro-segments in the preferred micro-segment sequence are then concatenated to produce synthesized speech.

Some embodiments of the present invention therefore enable speech synthesis using micro-segments and a sequence of acoustic parameters representing a target acoustic model, rather than using phones or diphones. A micro-segment can be any length of a speech segment, but is generally shorter than a phone or diphone. For example, a micro-segment can be a 20ms speech frame, whereas a speech segment of a phone usually comprises several such speech frames. Since speech segments synthesized by concatenating micro-segments can provide more frequency and prosodic variations than speech segments synthesized by concatenating phones or diphones, overall sound quality of Text To Speech (TTS) systems can be improved.

Referring to FIG. 2, a flow diagram illustrates a method 200 for synthesizing speech from an input string 205, according to some embodiments of the present invention. First, the input string 205 is processed to provide a sequence 230 of acoustic parameters. A sequence 240 of sets 235 of candidate micro-segments is then generated from a speech library using the sequence 230 of acoustic parameters. A preferred micro-segment sequence 245 is then determined from the sequence 240 of the sets 235 of candidate micro-segments for the

sequence 230 of acoustic parameters. Finally, micro-segments in the preferred micro-segment sequence 245 are then concatenated to produce a synthesized speech signal 250. For example, speech frames 255 corresponding to the micro-segment descriptions in the preferred micro-segment sequence 245 can be loaded
5 into the RAM 104 of the mobile telephone 100, and then concatenated and played over the communications speaker 122 to produce the synthesized speech signal 250.

Referring to FIG. 3, a flow diagram further illustrates a general method 300 for synthesizing speech from an input string, according to some embodiments
10 of the present invention. At step 305, the input string is processed to provide a sequence of acoustic parameters. For example, an acoustic parameter of the sequence 230 of acoustic parameters can comprise a spectrum parameter, a pitch parameter, and an energy parameter.

According to some embodiments of the present invention, the acoustic
15 parameters, also referred to as target speech units, are generated from the input string using prosodic positions. For example, a prosodic position can comprise a position of a syllable in a word and a position of the word in a sentence.

A spectrum parameter can be modeled using a known spectral feature representation method, such as a Linear Predictive Coding (LPC) method, a Line
20 Spectral Pairs (LSP) method, or a Mel-Frequency Cepstral Coefficient (MFCC) method. Therefore, using prosodic positions, spectrum parameters of phones can be determined. For example, a positional spectrum model, such as a Gaussian Mixture Model (GMM), can be used to map acoustic features of phones, such as prosodic position, to spectrum parameters. The pitch parameter can be
25 determined using a pitch model that defines pitch contours of syllables based on prosodic positions of the syllables. The pitch model can comprise pitch contour models, such as WO_stress, WO_unstress, WF_stress, WF_unstress, or WS.

For the energy parameter, different strategies can be used for a voiced part and an un-voiced part of a syllable. For a voiced part, energy contour patterns can

be defined for syllables. Different energy contour patterns can be defined using a position of a cv-like unit in the syllable and/or a condition concerning whether the syllable is a stress syllable. For an un-voiced part, energy contour patterns can be defined for phonemes. Each (un-voiced) phoneme can have one or more energy
5 contour patterns. An energy contour of an un-voiced phoneme can depend on the position of the phoneme in a syllable and the position of the syllable in a word. To reduce an amount of memory required, some (un-voiced) phonemes can share a same energy contour pattern if the phonemes have similar positions and similar articulation manner. For example, phonemes “s”, “sh”, and “ch” can share a same
10 energy contour pattern, and similarly, “g”, “d”, and “k” can share another same energy contour pattern.

At step 310, a sequence of sets of candidate micro-segments is generated from a speech library 315 using the sequence of acoustic parameters. According to some embodiments of the present invention, the sets of the candidate micro-
15 segments can be generated using a target cost function and a duration model. For example, the target cost function can be a weighted sum of a spectrum cost, a pitch cost, and an energy cost. Lower target cost can mean that acoustic characteristics of a candidate micro-segment closely match an acoustic parameter. For example, for each acoustic parameter in the sequence 230 of acoustic
20 parameters, the mobile telephone 100 can search the speech library 315 to find a set of candidate micro-segments (e.g., speech frames) having acoustic characteristics that closely match the acoustic parameter and an estimated duration of the acoustic parameter. Such closely matching speech frames then can be selected to generate the sequence 240 of the sets 235 of candidate micro-segments.

25 In order to reduce processing time, speech frames in the speech library 315 can be classified into several sets of speech frames using prosodic positions of the speech frames, and the candidate micro-segments can be searched in one of the sets of speech frames that closely matches a prosodic position of the acoustic parameter.

At step 320, a preferred micro-segment sequence is determined from the sets of candidate micro-segments for the sequence of acoustic parameters. For example, a Viterbi algorithm can be used to determine the preferred micro-segment sequence 245, and a path cost function of the Viterbi algorithm can be a
 5 sum of a target cost function and a concatenation cost function.

According to some embodiments of the present invention, the target cost function can be a weighted sum of a spectrum cost function, a pitch cost function, and an energy cost function. For example, the spectrum cost function can be a measure of a degree of difference in spectral features between a candidate micro-
 10 segment and an acoustic parameter, also referred to as a target micro-segment, in the sequence 230 of acoustic parameters. Similarly, the pitch cost function and the energy cost function can measure degrees of difference in pitch and energy features between an acoustic parameter and a candidate micro-segment, respectively. For example, the target cost function can be defined as follows:

$$15 \quad C^T(u_{i,k}) = K_S^T C_S^T(u_{i,k}) + K_P^T C_P^T(u_{i,k}) + K_E^T C_E^T(u_{i,k}) \quad (\text{Eq. 1})$$

where $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence 230 of acoustic parameters, $C^T(u_{i,k})$ is the target cost function, $C_S^T(u_{i,k})$ is a spectrum cost function, $C_P^T(u_{i,k})$ is a pitch cost function, $C_E^T(u_{i,k})$ is an energy cost function, and K_S^T , K_P^T , and K_E^T are weight values.

20 The concatenation cost function can be a weighted sum of a spectrum difference function, a pitch difference function, and an energy difference function. The spectrum difference function can measure a degree of difference in spectral features between two adjacent micro-segments. Similarly, the pitch difference function and the energy difference function can measure degrees of difference in
 25 pitch and energy features between two adjacent micro-segments, respectively. For example, the concatenation cost function can be defined as follows:

$$C^C(u_{i-1,j}, u_{i,k}) = K_S^C C_S^C(u_{i-1,j}, u_{i,k}) + K_P^C C_P^C(u_{i-1,j}, u_{i,k}) + K_E^C C_E^C(u_{i-1,j}, u_{i,k}) \quad (\text{Eq. 2})$$

where $u_{i-1,j}$ is a j -th candidate micro-segment of an $(i-1)$ -th acoustic parameter in

the sequence 230 of acoustic parameters, $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence 230 of acoustic parameters, $C^C(u_{i-1,j}, u_{i,k})$ is the concatenation cost, $C^S(u_{i-1,j}, u_{i,k})$ is a spectrum difference function between $u_{i-1,j}$ and $u_{i,k}$, $C^P(u_{i-1,j}, u_{i,k})$ is a pitch difference function between $u_{i-1,j}$ and $u_{i,k}$, $C^E(u_{i-1,j}, u_{i,k})$ is an energy difference function between $u_{i-1,j}$ and $u_{i,k}$, and K^C_S , K^C_P , and K^C_E are weight values.

At step 325, micro-segments in the preferred micro-segment sequence are then concatenated to produce synthesized speech.

Referring to FIG. 4, a general flow diagram illustrates sub-steps of the step 305 of the method 300 of processing the input string to provide the sequence of acoustic parameters, according to some embodiments of the present invention. At step 405, the input string is processed to provide a phoneme sequence. For example, the input string 205 can be a text message or an email message received at the mobile telephone 100, and the phoneme sequence can be a string representing pronunciation of the text message in a phonetic alphabet.

At step 410, syllable boundaries are then determined in the phoneme sequence to provide a syllable sequence. For example, an English word can comprise several syllables, and boundaries of such syllables in the word then can be determined to provide the syllable sequence. For example, a phoneme sequence “ihksplehn” concerning the English word “explain” can be divided into a syllable sequence comprising two syllables, such as “ihk” and “splehn”.

At step 415, sub-syllable units are then identified in the syllable sequence to provide a sub-syllable sequence. The sub-syllable units can be equal or smaller than syllables, and can be cv-like speech units, which can comprise a consonant and a vowel. Thus, the sub-syllable sequence can comprise cv-like speech units and consonants. For example, two cv-like speech units (“ih” and “lehn”) can be identified in a syllable sequence (“ihk” + “splehn”). A corresponding sub-syllable sequence then can be (“ih” + “k” + “s” + “p” + “lehn”).

According to some embodiments of the present invention, using cv-like speech units to represent pronunciation of an input text can reduce a number of basic units needed to describe words. For example, a lexicon comprising 202,000 words may comprise 24,980 syllables, and only 6,707 cv-like units.

- 5 At step 420, the sub-syllable sequence is then processed to provide a micro-segment description sequence. For example, by estimating a duration of each element in the sub-syllable sequence using a duration model, a number of micro-segments needed to synthesize speech for each element can be estimated. For example, consider the following cv-like speech unit (a sub-syllable): ih.
- 10 If an estimated duration of the cv-like speech unit is approximately equal to five micro-segments, the sub-syllable can be mapped to five micro-segment descriptions as follows: $ih_f ih_f ih_f ih_f ih_f$, where ih_f is a micro-segment description.

- According to some embodiments of the present invention, an estimated
- 15 duration of a sub-syllable can be obtained by applying a duration model comprising average durations of phones and prosodic attributes of phones. For example, an estimated duration of a phone p can be obtained according to the following equation:

$$L_p = k \times L_{avg} \quad (\text{Eq. 3})$$

- 20 where L_p is the estimated duration of the phone p , L_{avg} is an average phone duration of the phone p , and k is a prosodic attribute coefficient obtained from factors comprising a number of phones in a syllable containing the phone p , a number of syllables in a word containing the syllable, and a type of the phone p .

- At step 425, the micro-segment description sequence is then processed to
- 25 provide the sequence of acoustic parameters. For example, each micro-segment description in the micro-segment description sequence can be mapped to an acoustic parameter describing the micro-segment description's acoustic characteristics such as spectrum (frequency characteristics) and prosodic characteristics (pitch, energy, or duration). The micro-segment description

sequence can comprise micro-segment descriptions, each of which is a description of a speech micro-segment that is usually smaller than a phone. For each micro-segment description in the micro-segment description sequence, an acoustic parameter can be estimated using an acoustic model. For example, the acoustic parameter can comprise a spectrum parameter s_n , a pitch parameter p_n , and an energy parameter e_n .

Referring to FIG. 5, a diagram illustrates a pitch model comprising five normalized pitch contour models: WO_stress 505, WO_unstress 510, WF_stress 515, WF_unstress 520, and WS 525, as used according to some embodiments of the present invention. The WO_stress 505 pitch contour model defines a pitch contour for stressed syllables positioned at a beginning or middle of words having multiple syllables. The WO_unstress 510 pitch contour model defines a pitch contour for unstressed syllables positioned at a beginning or middle of words having multiple syllables. The WF_stress 515 pitch contour model defines a pitch contour for stressed syllables positioned at an end of words having multiple syllables. The WF_unstress 520 pitch contour model defines a pitch contour for unstressed syllables positioned at an end of words having multiple syllables. The WS 525 pitch contour model defines a pitch contour for syllables in words having only one syllable.

Advantages of some embodiments of the present invention therefore include improved sound quality of synthesized speech. Speech segments synthesized by concatenating micro-segments can provide improved speech continuity and more prosodic variations than speech segments synthesized by concatenating phones or diphones. Overall sound quality of TTS systems therefore can be improved, particularly in resource constrained handheld devices such as mobile telephones and personal digital assistants (PDAs).

It will be appreciated that embodiments of the invention described herein may be comprised of one or more conventional processors and unique stored program instructions that control the one or more processors to implement, in

conjunction with certain non-processor circuits, some, most, or all of the functions of synthesizing speech from an input string as described herein. The non-processor circuits may include, but are not limited to, a radio receiver, a radio transmitter, signal drivers, clock circuits, power source circuits, and user input devices. As such, these functions may be interpreted as steps of a method for synthesizing speech from an input string. Alternatively, some or all functions could be implemented by a state machine that has no stored program instructions, or in one or more application specific integrated circuits (ASICs), in which each function or some combinations of certain of the functions are implemented as custom logic. Of course, a combination of the two approaches could be used. Thus, methods and means for these functions have been described herein. Further, it is expected that one of ordinary skill, notwithstanding possibly significant effort and many design choices motivated by, for example, available time, current technology, and economic considerations, when guided by the concepts and principles disclosed herein will be readily capable of generating such software instructions and programs and ICs with minimal experimentation.

In the foregoing specification, specific embodiments of the present invention have been described. However, one of ordinary skill in the art appreciates that various modifications and changes can be made without departing from the scope of the present invention as set forth in the claims below. Accordingly, the specification and figures are to be regarded in an illustrative rather than a restrictive sense, and all such modifications are intended to be included within the scope of present invention. The benefits, advantages, solutions to problems, and any element(s) that may cause any benefit, advantage, or solution to occur or become more pronounced are not to be construed as critical, required, or essential features or elements of any or all the claims. The invention is defined solely by the appended claims including any amendments made during the pendency of this application and all equivalents of those claims.

We claim:

1. A method for synthesizing speech from an input string, the method comprising:
 - processing the input string to provide a sequence of acoustic parameters;
 - 5 generating from a speech library a sequence of sets of candidate micro-segments using the sequence of acoustic parameters;
 - determining a preferred micro-segment sequence from the sequence of sets of candidate micro-segments for the sequence of acoustic parameters; and
 - concatenating micro-segments in the preferred micro-segment sequence to
10 produce synthesized speech.
2. The method of claim 1, wherein processing the input string to provide the sequence of acoustic parameters comprises:
 - processing the input string to provide a phoneme sequence;
 - 15 determine syllable boundaries in the phoneme sequence to provide a syllable sequence;
 - identifying sub-syllable units in the syllable sequence to provide a sub-syllable sequence;
 - generating a micro-segment description sequence from the sub-syllable
20 sequence; and
 - processing the micro-segment description sequence to provide the sequence of acoustic parameters.
3. The method of claim 2, wherein the micro-segment description sequence
25 is generated from the sub-syllable sequence using a duration model comprising average durations of phones and prosodic attributes of phones.

4. The method of claim 2, wherein the sub-syllable sequence comprises one or more of a cv-like speech unit or a phone.
- 5 5. The method of claim 1, wherein an acoustic parameter of the sequence of acoustic parameters comprises a spectrum parameter, a pitch parameter, and an energy parameter.
6. The method of claim 1, wherein sets of candidate micro-segments are
10 selected from the speech library using a target cost function and a duration model.
7. The method of claim 6, wherein the target cost function is a weighted sum of a spectrum cost, a pitch cost, and an energy cost.
- 15 8. The method of claim 1, wherein the preferred micro-segment sequence is determined from the sets of candidate micro-segments for the sequence of acoustic parameters using a Viterbi algorithm.
9. The method of claim 8, wherein the Viterbi algorithm comprises a path
20 cost function that is a sum of a target cost function and a concatenation cost function.
10. The method of claim 9, wherein the target cost function is a weighted sum of a spectrum cost function, a pitch cost function, and an energy cost function.

25

11. The method of claim 9, wherein the concatenation cost function is a weighted sum of a spectrum difference function, a pitch difference function, and an energy difference function.

5 12. The method of claim 10, wherein the target cost function is defined as follows:

$$C^T(u_{i,k}) = K_S^T C_S^T(u_{i,k}) + K_P^T C_P^T(u_{i,k}) + K_E^T C_E^T(u_{i,k})$$

where $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence of acoustic parameters, $C^T(u_{i,k})$ is the target cost function, $C_S^T(u_{i,k})$ is a spectrum cost function, $C_P^T(u_{i,k})$ is a pitch cost function, $C_E^T(u_{i,k})$ is an energy cost function, and K_S^T , K_P^T , and K_E^T are weight values.

13. The method of claim 11, wherein the concatenation cost function is defined according to the following equation:

$$15 \quad C^C(u_{i-1,j}, u_{i,k}) = K_S^C C_S^C(u_{i-1,j}, u_{i,k}) + K_P^C C_P^C(u_{i-1,j}, u_{i,k}) + K_E^C C_E^C(u_{i-1,j}, u_{i,k})$$

where $u_{i-1,j}$ is a j -th candidate micro-segment of an $(i-1)$ -th acoustic parameter in the sequence of acoustic parameters, $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence of acoustic parameters, $C^C(u_{i-1,j}, u_{i,k})$ is the concatenation cost for concatenating $u_{i-1,j}$ and $u_{i,k}$, $C_S^C(u_{i-1,j}, u_{i,k})$ is a spectrum difference function between $u_{i-1,j}$ and $u_{i,k}$, $C_P^C(u_{i-1,j}, u_{i,k})$ is a pitch difference function between $u_{i-1,j}$ and $u_{i,k}$, $C_E^C(u_{i-1,j}, u_{i,k})$ is an energy difference function between $u_{i-1,j}$ and $u_{i,k}$, and K_S^C , K_P^C , and K_E^C are weight values.

14. The method of claim 5, wherein the pitch parameter is one of following pitch models: WO_stress, WO_unstress, WF_stress, WF_unstress, or WS.

15. The method of claim 5, wherein the energy parameter comprises a voiced part and an un-voiced part.

- 5 16. The method of claim 3, wherein the duration model is defined by the following equation:

$$L_p = k \times L_{avg}$$

where L_p is an estimated duration of a phone p , L_{avg} is an average phone duration of the phone p , and k is a prosodic attribute coefficient obtained from factors

- 10 comprising a number of phones in a syllable containing the phone p , a number of syllables in a word containing the phone p , and a type of the phone p .

AMENDED CLAIMS

received by the International Bureau on 11 September 2008 (11.09.08)

1. A method for synthesizing speech from an input string, the method
5 comprising:
 - processing the input string to provide a phoneme sequence;
 - determining syllable boundaries in the phoneme sequence to provide a syllable
sequence;
 - identifying sub-syllable units in the syllable sequence to provide a sub-syllable
10 sequence;
 - generating a micro-segment description sequence from the sub-syllable
sequence; and
 - processing the micro-segment description sequence to provide a sequence of
acoustic parameters;
 - 15 generating from a speech library a sequence of sets of candidate micro-
segments using the sequence of acoustic parameters, each of the candidate micro-
segments comprises speech acoustic parameters;
 - determining a preferred micro-segment sequence from the sequence of sets of
candidate micro-segments for the sequence of acoustic parameters; and
 - 20 concatenating micro-segments in the preferred micro-segment sequence to
produce synthesized speech.
2. The method of claim 1, wherein the micro-segment description sequence is
generated from the sub-syllable sequence using a duration model comprising average
25 durations of phones and prosodic attributes of phones.
3. The method of claim 1, wherein the sub-syllable sequence comprises one or
more of a cv-like speech unit or a phone.

4. The method of claim 1, wherein an acoustic parameter of the sequence of acoustic parameters comprises a spectrum parameter, a pitch parameter, and an energy parameter.

5

5. The method of claim 1, wherein sets of candidate micro-segments are selected from the speech library using a target cost function and a duration model.

6. The method of claim 5, wherein the target cost function is a weighted sum of a spectrum cost, a pitch cost, and an energy cost.

10

7. The method of claim 1, wherein the preferred micro-segment sequence is determined from the sets of candidate micro-segments for the sequence of acoustic parameters using a Viterbi algorithm.

15

8. The method of claim 7, wherein the Viterbi algorithm comprises a path cost function that is a sum of a target cost function and a concatenation cost function.

9. The method of claim 8, wherein the target cost function is a weighted sum of a spectrum cost function, a pitch cost function, and an energy cost function.

20

10. The method of claim 8, wherein the concatenation cost function is a weighted sum of a spectrum difference function, a pitch difference function, and an energy difference function.

25

11. The method of claim 9, wherein the target cost function is defined as follows:

$$C^T(u_{i,k}) = K_S^T C_S^T(u_{i,k}) + K_P^T C_P^T(u_{i,k}) + K_E^T C_E^T(u_{i,k})$$

where $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence of acoustic parameters, $C^T(u_{i,k})$ is the target cost function, $C_S^T(u_{i,k})$ is a

5 spectrum cost function, $C_P^T(u_{i,k})$ is a pitch cost function, $C_E^T(u_{i,k})$ is an energy cost function, and K_S^T , K_P^T , and K_E^T are weight values.

12. The method of claim 10, wherein the concatenation cost function is defined according to the following equation:

$$10 \quad C^C(u_{i-1,j}, u_{i,k}) = K_S^C C_S^C(u_{i-1,j}, u_{i,k}) + K_P^C C_P^C(u_{i-1,j}, u_{i,k}) + K_E^C C_E^C(u_{i-1,j}, u_{i,k})$$

where $u_{i-1,j}$ is a j -th candidate micro-segment of an $(i-1)$ -th acoustic parameter in the sequence of acoustic parameters, $u_{i,k}$ is a k -th candidate micro-segment of an i -th acoustic parameter in the sequence of acoustic parameters, $C^C(u_{i-1,j}, u_{i,k})$ is the

15 concatenation cost for concatenating $u_{i-1,j}$ and $u_{i,k}$, $C_S^C(u_{i-1,j}, u_{i,k})$ is a spectrum difference function between $u_{i-1,j}$ and $u_{i,k}$, $C_P^C(u_{i-1,j}, u_{i,k})$ is a pitch difference function between $u_{i-1,j}$ and $u_{i,k}$, $C_E^C(u_{i-1,j}, u_{i,k})$ is an energy difference function between $u_{i-1,j}$ and $u_{i,k}$, and K_S^C , K_P^C , and K_E^C are weight values.

13. The method of claim 4, wherein the pitch parameter is one of following pitch models: WO_stress, WO_unstress, WF_stress, WF_unstress, or WS.

20

14. The method of claim 4, wherein the energy parameter comprises a voiced part and an un-voiced part.

15. The method of claim 2, wherein the duration model is defined by the following equation:

$$L_p = k \times L_{avg}$$

where L_p is an estimated duration of a phone p , L_{avg} is an average phone
5 duration of the phone p , and k is a prosodic attribute coefficient obtained from factors comprising a number of phones in a syllable containing the phone p , a number of syllables in a word containing the phone p , and a type of the phone p .

10

1/4

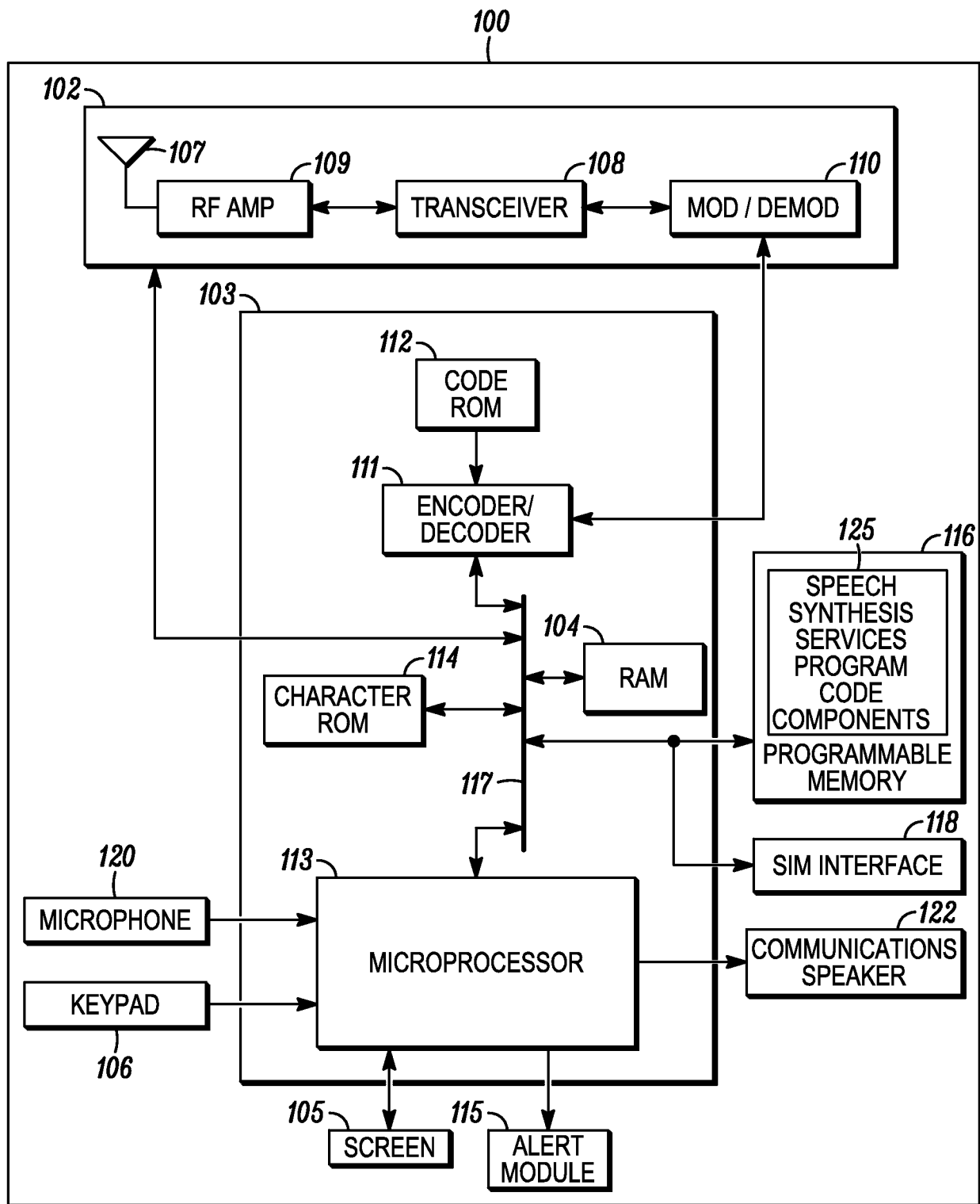


FIG. 1

2/4

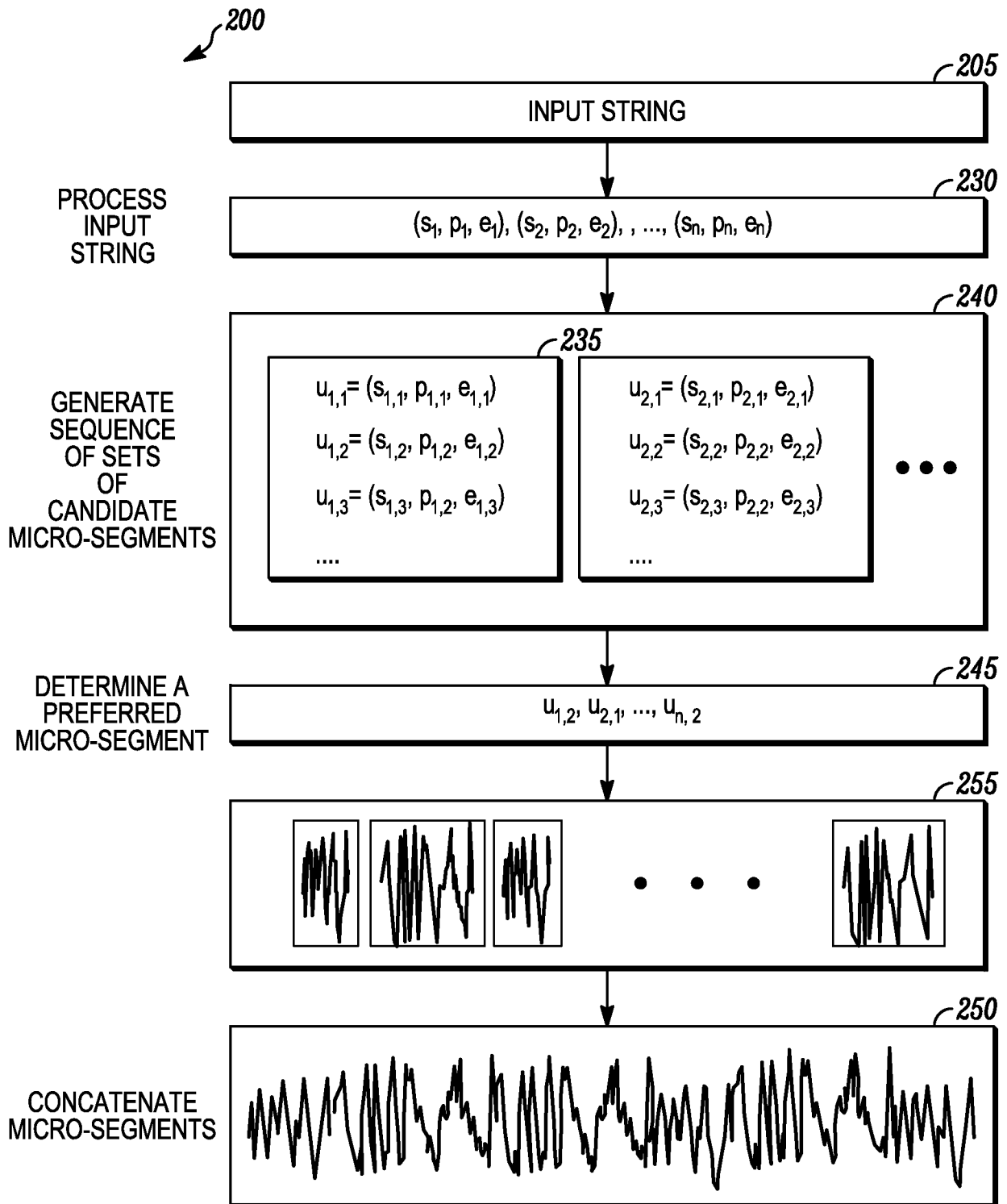
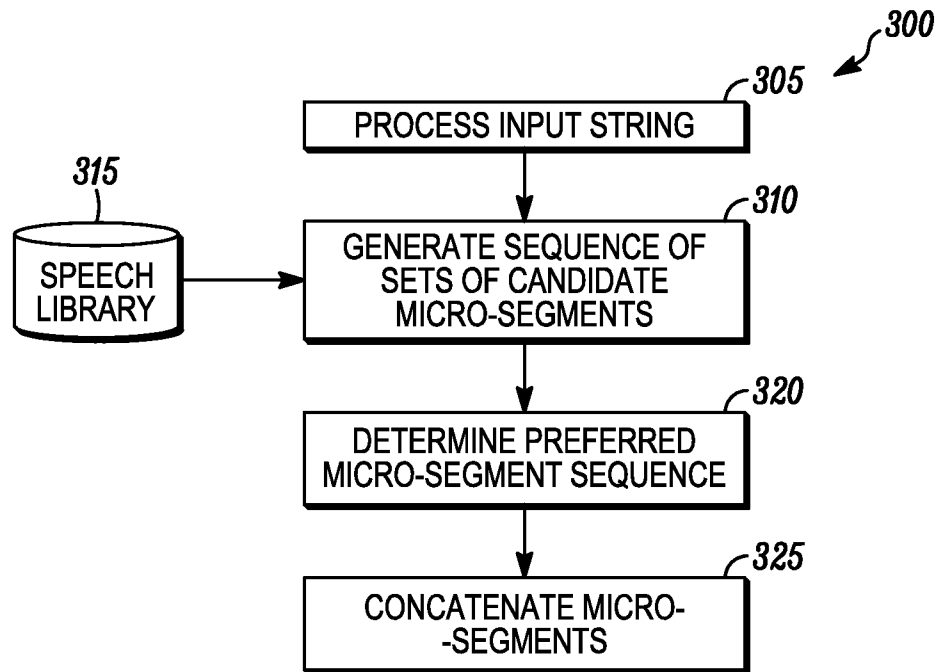
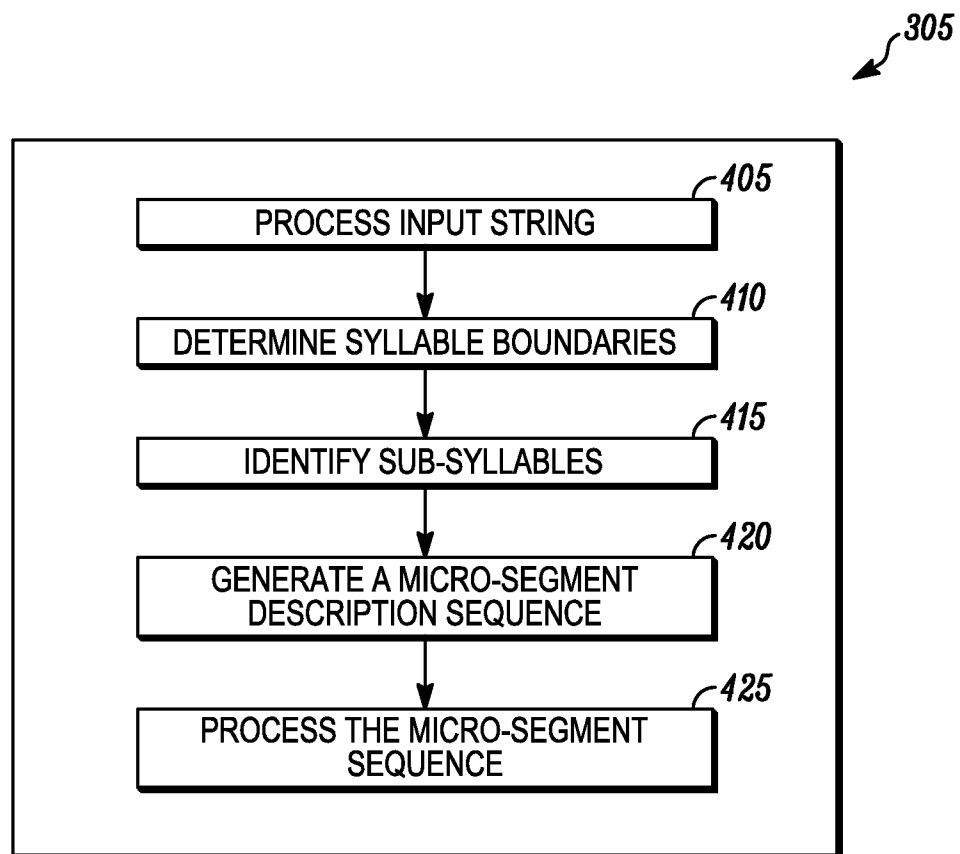


FIG. 2

3/4*FIG. 3**FIG. 4*

4/4

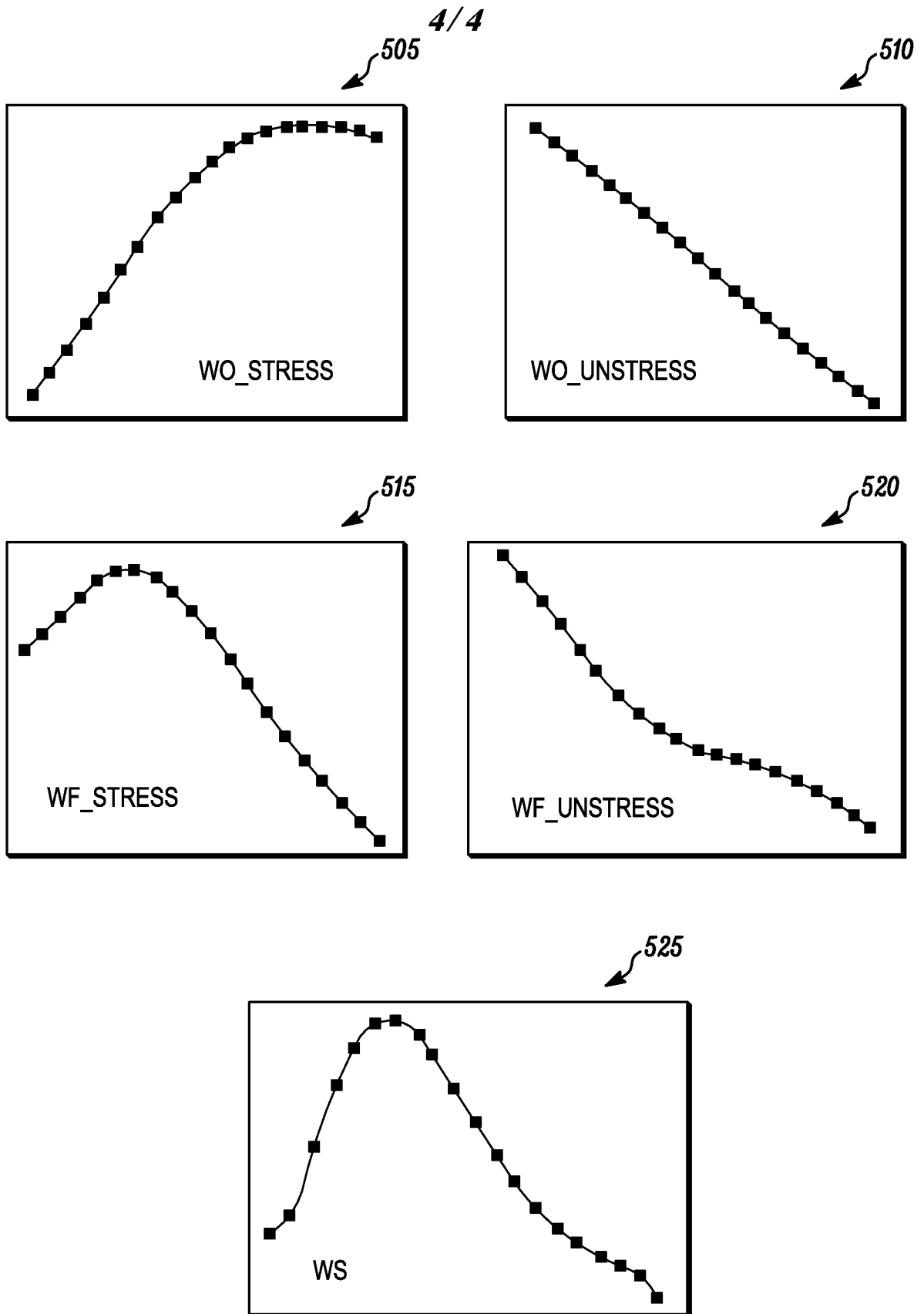


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2008/062822

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L13/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7 082 396 B1 (BEUTNAGEL ET AL) 25 July 2006 (2006-07-25)	1,5-16
Y	abstract; figure 4 column 1, lines 24-25 column 2, lines 47-56 column 3, line 57 - column 4, line 37 column 5, lines 16-61 column 6, lines 6-61 column 7, lines 1-44 column 8, lines 17-19	2-4
X	US 6 173 263 B1 (CONKIE) 9 January 2001 (2001-01-09) abstract; figure 6 column 4, lines 22-24 column 5, lines 5-20 column 6, lines 62-65	1,5-9
	----- -/-- -----	

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *&* document member of the same patent family

Date of the actual completion of the international search

7 July 2008

Date of mailing of the international search report

14/07/2008

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Santos Luque, Rocio

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2008/062822

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EL-IMAM Y A: "AN UNRESTRICTED VOCABULARY ARABIC SPEECH SYNTHESIS SYSTEM" IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, IEEE INC. NEW YORK, vol. 37, no. 12, 1 December 1989 (1989-12-01), pages 1829-1845, XP000099485 ISSN: 0096-3518	2-4
A	pages 1838-9	1,5-16
A	BLOUIN C ET AL: "Concatenation cost calculation and optimisation for unit selection in tts" SPEECH SYNTHESIS, 2002. PROCEEDINGS OF 2002 IEEE WORKSHOP ON 11-13 SEPT. 2002, PISCATAWAY, NJ, USA, IEEE, 11 September 2002 (2002-09-11), pages 231-234, XP010653653 ISBN: 978-0-7803-7395-2 page 232; table 1	1-16
Y	GB 2 313 530 A (ATR INTERPRETING TELECOMMUNICA [JP]) 26 November 1997 (1997-11-26) abstract; figure 3	1,5-13
Y	BENZMULLER R ET AL: "Microsegment synthesis-economic principles in a low-cost solution" ICSLP 96. PROCEEDINGS., FOURTH INTERNATIONAL CONFERENCE, vol. 4, 3 October 1996 (1996-10-03), pages 2383-2386, XP010238145 PHILADELPHIA, PA, USA ISBN: 978-0-7803-3555-4 the whole document	1,5-13
A	US 6 308 156 B1 (BARRY WILLIAM [DE] ET AL) 23 October 2001 (2001-10-23) abstract	1-16
A	BLACK A W ET AL: "OPTIMISING SELECTION OF UNITS FROM SPEECH DATABASES FOR CONCATENATIVE SYNTHESIS" 4TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '95, vol. 1, 18 September 1995 (1995-09-18), pages 581-584, XP000854776 MADRID, SPAIN the whole document	1-16

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2008/062822

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7082396	B1	25-07-2006	NONE
US 6173263	B1	09-01-2001	NONE
GB 2313530	A	26-11-1997	NONE
US 6308156	B1	23-10-2001	AT 183010 T 15-08-1999
		WO 9734291 A1 18-09-1997	
		DE 19610019 A1 18-09-1997	
		EP 0886853 A1 30-12-1998	