

MIXED-LINGUAL TEXT TO SPEECH

FIELD OF THE INVENTION

5

The present invention relates generally to Text-To-Speech (TTS) synthesis. The invention is particularly useful for letter to sound conversion for synthesized pronunciation of mixed-lingual text.

10

BACKGROUND OF THE INVENTION

Text to Speech (TTS) conversion, often referred to as concatenated text to speech synthesis, allows electronic devices to receive an input text string and provide a converted representation of the string in the form of synthesized speech. However, a device that may be required to synthesize speech originating from a non-deterministic number of received text strings will have difficulty in providing high quality realistic synthesized speech. One difficulty is based on letter to sound conversion in which identical letters or groups of letters may have different sounds and vowel stress/emphasis depending on other adjacent letters and position in a text segment to be synthesized.

25

Mixed-lingual TTS synthesis is emerging to find its applications in numerous domains such as call centre and handheld device etc. Hereafter mixed-lingual TTS means a speech synthesis technique for two or more languages that can be identified and detected from analysis of strings of text. When considering mixed-

30

lingual TTS for text strings of two languages, one dominant language, comprising the majority of the text, is usually termed as a primary language and the other is a secondary language. Typically, the primary language such as an Asian language (Mandarin, Cantonese, Japanese, Korean, Thai etc.) is mixed with English.

In a TTS system based on concatenation synthesis, an utterance is synthesized by concatenating several speech fragments together. Although concatenative speech synthesis has been identified as a promising approach to realize practical mixed-lingual TTS systems, identifying phonetic units, phonemes and their corresponding acoustic units is problematic. Also, the audible discontinuity at each concatenating points or between two successive units affects the synthesized speech quality. Further, both speech smoothing at each concatenating point and the smoothing of transitions between synthesized speech segments of two languages affects quality.

In this specification, including the claims, the terms 'comprises', 'comprising' or similar terms are intended to mean a non-exclusive inclusion, such that a method or apparatus that comprises a list of elements does not include those elements solely, but may well include other elements not listed.

SUMMARY OF THE INVENTION

5 According to one aspect of the invention there is provided a method for mixed-lingual text to speech synthesis, the method including:

 Receiving a mixed-lingual text string of at least two languages;

10 Separating text in the text string into either: (i) at least one character string of syllable based characters each of the characters representing a single syllable; (ii) or at least one word string comprising one or more words
15 formed by letters of the alphabet, at least some of the words representing more than one syllable;

 assigning phonetic units for each of the syllable based characters;

 assigning phonemes for each of the words, one
20 or more of the phonemes forming individual syllables;

 selecting individual acoustics units for both the phonetic units and phonemes; and

 concatenating the individual acoustics units
25 into concatenated ordered acoustics units having an order corresponding to the mixed-lingual text string; and

 performing speech synthesis using the concatenated ordered acoustics units.

30

 Suitably, the separating text also identifies which of the two languages is the primary language in the text string and which is the secondary language.

Preferably, the text comprising the primary language includes one or more dummy spacers indicating the position of text from the secondary language.

5 Suitably, the assigning phonetic units includes segmentation by maximum matching so that words in a lexicon match with the maximum number of consecutive Characters matching segments of the text of syllable based characters, the segmentation providing maximum
10 matched words.

 Preferably, the assigning phonetic units includes determining phrase boundaries for the maximum matched words to identify natural speech pauses that are not
15 explicitly identified by punctuation.

 Suitably, the assigning phonetic units includes identifying the phonetic units for each character having regard for which of the maximum matched words
20 the character comprises. Suitably, there is prosodic information associated with each of the phonetic units.

 Preferably, the assigning phonemes includes identifying one or more natural boundaries that are not explicitly identified by punctuation. Suitably, the
25 natural boundaries are identified by predictive word bundling based on part of speech patterns. Preferably, the part of speech patterns are identified using rules, the rules comprising:

30	Adjective + Noun	-> natural boundary;
	Number + quantity	-> natural boundary;
	Verb + Noun	-> natural boundary;
	Adverb + Verb	-> natural boundary;
	Noun + Noun	-> natural boundary;
35	Number + Noun	-> natural boundary;
	Verb + Verb	-> natural boundary; and
	Noun + Conjunction + Noun	-> natural boundary.

Suitably, the assigning phonemes includes segmenting words into sub-words forming a sub-word sequence, and identifying syllable boundaries to
 5 thereby assign the one or phonemes to individual syllables.

The identifying syllable boundaries preferably includes identifying each sub word as Consonant (C); a
 10 Vowel (V); and a Consonant group(CG).

Suitably, identifying syllable boundaries includes the following rules:
 VCV -> V|CV; VV->V|V; VCCV-> VC|CV when CC is not in
 15 CG; VCCV->V|CCV when CC is in CG; VCCCV->VC|CCV;
 VCCCCV-> VC|CCCCV,

Where “|” represents a syllable boundary.

Preferably, during concatenating, the acoustic
 20 units are inserted into the respective dummy spacers indicating the position of text from the secondary language.

Suitably, the concatenated ordered acoustic units
 25 undergo enhancing comprising merging excitation signals associated with the acoustic units at a concatenating boundary.

BRIEF DESCRIPTION OF THE DRAWINGS

5 In order that the invention may be readily understood and put into practical effect, reference will now be made to a preferred embodiment as illustrated with reference to the accompanying drawings in which:

10 Fig. 1 is a schematic block diagram of an electronic device in accordance with the present invention;

15 Fig. 2 is flow diagram illustrating a method for mixed-lingual text to speech synthesis that can be performed on the electronic device of Fig. 1; and

Fig. 3 illustrates an example of a Direct Acyclic Grap (DAG).

20 DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT OF THE INVENTION

Referring to Fig. 1, there is illustrated an electronic device 100, in the form of a radio-telephone, comprising a device processor 102
25 operatively coupled by a bus 103 to a user interface 104 that is typically a touch screen or alternatively a display screen and keypad. The electronic device 100 also has an utterance corpus 106, a speech synthesizer 110, Non Volatile memory 120, Read Only Memory 118 and
30 Radio communications module 116 all operatively coupled to the processor 102 by the bus 103. The speech synthesizer 110 has an output coupled to drive a speaker 112. The corpus 106 includes representations

of words or phonemes and associated sampled, digitized and processed utterance waveforms PUWs. In other words, and as described below, the Non Volatile memory 120 (memory module) in use provides for mixed-lingual Text-To-Speech (TTS) synthesis. Also the waveform utterance corpus 106 comprises sampled and digitized utterance waveforms in the form of phonemes and stress/emphasis of prosodic features.

As will be apparent to a person skilled in the art, the radio frequency communications unit 116 is typically a combined receiver and transmitter having a common antenna. The radio frequency communications unit 116 has a transceiver coupled to antenna via a radio frequency amplifier. The transceiver is also coupled to a combined modulator/demodulator that couples the communications unit 116 to the processor 102. Also, in this embodiment the non-volatile memory 120 (memory module) stores a user programmable phonebook database Db and Read Only Memory 118 stores operating code (OC) for device processor 102

Referring to Fig. 2, there is illustrated a method 200 for mixed-lingual text to speech synthesis. After a start step 210, a step 220 of receiving a mixed-lingual text string TS from the memory 120 is performed. The mixed-lingual text string TS may have originated from a text message received by module 116 or by any other means. Also, the mixed-lingual text string TS as described in this specification is typically formed from at least two languages, and by way of example a combination of Chinese Characters and alphabetical characters is described below; also the

text string TS may include digits 0 to 9 and/or punctuation marks.

5 A separating text step 230 then provides for
Separating text in the text string into either: (i) at
least one character string of syllable based characters
(SBC) each of the characters representing a single
syllable; (ii) or at least one word string comprising
10 one or more words hereafter also referred to as string
of words (SOW) formed by letters of the alphabet, at
least some of the words representing more than one
syllable. The separating text step 230 is performed by
unicoding that is known in the art. The unicode is
structured such that English Characters (letters) are
15 encoded so that each character is represented by a
number from 0000-007f and Chinese Characters are
encoded so that each character is stored by a number
from 4e00-9fa5.

20 The separating text step also identifies which of
the two languages is the primary language in the text
string and which is the secondary language. This is
achieved in this embodiment (English & Chinese mixed-
lingual text) by counting Chinese characters and
25 English words. If there are more Chinese Characters
than English words then the primary language will be
Chinese and the secondary language English.
Alternatively, if there are more English words than
Chinese Characters then the primary language will be
30 English and the secondary language Chinese.

To further explain the method 200 consider the
following four text strings in the form of four
sentences.

Text string 1: “英文单词 mention 可翻译成中文的提及”
 Text string 2: “Watson 是一位成功的商人”
 Text string 3: “大儿子目前住在 Chicago”
 5 Text string 4: “中国政府决定参加 APEC 会议”.

After the separating text step 230 for above four text strings, there are four strings of syllable based characters (SBC) from the primary language, these are
 10 as follows:

syllable based character string1 (SBC1):
 “英文单词 BLANK1 可翻译成中文的提及”
 syllable based character string2 (SBC2):
 “BLANK2 是一位成功的商人”
 15 syllable based character string3 (SBC2):
 “大儿子目前住在 BLANK3”
 syllable based character string4 (SBC2):
 “中国政府决定参加 BLANK4 会议;

20 There are also four strings of words (SOW), from the secondary language, each comprising one or more words, and these are identified as follows:

string of words1 (SOW1) -> Blank1: “mention”
 string of words2 (SOW2) -> Blank2: “Watson”
 25 string of words3 (SOW3) -> Blank3: “Chicago”
 string of words3 (SOW4) -> Blank4: “APEC”.

From the above, since there are more Chinese Characters than English words then the primary language
 30 will be Chinese and the secondary language English. This is apparent since the English words are represented as blanks in the strings of syllable based characters (SBC). Thus, the text comprising the primary language includes one or more dummy spacers

(the blanks BLANK1 TO BLANK4) indicating the position of text from the secondary language.

5 Considering the Chinese characters that are the
 strings of syllable based characters (SBC1 - SBC 4),
 the method 200 performs a step of assigning (240) for
 assigning phonetic units for each of the syllable based
 characters (SBC). The step of assigning (240) can
 10 essentially be broken down into three sub-steps these
 are: (a) text analysing; (b) Determining phrase
 boundaries; and (c) identifying phonetic units.

15 Sub-step (a) text analyzing essentially performs
 segmentation and normalization of Chinese Characters by
 using a word lexicon. This lexicon comprises over
 80,000 words part of which is shown in listing 1 below.

20 董
 董事
 董事长
 董事会
 董事局
 懂
 懂得
 懂事
 懂行
 25 ...
 英
 英文

Listing 1. Word Lexicon

30 The segmentation selects segments (words) by
 maximum matching so that words in the lexicon (stored
 in memory 120) match with the maximum number of
 consecutive Characters matching segments of the text of
 35 syllable based characters (SBC). For example,
 considering the syllable based character string1

(SBC1): “英文单词 BLANK1 可翻译成中文的提及”. Segmentation proceeds from the first to last character (sentence initial to sentence final). Hence, segmentation proceed in the following sequence: 英 is found in the
 5 lexicon, 英文 is found in the lexicon, 英文单 is not found in the lexicon. So the first segment or word is 英文. Similarly, 单 is found in the lexicon, 单词 is found in the lexicon and since the next word is a blank then the second segment or word is 单词. The blank
 10 “BLANK1” is skipped (not analysed for segmentation). Next, 可 is found in the lexicon, 可翻 is not found in the lexicon, so the third segment or word is 可. Similarly, the rest of the segments or words are 翻译, 成, 中文, 的 and 提及. The same process of segmentation is performed on the syllable based character strings (SBC2
 15 - SBC4) and provides for Maximum Matched Words (MMW) as will be apparent to a person skilled in the art.

Normalization, is useful for special characters
 20 such as digits and punctuation marks. Normalization uses rules to identify these special characters that have a special meaning or their corresponding phonetic units can differ depending on the context of these special characters. Hence, normalization provides
 25 normalized characters from the special characters. Also, normalization is effected by a rule base, part of which is as follows:

30 Dot Rules:

Full stop - is an end of sentence such as -> 上海是中国最大的城市之一.

35 Decimal point - is normalized to have the sound “point” such as ->
 1.234 - > 一点二三四.

Dot in a URL is normalized to have the sound "dot"
such as -> www.mot.com([www 点 mot 点 com](http://www.mot.com), www dot mot dot com)

5 Colon Rules:

Sentence break, 这有

Ratio: 比分是 2:2, (should be converted into 比)

Time: 10: 05

10

Slash Rules:

Fraction flag: 2/3 (三分之二)

Date flag: 97/10 (九七年十月, October, 1997)

15

Comma Rules:

Number identifier: 123,456 (十二万三千四百五十六, one hundred twenty three thousand three hundred fifty six)

Digit string rules:

20

Digit: 123456 (一二三四五六, one two three four five six)

Sub step (b) Determining phrase boundaries performs predictive word bundling (phrasing) using maximum matching as will be apparent to a person skilled in the art. The purpose of determining the phrase boundaries is to identify natural boundaries (natural speech pauses) that are not explicitly identified by punctuation in strings of the Maximum Matched Words (MMW). The predictive word bundling is based on part of speech patterns and uses the following rules:

	Adjective + Noun	-> natural boundary;
	Number + quantity	-> natural boundary;
35	Verb + Noun	-> natural boundary;
	Adverb + Verb	-> natural boundary;
	Noun + Noun	-> natural boundary;
	Number + Noun	-> natural boundary;
	Verb + Verb	-> natural boundary; and
40	Noun + Conjunction + Noun	-> natural boundary.

The Noun + Noun rule is recursive so more than two Nouns in sequence also provides a natural boundary (e.g. Noun + Noun + Noun; Noun + Noun + Noun + Noun etc.). Also, the Verb + Verb rule is recursive.

5

The natural boundaries identify emphasis and pauses. Hence, for the segmented syllable based character string1 (SBC1), the natural boundaries are identified as follows:

10

英文 单词: noun + noun

可 翻译 成: verb + verb + verb

15

中文 的 提及: noun + conjunction + noun

So the natural boundaries (phrasing) is:

英文 单词| BLANK1| 可 翻译 成| 中文 的 提及.

20

Where, “|” represents a phrase boundary.

Furthermore, the natural boundaries (phrasing) for SBC2 to SBC4 are:

25

For SBC2 after segmenting ->

BLANK2, 是, 一位, 成功, 的, 商人

After phrasing ->

BLANK2 | 是|一位 成功 的 商人

30

For SBC3 after segmenting ->

大, 儿子, 目前, 住, 在, BLANK3

After phrasing ->

大 儿子| 目前 住 在| BLANK3

35

For SBC4 after segmenting ->

中国, 政府, . 决定, . 参加, BLANK4, 会议

After phrasing ->

中国 政府 决定|参加|BLANK4 会议

5

Sub step (c) identifying phonetic units uses the results from Sub step (a) and sub step (b) to identify the phonetic units for each of the syllable based characters having regard for which of the maximum
10 matched words the character comprises. This sub step also identifies phonetic units for the natural breaks and normalized characters.

If a unit in the speech inventory stored in the
15 non volatile memory 120 is matched with the target phonetic unit in prosodic position and contextual information, this unit is selected. For instance, for the example Text String1 TS1 that has undergone predictive word bundling (phrasing) to have the natural
20 boundaries identified as:

英文单词|BLANK1|可 翻译成|中文的提及, then the phonetic units selected are shown in table 1. In table 1, the number associated with the prosodic position affects the pitch duration and energy (e.g. RMS power)
25 with the associated phonetic unit. In this regard, number 0 means the phonetic unit is at the beginning of a sentence; number 1 means the phonetic unit is at the end of a sentence; number 2 means the phonetic unit is at the beginning of a phrase; number 3 means the
30 phonetic unit is in the middle of a phrase; and number 4 means the phonetic unit is in the end of a phrase.

Character	Phonetic Unit	Prosodic Position
英	ying1	0
文	wen2	3
单	dan1	3
词	ci2	4
可	ke3	2
翻	fan1	3
译	yi4	3
成	cheng2	4
中	zhong1	2
文	wen2	3
的	de5	3
提	ti2	3
及	ji2	1

Table 1, examples of phonetic units and associated prosodic positions for Chinese characters

5

After the step of assigning (240) the method performs a selecting step (250) for selecting individual acoustics units for the phonetic units. (For syllable-based language such as Chinese, phonetic units with prosodic position and contextual information to perform unit selection is identified.

10

The unit selection procedure for a syllable-based language can be described as follows. A syllable with tone and prosodic position is searched in an inventory stored in the memory 120 and the acoustic unit selected. If not found, the method will relax the condition for prosodic position and will select a syllable with the same tone and a different prosodic position.

15

20

Now returning to the separating text step 230 and considering the English strings of words (SOW), these words are processed at an assigning step (260) that provides for assigning phonemes for each of the words, where one or more of the phonemes form individual syllables. The step of assigning (260) can essentially be broken down into four sub-steps these are: (a) normalizing; (b) Determining phrase boundaries; (c) letter to sound conversion; and (d) determining syllable boundaries.

Sub-step (a) performs normalizing that is essentially the same normalizing as described above but using English text in the rule base.

Sub-step (b) Determining phrase boundaries is performed by a process described in Chinese Patent application filed under number 03132708.7 that is incorporated into this specification by reference. In summary, the phrase boundaries are determined by statistical analysis and heuristics selected from a large training set of over 90,000 transcriptions. These heuristics are stored in the non-volatile memory 120. The purpose of the phrase boundaries is to identify natural boundaries (natural speech pauses) that are not explicitly identified by punctuation.

Sub-step (c) letter to sound conversion addresses the fact that identical letters or groups of letters may have different sounds and vowel stress/emphasis depending on other adjacent letters and position in a text segment to be synthesized. Sub-step (c) letter to sound conversion is performed by a process described in Chinese Patent application filed under number

03132709.5 that is incorporated into this specification by reference. However, in summary letter to sound conversion involves segmenting words into sub-words, the sub-words forming a sub-word sequence with one or more of the sub-words comprising at least two letters. The sub-word sequence is determined by analysis of all possible sub-words that could comprise a selected word. This analysis is performed by use by construction of a Direct Acyclic Graph (DAG).

Consider, for example, if a selected word "mention" was being processed by step 260, then the Direct Acyclic Graph DAG of Fig. 3 is constructed with all possible sub-words that could comprise the selected word "mention". With each sub-word a pre-defined weight WT is provided, for example the sub-word "ment", "men" and "tion" have respective weights 88, 86 and 204. Accordingly, the DAG is traversed and the sub-words with the maximum combined (summed) weights WT that form the selected word are selected to be identified as phonemes as described in Chinese Patent application filed under number 03132709.5. In the case for the word "mention" the sub-words "men" and "tion" would be selected. and thereafter phonemes can be identified for these sub-words once syllable boundaries are identified.

The sub-step (d) identifying syllable boundaries is performed using the following rules:

A Consonant (C) comprise any of the following sub-words: TS, b, ch, d, dh, f, g, hh, k, l, m, n, ng, , p, r, s, sh, t, th, v, w, y, z, zh;

A Vowel (V) comprise any of the following sub-words: aa, ae, ao, aw, ax, ay, eh, er, ey, ih, iy, ow, oy, uh, uw;

A Consonant group (CG) comprise any of the following sub-words:

5 bl, br, by, dr, dy, fl, fr, fy, gl, gr, gw, gy, kl, kr, kw, ky, ly, my, ny, pl, pr, py, sk, sl, sm, sn, sp, st, tr.

Boundary Rule 1: VCV -> V|CV;

Boundary Rule 2: VV->V|V;

10 Boundary Rule 3: VCCV-> VC|CV when CC is not in the CG;

Boundary Rule 4: VCCV->V|CCV when CC is in the CG;

Boundary Rule 5: VCCCV-> VC|CCV;

Boundary Rule 6: VCCCCV-> VC|CCCCV.

Where "|" represents a syllable boundary.

15

After the sub-step (d) identifying syllable boundaries, the sub-words such as "men" and "tion", phonemes are identified for these sub-words. Hence, the identifying step identifies phonemes for each word comprising sub-words and, as will be apparent to a person skilled in the art, more than one phoneme can form an individual syllable. For instance, for the examples of the text string 1 to text string 3, the phonemes and associated prosodic information would be selected. based on identification of the natural boundaries. Such phonemes and prosodic information (prosodic position) are shown in table 2 where M has the meaning that the word is at the beginning or middle of a sentence (i.e. not the sentence's end); F means that the word is at the end (Final Position) of a sentence; 0 means that the syllable is at the beginning or middle of a word (i.e. not the word's end); and F means that the syllable is at the end (Final Position) of a word.

20

25

30

Word	Phoneme	Prosodic position
Mention	m'ehn	MO
Mention	shaxn	MF
Watson	wat	MO
Watson	son	MO
Chicago	ch	FO
Chicago	kar	FO
Chicago	gow	FF

Table 2, examples of phonemes and associated prosodic positions for English text

5

After the assigning step 260 a step of selecting
 270 provides for selecting individual acoustic units
 based on a Variable Length Unit (VLU) model that
 comprises one or more phonemes, whereas the acoustic
 units for the Chinese are not variable in length as
 each acoustic unit comprises a syllable.

10

For the VLUs there are 5 possible classes:

Class 1: regular consonant + vowel, semi-vowel +
 vowel, nasal + vowel. Regular consonant includes
 stop, fricative and dental etc.

15

Class 2: Class 1 + nasal/semi-vowel.

Class 3: sub-CV. It is a part of Class 1 or Class 2.

20

Class 4: consonant initial and consonant initial
 group. Consonant initial is a consonant acting as
 syllable initial while consonant initial group mainly
 refers to a stop followed by a dental. (For example,
 s+ t)

Class 5: consonant final and consonant final group.

25

From the above, if a word has two consecutive vowels or a semivowel preceded and followed by a vowel, maximum matching is conducted in the speech inventory stored in memory 120. If there is no match, the method will search syllable by syllable and assign acoustic units. In this embodiment, maximum matching is always performed. The purpose of this acoustic unit selection is to reduce concatenation points and select acoustic units with the same or similar prosodic constraints.

After the steps of selecting 250 and 270 the resulting acoustic units are merged at a concatenating step 275, where individual acoustic units are concatenated into ordered acoustic units in an order corresponding to the mixed lingual text processed by the method 200. In the process of concatenating, and using the examples of text strings 1 to 4, the acoustic units formed by the VLU (the phonemes) are inserted into the respective dummy spacers (BLANK1 to BLANK4) in the corresponding text string.

During the process of concatenating, a parametric bit-stream identified by the acoustic units is firstly decoded into frames comprising two major components: vocal tract parameters LSP (Line Spectral Pairs) that includes an $LSP[i]$ vectors for each frame; and an excitation signal represented by an excitation stream $Exc[i][j]$ (i is index of frame and j is the index of sample in a frame). A rough pitch estimation $OT[i]$ of each frame i is also obtained from an adaptive lag parameter of the bit-stream. As will be apparent to a person skilled in the art, one acoustic unit may comprise several frames. The above three parameters are

presented frame by frame for each acoustic unit. Also, for each frame, the parameter $LSP[i]$ is a vector and includes 10 elements examples of which are: [1227, 2027, 3884, 9108, 12892, 16454, 19866, 23019, 26304, 29553]. Further, the parameter $Exc[i][j]$ is vector including 240 elements examples of which are: [0, 0, -1, -3, 0, 2, 4, 10, 6, -1, ...]; and the parameter $OT[i]$ is a integer within the range [40, 150] such as 78.

Next, enhancing is performed at an enhancing step where the concatenated ordered acoustic units undergo enhancing which involves:

- (a) Smoothing a pitch estimation $OT[i]$ of each frame at a concatenating boundary of two adjacent acoustic units. The pitch values of two units at a boundary are linearly interpolated to obtain smoothing new values $OT[T]$ (for an end frame of a front or preceding unit) and $OT[H_u]$ (for the start frame of the next unit);
- (b) Merging the excitation signal at concatenating boundary of two adjacent acoustic units. The first sub-frame excitation signal of the successive unit is merged into a new sub-frame with the last one sub-frame of the front unit by triangular window;
- (c) Interpolating the $LSP[i]$ vectors at the concatenating boundary of adjacent acoustic units. Each $LSP[i]$ vector corresponding to the merged excitation signal is obtained by interpolating the last $LSP[i]$ vector of the front unit and the first $LSP[i]$ vector of the succeeding unit. A concatenated $LSP[i]$ vector series is therefore constructed.

At a performing speech synthesis step 285, the concatenated LSP[i] vector series of successive acoustic units are converted to predictive coefficients $a[i][j]$ (i is the index of frames and j is the order of prediction). Smoothing is reconstructed by feeding the concatenated smoothing excitation series comprising $OT[T]$, $OT[H_v]$ etc to the Speech Synthesizer 110 thereby the multi-lingual text string TS is synthesized as speech by transforming the predictive coefficients $a[i][j]$ and associated smoothing excitation series into a waveform synthesizing speech. Also, the as will be apparent to a person skilled in the art, stress (primary, secondary or no stress as appropriate) on vowels may also used to provide an improved synthesized speech quality by appropriate stress emphasis.

After the performing speech synthesis 285 the method 200 terminates at an end step 290. It will be apparent that although Chinese was identified as the primary language for the purpose of example and explanation, English could also be the primary language. When English is the primary language, the dummy word spacers are included in the English text indicating the position of text from the secondary language (e.g. Chinese). Furthermore, during concatenating the acoustic units are inserted into the respective dummy spacers indicating the position of text from the secondary language. It will also be apparent that although steps 240 and 250 were described before steps 260 and 270, steps 260 and 270 could be effected before or concurrently with steps 240 and 250.

Advantageously, the present invention provides for mixed-lingual TTS synthesis that can be used on electronic devices such as handheld or portable devices such as laptop computers, Personal Digital Assistants,
5 Cellular telephones plus many other devices.

The detailed description provides a preferred exemplary embodiment only, and is not intended to limit the scope, applicability, or configuration of the
10 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
15 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:

- 5 1. A method for mixed-lingual text to speech synthesis, the method including:
 Receiving a mixed-lingual text string of at least two languages;
 Separating text in the text string into either: (i) at least one character string of
10 syllable based characters each of the characters representing a single syllable; (ii) or at least one word string comprising one or more words formed by letters of the alphabet, at least some of the words representing more than one syllable;
15 assigning phonetic units for each of the syllable based characters;
 assigning phonemes for each of the words, one or more of the phonemes forming individual syllables;
20 selecting individual acoustics units for both the phonetic units and phonemes; and
 concatenating the individual acoustics units into concatenated ordered acoustics units having an order corresponding to the mixed-lingual text
25 string; and
 performing speech synthesis using the concatenated ordered acoustics units.
- 30 2. A method as claimed in claim 1, wherein the separating text identifies which of the two languages is the primary language in the text string and which is the secondary language.

3. A method as claimed in claim 1, wherein the text comprising the primary language includes one or more dummy spacers indicating the position of text from the secondary language.

5

4. A method as claimed in claim 1, wherein the assigning phonetic units includes segmentation by maximum matching so that words in a lexicon match with the maximum number of consecutive Characters matching segments of the text of syllable based characters, the segmentation providing maximum matched words.

10

5. A method as claimed in claim 4, wherein the assigning phonetic units includes determining phrase boundaries for the maximum matched words to identify natural speech pauses that are not explicitly identified by punctuation.

15

6. A method as claimed in claim 5, wherein, the assigning phonetic units includes identifying the phonetic units for each character having regard for which of the maximum matched words the character comprises.

20

7. A method as claimed in claim 6, wherein there is prosodic information associated with each of the phonetic units.

25

8. A method as claimed in claim 6, wherein the assigning phonemes includes identifying one or more natural boundaries that are not explicitly identified by punctuation.

30

9. A method as claimed in claim 8, wherein the natural boundaries are identified by predictive word bundling based on part of speech patterns.

5 10. A method as claimed in claim 9, wherein the part of speech patterns are identified using rules, the rules comprising:

10	Adjective + Noun	-> natural boundary;
	Number + quantity	-> natural boundary;
	Verb + Noun	-> natural boundary;
	Adverb + Verb	-> natural boundary;
	Noun + Noun	-> natural boundary;
	Number + Noun	-> natural boundary;
15	Verb + Verb	-> natural boundary; and
	Noun + Conjunction + Noun	-> natural boundary.

20 11. A method as claimed in claim 8, wherein the assigning phonemes includes segmenting words into sub-words forming a sub-word sequence, and identifying syllable boundaries to thereby assign the one or phonemes to individual syllables.

25 12. A method as claimed in claim 11, wherein the identifying syllable boundaries preferably includes identifying each sub word as Consonant (C); a Vowel (V); and a Consonant group(CG).

30 13. A method as claimed in claim 11, wherein identifying syllable boundaries includes the following rules:

VCV -> V|CV; VV->V|V; VCCV-> VC|CV when CC is not in CG; VCCV->V|CCV when CC is in CG; VCCCV->VC|CCV; VCCCCV-> VC|CCCV,

35 where "|" represents a syllable boundary.

14. A method as claimed in claim 1, wherein during concatenating the acoustic units are inserted into the respective dummy spacers indicating the position of text from the secondary language.

5

15. A method as claimed in claim 11, wherein the concatenated ordered acoustic units undergo enhancing comprising merging excitation signals associated with the acoustic units at a concatenating boundary.

10