Suggestions on Tone and Word Boundary of Mandarin for SSML

LOU Xiaoyan, LI Jian
Research and Development Center, Toshiba China
{louxiaoyan, lijian@rdc.toshiba.com.cn}

Abstract

In this position paper, some suggestions are proposed so that the SSML can process tonal language like Mandarin and et.al. Since in Mandarin word boundary is not explicit like in English, the suggestion of word boundary for Mandarin is also proposed in this paper.

Suggestion

1．Tone

Mandarin is a tonal language. In Mandarin, each Chinese character is pronounced as one syllalbe. The same syllables with different tones take different meanings. Tones are as important and essential as phonemes in Mandarin. There are 5 tones in Mandarin called as “yin ping(阴平)”, “yang ping(阳平)”, “shang sheng(上声)”, “qu sheng(去声)” and “qing sheng”(轻声) respectively. Tone “qing sheng” is also called as neutral tone.

In Mandarin, some characters change their original tones in some special cases. This phenomenon is called as tone sandhi. For example, if the tones of two successive Chinese characters in a prosodic word are both tone 3, the tone of previous Chinese character will change to tone 2. The synthesizer should consider tone sandhi and give the changed tones.

Table 1 shows the 5 tones in Mandarin. Column 3 gives the suggested value of tone in Mandarin for easier input to a computer..

Table 1

tone	Shape of pitch contour	suggested Value of the tone
阴平	High	1
阳平	Low-high	2
上声	High-low-high	3
去声	High-low	4
轻声	Low	5

As an important part of pronunciation in tonal language, tone(s) should also be offered with phoneme sequence to distinguish syllable(s). For easier recognition, a pinyin sequence, a sequence of phoneme and tone of format as shown below, is suggested. Slash seperates the pinyin of one character from others.

Figure 1: An example of pinyin sequence.

Text: 大都会（dàdūhuì）
Pinyin sequence: /d a 4/d u 1/h ui 4/

- Non-markup behavior: For a tonal language synthesizer, it must derive tone(s) for each syllable if no tonal markup is used. It can be achieved by looking up a pronunciation dictionary (which maybe language-dependent) and applying rules to determine tone inflection.

- Markup support: In SSML 1.0, elements **phoneme**, **lexicon** and **say-as** are optional in step "text-to-phoneme" to provide the content creator with explicit control over pronunciation. In a tonal language, tone, as an important part of pronunciation, should also be markuped optionally when the content creator hopes to set tones. And especially, it may be difficult to deal with tone sandhi for a synthesizer. Using tone markup, the content creator can set changed tone directly and get speech with correct tones. So two following means are suggested to indicate the tone of syllable by introducing the new markups.

In **solution 1**, **tone** element is suggested to render the tone information of the contained text. This **tone** element is optional. In this solution, an attribute to specify the tone is required, This attribute is named as the **detail** attribute.

Figure 2: An example of solution 1

Example of solution 1:

```
<tone detail="2 4">不要</tone>
```

In the example of solution 1, the value of **detail** attribute is "2 4", it indicates that the tone value of "不" is "2" and the tone value of "要" is "4".

In this solution, the tone element can only contain text (no element).

Solution 2 attempts to add value "t" and "pt" to the **alphabet** atrribute of **phoneme** element. The Pinyin sequence as shown in Figure 1 is suggested to be used as the value of the **ph** attribute for better understanding. In this case, the synthesizer should support these pinyin definitions.

Figure 3: An example of solution 2

Example of solution 2:

```
<phoneme alphabet="t" ph="2 4">不要</phoneme>
<phoneme alphabet="pt" ph="/bu 2/yao 4/">不要</phoneme>
```

When the value of **alphabe**t is "pt", the full pinyin sequence with tone should be given as the value of **ph**. And the lexicon presentation of **ph** is shown as following.

/phoneme-sequence1 tone1/phoneme-sequence2 tone2/

When the value of **alphabet** is "t", only the tone sequence is give as the value of **ph**. Then the lexicon presentation of **ph** is shown as following.

/tone1/tone2/…

When the value of **alphabet** is "t", the phoneme sequence should be derived by the synthesizer automatically.

The solutions introduced above are intended to provide tone markups for Chinese character(s). The tone strings given by the markups cannot be changed by the text normalization step or by the result of looking up the lexicon.

When the value of tone is not supported by a synthesizer, or the length of the tone sequence is unequal to the number of characters in the contained text, it must render the contained text as if no markup for tone is used.

2. Word boundary

Chinese sentences are composed with strings of Chinese characters without blanks or spaces to specify the word boundaries. Chinese word is the basic unit for sentence parsing and understanding. Therefore the first step of processing Chinese sentences is to identify the word boundries. The synthesizer may have difficulties to identify these words (1) complex words, such as reduplications, derived words etc., (2)proper names, (3)the ambiguous word segmentation. The ambiguity is caused by the different meanings of the same character string takes in different context. An example of ambiguity is given below.

Figure 4: An example of segmentation ambiguity

a: 上海是个***大都会***。 → 上海 是 个 大 都会。(Shanghai is a ***metropolis***)
b: 上海人***大都会***那么说 → 上海人 大都 会 那么 说。(***most*** Shanghainese ***will*** say something like that)

A Chinese character sequence “大都会” appears both in sentence (a) and (b). In sentence (a), it is a word and its meaning is metropolis, while in sentence (b), it is separated as two words and it means most people will do something. And the Chinese character “都” in sentence (a) pronounced as “du1” while it pronounced as “dou1” in sentence (b). So it is very important to get the correct word boundary for Chinese.

- Non-markup behavior: the synthesizer should attempt to determine the word or phrase boundary using language-specific knowledge.
- Markup support: A **w** element is suggested to be introduced in SSML to indicate the word boundary in Chinese. And this element is thought to be useful to eliminate the ambiguities. At the same time **xml:lang** is also suggested to be used as one attribute of this **w** element.

The use of **w** element is optional.

The **w** element is suggested to render the contain text only and the following elements can be : **audio**, **emphasis**, **mark**, **phoneme**, **prosody**, **say-as**, **sub**, **voice** and **t**(if defined).

Figure 5: One example to makup **w** element.

```
<w>都会</w>
```

An optional attribute is recommended to be defined for the **w** element, and here names the attribute as **detail**. And this **detail** attribute is optional. The legal values of the **detail** attribute depend on the number of Chinese characters of the contained text when the contained text is Chinese. The default value of this attribute is the total number of the Chinese characters. It means the contained text will be regarded as one word.

Figure 6: Another example to makup **w** element

```
<w value=“3 2 1”>上海人大都会</w>
```

In the example shown above, the phrase is seperated into three words, and the number of Chinese characters in these words are 3, 2 and 1 respecitively. So the phrase should be interpreted as “上海人”, “大都” and “会”.

When the value of the **detail** attribute is unsupported by a synthesizer, it must render the contained text as if no value were specified for the **detail** attribute.

Influence on SSML 1.0

In the previous section, we provide two ways to represent tone and word boundary by introducing new elements or attributes. This part is to discuss the possible influences of these new markups on the SSML 1.0.

1． Influence on speech synthesizing steps

   As described above, word is the basic unit for sentence parsing and understanding. So when dealing with Mandarin, the synthesis preocess should segment words in the input text first then parse textand analyze structure.

2． Element and attribute

   We provide the author of content the ability to represent tone and word boundary by introducing new markups. And here the relation between the markups defined in SSML 1.0 and these new markups will be discussed.

   a) The relation between SSML 1.0 markups and tone markup

   In the previous section, two solutions are discussed to markup tone. When the **tone** element is used, the **tone** element can only contain text(no element). And it can be enclosed by these elements: **p**, **s**, **w**(if defined). If the **phoneme** element is modified to adopt tone markup, the relation between the **phoneme** element and others should not be changed.

   b) The relation between SSML 1.0 markups and the word segmentation markup

   The **w** element is also introduced in this paper. And the **p** and **s** element defined in SSML 1.0 may be followed by the **w** element. And the **w** element can be followed by **audio**, **emphasis**, **mark**, **phoneme**, **prosody**, **say-as**, **sub**, **voice** and **t**(if defined)