SSML Extensions for Chinese Voice Browsing

Helen MENG, Wai-Kit LO, Tien-Ying FUNG, Yuk-Chi LI and Zhiyong WU
The Chinese University of Hong Kong,
Hong Kong SAR, China

{hmmeng, wklo, tyfung, ycli, zywu@se.cuhk.edu.hk}

Abstract

In this paper, we will summarize our suggestion for extending the Speech Synthesis Markup Language (SSML) specification to better support Chinese voice browsing. Three elements (**`<phrase>`**, **`<word>`**, **`<tone>`**), one attribute (**`dialect-accent`**) and six attribute values (for the **`<say-as>`** element) are proposed to tackle the unique challenges presented by the Chinese language, such as the richness in dialects, the monosyllabic and tonal nature, the lack of word boundaries and the ambiguity introduced by multiple possible syllable pronunciations to a character.

1. Introduction

This document focuses on extending the Speech Synthesis Markup Language (SSML) specification by proposing new elements, attributes and attribute values for better support of Chinese voice browsing.

The current version of SSML 1.0 is generic for multiple languages. However, Chinese has certain features that differ from western languages. For example, it is very rich in dialects, it is monosyllabic and tonal, it lacks explicit word boundaries and often has ambiguities due to homographs — multiple possible pronunciations to a single character. These characteristics present additional challenges in text-to-speech (TTS) synthesis for Chinese and additional extensions to SSML will be of benefit for TTS implementation. To handle these characteristics, we propose three additional elements including **phrase**, **word** and **tone**; a **dialect-accent** attribute, and six attribute values for the **`interpret-as`** attribute of the **say-as** element are suggested.

2. Characteristic of the Chinese Language

2.1. Chinese Dialects and Accents

Chinese is rich in dialects (e.g. Mandarin, Cantonese, Hakka, Min, etc.). Influenced by their native dialects, people from different regions of China speak with different accents. For example, Beijing Mandarin sounds significantly different from Guangdong Mandarin because of their long geographical distance. For the same dialect, people have preferences on which accent to listen to based on their habits. Different accent also imply which region the speaker is from.

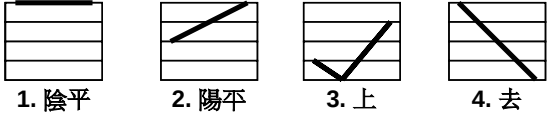
2.2. Pronunciation Transcription for Chinese dialects

Different phonetic transcription schemes are adopted for different dialects when transcribing Chinese characters into tonal syllables using Roman alphabets. For Mandarin, the "pinyin" (漢語拼音) scheme is used and for Cantonese, the "jyutping" (粵拼) developed by the Linguistic Society of Hong Kong (LSHK) is employed. Since Chinese is monosyllabic and tonal, pronunciation of a Chinese character can be represented by a syllable and a tone. When transcibing Chinese characters, a pair of "/" will be used to enclose the syllables. For example, the pronunciation of 銀行 (bank) in Mandarin is /yin2 hang2/ and it is transcribed as /ngan4 hong4/ in Cantonese.

Tones in Chinese dialects are realized as different pitch profile and duration for the syllables. A 1-digit number is used to describe the tone of a syllable. Different dialects have different number of tones. There are four tones in Mandarin and six or nine tones in Cantonese. Figures 1 to 3 illustrate the tone systems for Mandarin and Cantonese.

Figure 1 shows the schematics for pitch profiles of the four tones in Mandarin, namely Yin ping (陰平), Yang ping (陽平), Shang (上), Qu (去). There is also an additional neutral tone that has no fixed tone height and shape.

Fig. 1 Pitch profiles for tones in Mandarin.



1. 陰平 /yin ping/, high level | **2. 陽平 /yang ping/, low level** | **3. 上 /shang/, rising** | **4. 去 /qu/, going**

Fig. 2 Pitch profiles for non-entering tones in Cantonese.

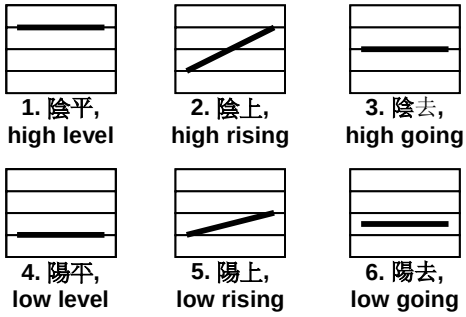
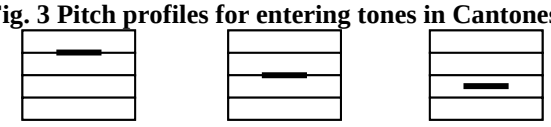


Fig. 3 Pitch profiles for entering tones in Cantonese.



Figures 2 and 3 show the schematics for the pitch profiles of the nine tones in Cantonese, which are categorized as non-entering (tones 1 to 6) and entering tones (tones 7 to 9). The pitch levels and profiles of entering tones 7, 8 and 9 are identical to the non-entering counterparts, tones 1, 3, and 6, respectively. However, entering tones are shorter in duration. For simplicity, a six-tone system is applied in some Cantonese transcription schemes (e.g. "jyutping").

2.3. Text Processing of Written Chinese

Word segmentation is very important in Chinese text processing. In a written Chinese sentence, all characters pack together without explicit word or phrase boundaries. In general, we can view a Chinese word (字詞) as a meaningful unit composed of one or more characters. The Chinese phrase (短語) is a unit composed of one or more words. It is crucial to correctly identify word or phrase boundaries not only for prosodic control, but also for correct selection among multiple pronunciation alternatives. The following examples show changes in pronunciations of the Chinese character 行 in both Mandarin and Cantonese dependeing on the lexical character context.

Mandarin
行 as /**hang2**/: Bank: 銀行 /yin2 **hang2**/
行 as /**xing2**/: Pedestrian: 行人 /**xing2** ren2/
Cantonese
行 as /**hong4**/: Bank: 銀行 /ngan4 **hong4**/
行 as /**hang4**/: Pedestrian: 行人 /**hang4** jan4/

Given a sequence of Chinese characters 銀行人, we can have two different segmentations: 銀行/人 (in Mandarin: /yin2 hang2/ /ren2/, and in Cantonese: /ngan4 hong4/ /jan4/), or 銀/行人 (in Mandarin: /yin2/ /xing2 ren2/, in Cantonese: /ngan4/ /hang4 jan4/). Explicit insertion of word and phrase boundaries can therefore eliminate such ambiguity.

In addition, pronunciation of a character depends on the data object to which it belongs. For instance, the Cantonese pronunciation of 單 /daan1/ changes to /sin6/ when the character is used as a surname as illustrated below.

單 as /**daan1**/ in isolation Singular: **單** /daan1/
單 as /**daan1**/ in word Bicycle: **單**車 /daan1 ce1/
單 as /**sin6**/ in surname "Sin Siu-Ming": **單**小明 /sin6 siu2 ming4/

2.4. Features of Spoken Chinese that Affect Pronunciation

A syllable may carry different tones depending on its meaning, the context or the mode of speaking. The following examples show that in Cantonese, tones carried by the syllables /tong/ (糖), /soeng/ (上) and /man/ (文) change in various situations.

/tong/ with different meanings
tone 4: "sugar" **糖** /**tong4**/ add some sugar: 加**糖** /gaa1 **tong4**/
tone 2: "candy" **糖** /**tong2**/ eat a candy: 食**糖** /sik6 **tong2**/
/soeng/ in different contexts
tone 5: attend a class" 上堂 /**soeng5** tong4/
tone 6: "upstairs" 樓上 /lau4 **soeng6**/
/man/ in different modes of speaking
tone 4: "English" in formal mode 英文 /jing1 **man4**/
tone 2: "English" in colloquial mode 英文 /jing1 **man2**/

3. Proposed extensions related to Text Processing and Pronunciation

3.1. `"dialect-accent"` Attribute

To support the specification of dialects and accents in spoken Chinese, the **`dialect-accent`** attribute is proposed to be used together with **`xml:lang`** attribute defined by XML 1.0. The **`dialect-accent`** attribute is targeted to be

defined as optional in the DTD or XML schema for the backward compatibility with existing VoiceXML parsers.

The proposed value for this attribute is composed of two parts: a primary language subtag for dialect specification and an optional subtag for accent specification. The two parts are separated with a hyphen. The syntax conforms with RFC3066 and is described as follow:

`dialect-accent` =
primary-subtag ("-" optional-subtag)
primary-subtag = 2ALPHA
optional-subtag = 2ALPHA

The format of the *primary-subtag* is defined as a two-alphabet abbreviation of a dialect. For the *optional-subtag*, the format follows straightly the abbreviations of Chinese provinces, autonomous regions and special administrative regions listed in the EDU.CN Domain Policy (中國教育和科研計算機網 EDU.CN 網絡域名註冊辦法), which is defined by the China Education and Research Network Information Centre (CERNET 網絡信息中心) (EDU.CN Domain Policy).

The following table provides some possible values of the **`dialect-accent`** attribute corresponding to the values of the **`xml:lang`** attribute:

***xml:lang* values**	**Dialect**	**Accent**	***dialect-accent* values**
zh-CH	Mandarin	Beijing	MD-BJ
		Guangdong	MD-GD
		Taiwan	MD-TW
	Min	Fujian	MN-FJ
		Taiwan	MN-TW
zh-TW	Mandarin	Beijing	MD-BJ
		Taiwan	MD-TW
	Min	Fujian	MN-FJ
		Taiwan	MN-TW
zh-HK	Cantonese	Hong Kong	CT-HK
		Guangdong	CT-GD
	Mandarin	Hong Kong	MD-HK
		Beijing	MD-BJ
		Taiwan	MN-TW

The **`dialect-accent`** attribute can be used optionally in **`voice`**, **`speak`**, **`p`**, and **`s`** elements. The following example shows how the default dialect and accent can be set respectively to Mandarin and Beijing accent:

```
<p xml:lang="zh-CH"
   dialect-accent="MD-BJ">
   我從北京來的。</p>
```

3.2. "phrase" Element

The **`phrase`** element defines a Chinese phrase. The **`phrase`** element does not have any attribute and it can occur within the content of the **`s`** element. The following elements can occur within the content of the **`phrase`** element: **`audio`**, **`break`**, **`emphasis`**, **`mark`**, **`phoneme`**, **`prosody`**, **`say-as`**, **`sub`**, **`voice`**, **`word`**.

Consider the following Chinese poem:

明年逢春好不晦氣
終年倒運少有餘財

Since there is no punctuation present in the poem, readers need to determine the location of the phrase breaks in order to comprehend it. The poem can be segmented in two ways which lead to completely different interpretations — pessimistic and optimistic.

Pessimistic interpretation	
明年逢春/好不晦氣	break before 好不
終年倒運/少有餘財	break before 少有
Optimistic interpretation	
明年逢春好/不晦氣	break in between 好 and 不
終年倒運少/有餘財	break in between 少 and 有

In the pessimistic interpretation, the word 好不 means "very" and 少有 means "unlikely". When these words are attached to 晦氣("unlucky") and 餘財("savings"), they convey negative meanings:

好不 / 晦氣 ⇒ very unlucky
少有 / 餘財 ⇒ unlikely to have savings

However, in the optimistic interpretation, phrase break positions change the meanings of the characters: 好 ⇒ "good", 不 ⇒ "not", 少 ⇒ "unlikely" and 有 ⇒ "have". As a result, the sentences bring positive meaning.

逢春 (coming sprint)　好 (good) :
　⇒ the coming spring is good
不 (not) 晦氣 (unlucky)　: ⇒ lucky

倒運 (bad luck) 少 (unlikely) : ⇒ lucky
有 (have) 餘財 (savings) : ⇒ have savings

We can apply the <phrase> tag to explicitly define the desired meaning in the markup file through correct phrase breaking.

```
<!-- pessimistic interpretation -->
<phrase>明年逢春</phrase><phrase>好不晦氣
</phrase>
```

and

```
<!-- optimistic interpretation -->
<phrase>明年逢春好</phrase><phrase>不晦氣
</phrase>
```

3.3. "word" Element

The **word** element defines a Chinese word. It does not have any attribute and it can occur within the contents of **s** and **phrase** elements. The following elements can occur within the content of the **word** element: **audio**, **break**, **emphasis**, **mark**, **phoneme**, **prosody**, **say-as**, **sub**, **voice**.

Consider the sentence 這一晚會如常舉行. This sentence has three different interpretations with word boundaries located in different positions:

1	這一	晚會	如常		舉行
2	這一	晚會	如	常	舉行
3	這一晚	會	如常		舉行

In the 1st and 2nd segmentations, the respective meanings are "This banquet is held as usual." and "If this banquet is held frequently," respectively. It should be noted that under these segmentations, the pronunciation of the character 會 in Cantonese is /wui2/. However, in the third semgentation, the pronunciation is either /wui3/ (or /wui5/) because the character 會 is now a word by itself and thus has a different meaning and pronunciation. The meaning of the 3rd segementaion is "(An event) will be held as usual tonight."

The introduction of the **word** element not only defines the boundaries of words, but also helps to generate a correct pronunciation transcription during text processing for synthesis.

3.4. "tone" Element

The **tone** element allows authors to indicate a tone change for a Chinese syllable. This proposed element has one required attribute, **value**, to specify a desired tone. The acceptable values are of type **xsd:nonNegativeInteger** as in XML schema part 2 for datatypes. A number from 1 to 5 is used for for Mandarin and 1 to 6 for Cantonese.

Change of tones depends on the meaning and context of the character as well as the mode of speaking. The followings are some examples of using the tone element in various situations.

Tone changes depending on meanings	
tone 4: "sugar" **糖 /tong4/**	
	加<tone value="4">**糖**</tone>
tone 2: "candy" **糖 /tong2/**	
	食<tone value="2">**糖**</tone>
Tone changes depending on contexts	
	<tone value="5">上</tone>堂
	樓<tone value="6">上</tone>
Tone changes depending on modes of speaking	
[formal]	英<tone value="4">文</tone>
[colloquial]	英<tone value="2">文</tone>

4. Proposed Legal Values for the "interpret-as" Attribute

The **interpret-as** attribute is a required attribute of the **say-as** element which belongs to the *Document Structure, Text Processing and Pronunciation* category in SSML 1.0. It allows authors to indicate the type of text contained within the element. The **say-as** element also has two other optional attributes, **format** and **detail** to fulfil the need of different values set to the **interpret-as** attribute. To provide more flexibility to text verbalization in Chinese, the following values are proposed: **Chinese-name**, **fraction**, **measure**, **net**, **percentage**, and **ratio**.

4.1. "Chinese-name" Value

The "**Chinese-name**" value is proposed for the specification of Chinese names. This value is needed as a Chinese character may be pronounced differently when it appears in a name. For example, the character 單 as explained in Section 2.4.

When "Chinese-name" is used as the value for the **interpret-as** attribute, the **format** attribute is also required to specify the distribution of surname and given name. The syntax of the value of **format** is composed of two parts: alphabet "S" to represent a surname and alphabet "G" to represent a given name. Multiple "S"s and "G"s can be used to indicate a corresponding number of characters that represents surname and given name. The syntax is described below:

format = S*G*
"" indicates one or many occurences*

The following example shows the characters 單明明 can be correctly pronounced using the "Chinese-name" value. Consider the sentence:

單明明的成績單明明是給你拿走了
("It is obvious that you have taken away Sin Ming Ming's transcript.")

The first occurrence of 單明明 is a name, whereas the second occurrence of it should be segmented as (成績)單 (i.e. "transcript") and 明明 (i.e. "obvious"). Therefore, the correct pronunciations of the two occurrences should be /**sin6** ming4 **ming2**/ and /**daan1** ming4 **ming4**/ respectively.

```
<!-- the first occurrence of 單明明 is a
Chinese name -->
<say-as interpret-as="Chinese-name"
format= "SGG">
    單明明 </say-as>
    的成績單明明是給你拿走了
```

4.2. "fraction" Value

The "fraction" value is used when the contained text should be verbalized as a fraction. To verbalize a fraction in Chinese, the denominator is pronounced first, followed by the additional word 分之 (A 分之 B means B "out of" A) and the numerator at the end. For example, the fraction 3/4 is verbalized as 四分之三 (i.e. "three out of four").

When "fraction" is used as the value of the **`interpret-as`** attribute, the **`format`** and **`detail`** attributes are not required.

```
<!-- verbalize 3/4 as 四分之三 -->
<say-as interpret-as="fraction">3/4
</say-as> 個橙
<!-- three quarters of an orange. -->
```

4.3. "measure" Value

The "measure" value is used when the contained text should be verbalized as a measurement, such as 10cm, 5m and 30ml.

The amount in front of the measurement unit will be interpreted as number as in VoiceXML 2.0 (i.e. 10 is verbalized as ten but not one zero). The unit is translated and pronounced in Chinese, for example, "cm" is "厘米" /lei4 mai5/. The translations of measurement units are listed below:

mm	毫米	l	升	oz	安士
cm	厘米	mg	毫克	ft(s)	呎
m	米	g	克	in	吋
km	千米	kg	千克	yd	碼
ml	毫升	lb(s)	磅		

When "measure" is used as the value of the **`interpret-as`** attribute, the **`format`** and **`detail`** attributes are not required.

```
<!-- verbalize 180cm as 一百八十厘米 -->
<say-as interpret-as="measure">
180cm</say-as>
```

4.4. "net" Value

The "net" value is used when the contained text is a URL or an email address. Since URLs and email addresses are usually written in English, it is inappropriate for a Chinese text-to-speech synthesizer to convert the contained text into Chinese. In this case, the Chinese text-to-speech synthesizer should spell out the English alphabet.

When "net" is used as the value of the **`interpret-as`** attribute, the **`format`** attribute is required. The possible values for the **`format`** attribute are "**email**" and "**uri**".

```
<!-- spell out the contained URL -->
<!-- the whole sentence is verbalized as
H T T P colon slash slash W W W dot C U H
K dot E D U dot H K -->
<say-as interpret-as="net" format="uri">
   http://www.cuhk.edu.hk
</say-as>
```

4.5. "percentage" Value

The "percentage" value is used when the contained text should be verbalized as a percentage. To verbalize a percentage in Chinese, the words 百分之 (i.e. "out of a hundred") are added before the number. For example, the percentage "70%" is verbalized as 百分之七十 (i.e. "seventy out of a hundred").

When "percentage" is used as the value of the **`interpret-as`** attribute, the **`format`** and **`detail`** attributes are not required.

```
<!-- verbalize 70% as 百份之七十 -->
<say-as interpret-as="percentage">
70%</say-as>
```

4.6. "ratio" Value

The "ratio" value is used for the **`interpret-as`** attribute when the contained text should be verbalized as a ratio. To verbalize a ratio in Chinese, the word "比" (i.e. "to") is added. For example, the ratio 1:99 is verbalized as 一比九十九 (i.e. "one to ninety nine").

When "ratio" is used as the value of the **`interpret-as`** attribute, the **`format`** and **`detail`** attributes are not required.

```
<!-- verbalize 1:99 as 一比九十九 -->
<say-as interpret-as="percentage">
1:99</say-as>
```