SSML Extensions Aimed To Improve Asian Language TTS Rendering

Jilei Tian[*], Xia Wang[+], and Jani Nurminen[*]
Multimedia Technologies Laboratory
**Nokia Research Center, Finland*
+Nokia Research Center, China

1. INTRODUCTION

The Speech Synthesis Markup Language (SSML) is a standard way of producing content to be spoken by a speech synthesis system (also referred to as text-to-speech (TTS) system), led by the W3C Voice Browser Working Group**.** SSML is an XML-based markup language, which is aimed to control a TTS system in a variety of application contexts.

To make SSML more useful in the current and emerging markets, non-English languages should also be taken into account. The main objective of this proposal is to identify and prioritize extensions and additions to SSML that will improve the use of SSML for rendering Asian languages including Chinese languages, Indian languages, Thai, Vietnamese, etc.

2. POSITION

Both formant synthesis based and concatenative acoustic unit based TTS systems have been developled in Nokia. Many non-English languages have been considered in the development work, and Nokia's Mandarin Chinese TTS system is under continuous development within the TC-STAR framework (www.tc-star.org). To meet the needs of the TTS evaluations in TC-STAR, common interfaces for the input and all the internal modules have been carefully defined. SSML has been taken into use as the input format, and Nokia has proposed extensions related to the Asian language peculiarities.

3. PROPOSAL

3.1 Background on Asian languages

Asian languages are different from western languages from a few perspectives: first, many Asian languages are tonal languages, e.g. Chinese including different spoken variants called dialects, Vietnamese, Thai, are all languages in which tones play an important semantic role; second, many Asian languages are syllabic in nature, e.g. Chinese, Vietnamese, Thai, Hindi; third, many Asian languages (e.g. Chinese, Thai, Vietnamese, Japanese) do not have a proper boundary for words. These peculiarities cause special challenges for TTS systems, and consequently, require special treatment in SSML.

In the Asian language TTS systems, tones could be treated as suprasegmental features and controlled in the suprasegmental layer, or tones could be treated within the basic unit for concatenation of corpus-based TTS systems, e.g. syllable or final level. Different languages have different tones, and therefore SSML should have support for handling tonal aspects.

Tone Sandhi is another problem that should be considered to better support Asian languages. Although every syllable has a default tone when pronunuced in isolation (called a lexical tone), the tone might change inside a word due to the context. Such a phenomenon is called tone sandhi. Many of the tonal languages have tone sandhi. If the tone changes are not handled in a proper way, the word could be perceived as a totally differerent word.

In syllabic languages, syllables or sub-syllabic structures like initials and finals could be very natural units for the TTS systems (this is the case for Chinese, Vietnamese, Hindi, etc). Phonemes can be good units for western languages, but not very good for Asian languages. We should respect the characteristics of the language to be processed and give more flexibility in SSML to better support different languages.

Word segmentation, for Asian languages such as Chinese, Thai, Japanese, Vietnamese etc., is an essential part of the text processing for TTS.

Another trend in Asia is the use of loaned words in daily communication. If foreign words, URLs, e-mail addresses or acronyms are mixed with local texts, they should be taken care of with special efforts. We propose multilingual support in SSML.

3.2 SYLLABLE element <syllable>

For syllabic languages, it is natural to use syllables or tonal syllables as the basic unit of a TTS system. For some languages such as Chinese, Thai, Hindi, etc., the syllables written as in the local language might not offer the best way for computers to read. Orthorgraphic plus romanized form or transliteration would be a better representation for SSML. For Mandarin Chinese, we could use HanYu pinyin; for Cantonese, we could use Jyutping propsed by LSHK; for languages that do not have a standard Romanization or transliteration system, we could use proposals from local research society and provide room for later changes.

3.3 WORD element <word>

The word segmentation is a very crucial issue in languages that don't have word boundaries (e.g. Chinese) as described in section 3.1. Word element is also crucial for tone sandhi, in which the real tone and the default tone will both be documented. We proposed a scheme for tone sandhi in SpeeCon and LC-STAR project, i.e. <baseform romanization><default tone><real tone>. For example, Mandarin Chinese “五百” will be pronounced as “wu2 bai3” instead of “wu3 bai3” with default lexical tones. We propose to use “wu32 bai3” to mark the tone sandhi. The <word> element is proposed to enhance the word segmentation and tone sandhi, in case that automatic word segmentation does not work or user forces the system to take a certain word segment. Its attribute is the segmented word.

Inside <word> element, <pos> could be optional, but it could be used for determining the pronunciation of given word, etc, in the case where the POS tagging does not work or the user forces the system to use a certain POS tag. <break> can be used for defining the break strength at the boundary (character boundary, word boundary, prosodic phrase boundary, sentence boundary, etc).

For some attributes and their values, some additions are required. For example, <encoding> should include GB code, BIG5, etc. <phoneme ph=”…..”> should include Chinese PinYin with tone markers, such as “xia4 wu3”.

Example:

```
<?xml version="1.0" encoding="GB2312" ?>
<speak version="1.0" xml:lang="cn">
        <word word="下午">
          <pos>
            <NOUN />
          </pos>
          <break strength="weak" />
        </word>
</speak>
```

3.4 <prosody> element update

For tonal languages like Chinese, the pitch contours play a very important role in rendering TTS speech. The same phoneme sequence or the same baseform syllable with a different tone leads to completely different meanings. Therefore, it is recommended to enhance the descriptions on prosodic features, particularly on pitch. We propose to give opportunity to describe the prosody features in (time, value) format. This approach gives the possiblity to cover any prosodic needs.

Hereafter we use Mandarin Chinese as an example language.

As can be seen, <syllable> is used to define the given character. <frequency> and <energy> are introduced to describe prosodic features, pitch and volume, in the (time, value) format in order to have a better representation capability for prosodic features.

Example:

```
<?xml version="1.0" encoding="GB2312" ?>
<speak version="1.0" xml:lang="cn">
        <word word="下午">
          <pos>
           <NOUN />
           </pos>
          <syllable syl="下">
           <frequency>
                 <pair time="0" value="380" />
                 <pair time="80" value="363" />
                 <pair time="160" value="340" />
                 <pair time="240" value="301" />
           </frequency>
           <energy>
                 <pair time="267" value="74" />
           </energy>
           <break strength="none" />
          </syllable>
          <syllable syl="午">
           <frequency>
                 <pair time="0" value="290" />
                 <pair time="54" value="285" />
                 <pair time="108" value="285" />
                 <pair time="162" value="290" />
            </frequency>
          - <energy>
                 <pair time="181" value="71" />
            </energy>
              <break strength="weak" />
          </syllable>
        </word>
</speak>
```

One thing that needs to be noted is that the frequencies defined above offers an illustrative example of the pitch contour representation, and we are aware that the absolute frequency is not usable in practice due to the natural variations between different speakers. Some kind of relative frequencies would offer a better solution in practice. However, the (time, value) format could also be used in this case. Another

proposal is to define the tone contours out of the word level and to use those definitions in the word level; therefore we do not need to specify the frequencies for each word.

Tones are suprasegmental features. For parameter-based synthesis, tones are defined as another feature in the suprasegmental level, like duration and stress.

3.5 Multi-lingual support

Possible extensions to SSML also include better support for different aspects related to multilinguality. Here, multilingualility could be understood in one of the following senses:

- Multilingual texts running mixed together. Proper synthesis engines need to be selected for the texts in different languages;

- Loan words written in the local language. Arguments exist whether the words should be spoken using the local language or a foreign language. For example, Japanese will write all the loan words in Katakana, and Thai will also write the loan words in Thai.

- Abbreviated representations like “asap” (as soon as possible), “cu” (see you) and special Internet-based symbols like smiles. Different synthesis strategy is needed for those special things.

4. CONCLUSION

Some extensions to SSML have been proposed in section 3. The extensions reflect the XML development work for a Mandarin Chinese TTS system and the research on other Asian languages that we have carried out in Nokia. Some of proposed additions can definitely be combined into the current SSML standard, but the main aim has been to show the way we think and list some issues that we consider to be of particular importance from the viewpoint of Asian tonal and syllabic languages. Our proposals aim at initializing profound thinking and discussions.