Prosody and Style Controls in CU VOCAL using SSML and SAPI XML Tags

Tien-Ying Fung*, Yuk-Chi Li*, Helen Meng* and P. C. Ching**

*Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, **Digital Signal Processing Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

{tyfung, ycli, hmmeng@se.cuhk.edu.hk, pcching@ee.cuhk.edu.hk}

Abstract

CU VOCAL is a Cantonese text-to-speech (TTS) engine. We use a syllable-based concatenative synthesis approach to generate intelligible and natural synthetic speech in Cantonese. This paper reports on our recent enhancements in CU VOCAL to support user adjustments in prosody and style with the use of the Speech Synthesis Markup Language (SSML) [1, 2] in the input text. CU VOCAL was previously developed as a SAPI-compliant engine to enable easy integration with other applications. This paper also reports on our enhancements in the CU VOCAL SAPI engine to support the SAPI 5 XML tags [3].

1. Introduction

The Speech Synthesis Markup Language (SSML) Specification [1, 2] is a W3C standard that supports browsing of Web content by voice. SSML is designed to provide an XML-based markup language for assisting the generation of synthetic speech in Web and other applications. For example, a Cascading Style Sheet Level 3 working draft which defines aural properties that give control over rendering XML to speech using SSML have been proposed by W3C at the writing of this document [4]. The use of the style sheet properties on the Web for text-to-speech synthesis can be used together with Web form controls (e.g. buttons) to provide accessibility to visually impaired group. CU VOCAL thus is enhanced to support some SSML tags which are applicable to Chinese speech synthesis. CU VOCAL is a Cantonese text-to-speech (TTS) engine that we have developed using a syllable-based concatenative synthesis approach [5, 6]. CU VOCAL accepts free-form Chinese text as input, and generates highly natural Cantonese speech with a Hong Kong female voice. CU VOCAL also accepts Arabic numeric expressions and it is able to spell out English words by pronouncing individual English alphabets. This paper focuses on the recently supported input format of CU VOCAL, the “Prosody and Style” SSML tags and SAPI 5 XML tags (for CU VOCAL SAPI engine), and reports on how the tags are handled by CU VOCAL.

According to the SSML Specification Version 1.0, there are four tags which belong to the categories of “Prosody and Style”: prosody, emphasis, voice and break. CU VOCAL has been enhanced to support the first two – the prosody and the emphasis tags1. The prosody tag is used to control the pitch level, speaking rate and volume of the speech output, and the emphasis tag is used to request that the contained text being spoken with emphasis. To achieve the same effect, Microsoft’s Speech Application Programming Interface (SAPI) [7] has developed another markup, the SAPI 5 XML tags [3]. Four tags are used to allow user adjustment over prosody and emphasis in SAPI-compliant speech synthesis engine, they are pitch, rate, volume and emph. CU VOCAL supports this set of tags by converting them to the equivalent SSML tags to avoid duplication in implementation. The following sections introduce and discuss how the prosody and emphasis SSML tags are supported by CU VOCAL, as well as how the SAPI 5 XML tags are supported.

2. The Prosody Tag

The prosody tag in SSML 1.0 permits control of the pitch, speaking rate and volume of the speech output. The six optional attributes of this tag are pitch, contour, range, duration and volume:

- pitch -- set pitch relatively or by number of Hz,
- contour -- set actual pitch contour,
- range -- set pitch range,
- rate -- adjust speaking rate relative,
- duration -- adjust duration,
- volume -- change the volume of the element content

Currently, CU VOCAL supports the pitch, rate and volume attributes.

2.1 The pitch Attribute

The pitch attribute is used to adjust the pitch level of the synthetic speech. The usage of this attribute is

<prosody pitch="value"> contained text </prosody>

The legal values of this attribute include a number followed by the unit “Hz” (e.g. 16000Hz) and a relative change (e.g. +10Hz, -20%). This attribute also accepts pre-defined values:

1 The voice tag, which requests a change in speaking voice; and break tag, which controls the pausing or other prosodic boundaries between words, are not yet supported by CU VOCAL.
X-LOW, LOW, MEDIUM, HIGH, X-HIGH and DEFAULT. When the value is set to DEFAULT, CU VOCAL will not perform any pitch adjustment. For the other values, CU VOCAL performs pitch adjustment to the synthetic speech of the contained text to support the prosodic effects accordingly. Figure 1 shows the pre-defined values and the equivalent pitch values rendered by CU VOCAL. An example of the usage of this attribute with the sentence “我是女低音” (meaning: I am an alto) is shown in Figure 2.

<table>
<thead>
<tr>
<th>x-low</th>
<th>15kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>15.5kHz</td>
</tr>
<tr>
<td>medium</td>
<td>16.25kHz (a bit higher than default)</td>
</tr>
<tr>
<td>high</td>
<td>16.5kHz</td>
</tr>
<tr>
<td>x-high</td>
<td>17kHz</td>
</tr>
<tr>
<td>default</td>
<td>16kHz (no change)</td>
</tr>
</tbody>
</table>

Figure 1. The pre-defined values of the pitch attribute for the prosody tag and their equivalent pitch values rendered by CU VOCAL.

2.2 The rate Attribute

The rate attribute is used to adjust the speaking rate of the synthetic speech. The usage of the rate attribute is illustrated below:

```xml
<prosody rate="value"> contained text </prosody>
```

The legal values of this attribute include a relative change specified with a percentage or a number, e.g. +2% (faster by two percent), 2 (twice the default rate), 0.5 (half the default rate). The attribute also accepts pre-defined values: SLOW, MEDIUM, FAST, X-FAST and DEFAULT. Similar to the pitch attribute, DEFAULT refers to the original speaking rate of CU VOCAL.

When other values are used, CU VOCAL will adjust the speaking rate according to the value specified with the rate attribute. The speaking rate adjustment is achieved by performing elongation or elimination of pitch cycles to the syllables contained in the synthetic speech. To speed up the synthetic speech, CU VOCAL duplicates the pitch cycles and evenly adds them to the syllable nuclei. To facilitate the adding and dropping of pitch cycles, phoneme-base alignment and pitch tracking and marking are performed beforehand. Pitch tracking is achieved with the approach of Normalized Cross Correlation Function [8] using the Hidden Markov Model Toolkit [9] and pitch marking is achieved with dynamic programming.

Figure 3 shows the implementation of CU VOCAL for the pre-defined values. Figure 4 illustrates the use of this attribute with the sentence “我趕時間要走先” (meaning: I am in a hurry and have to leave now) is generated with a faster speaking speed in order to give an impression of being in a hurry, while “慢慢傾” (meaning: take your time) in the second phrase is generated with a slower speed.

2.3 The volume Attribute

The volume attribute allows user to adjust the volume of the synthetic speech of the contained text. The usage of the volume attribute is illustrated below:

```xml
<prosody volume="value"> contained text </prosody>
```

The legal values of this attribute include a number ranges from 0.0 to 100.0 or a relative change specified with a number (e.g., +10). This attribute also accepts pre-defined values: SILENT, X-SOFT, SOFT, MEDIUM, LOUD, X-LOUD and DEFAULT. CU VOCAL enables volume adjustment by the rising or lowering of the energy level of the synthetic speech of the contained text. The value 50 is set as the default volume (equivalent to DEFAULT), which is the original volume of CU VOCAL’s voice. The values 0 to 100 render volumes from silence to double that of the default volume. The mapping of the pre-defined values and the numeric values is defined as shown in Figure 5. Figure 6 illustrates an example of using the volume attribute with the sentence “現在已經係夜深，請將音量收細” (meaning: it is late night now, please turn down the volume).

```xml
<prosody volume="x-soft"> 請將音量收細 </prosody>
```

Figure 5. The mapping table of the pre-defined values and the numeric values of the rate attribute of the prosody tag defined by CU VOCAL.
(meaning: please turn down the volume) will be generated with a softer volume.

An integrated example of using the three attributes of the \texttt{prosody} tag is given in Figure 7.

![Example of using the three attributes of the \texttt{prosody} tag](image)

Figure 7. An integrated example illustrating the use of the \texttt{pitch}, \texttt{rate}, and \texttt{volume} attributes of the \texttt{prosody} tag in a single paragraph to produce different prosodic effects for different phrases. E.g. the last phrase “買平機是時候” has a larger volume and a higher pitch.

### 3. The Emphasis Tag

The \texttt{emphasis} tag indicates that the contained text will be spoken with emphasis. This tag comes with an optional attribute \texttt{level}, which specifies the strength of emphasis. The usage of the \texttt{emphasis} tag and its attribute is illustrated below:

\begin{verbatim}
<emphasis [level="value"]> contained text </emphasis>
\end{verbatim}

The pre-defined values for \texttt{level} are \texttt{STRONG}, \texttt{MODERATE}, \texttt{REDUCED} and \texttt{DEFAULT}. If \texttt{level} is not specified, the value \texttt{MODERATE} is used as default. CU VOCAL render emphasis by changing the loudness, speaking rate and pitch of the synthetic speech of the contained text. When \texttt{level} is set to \texttt{STRONG}, CU VOCAL generates a louder voice with higher pitch level and speaks slower; while \texttt{MODERATE} (equivalent to \texttt{DEFAULT}) is similar to \texttt{STRONG} but with a smaller extent. For \texttt{REDUCED}, CU VOCAL speaks faster and with a softer voice.

CU VOCAL implements the emphasis levels using the combination of adjustments in pitch, rate and volume with the three attributes of the \texttt{prosody} tag. The relations between the emphasis levels and the values of the \texttt{pitch}, \texttt{rate} and \texttt{volume} attributes is summarized in Figure 8. An example of using the “emphasis” tag is illustrated in Figure 9.

![Example of using the \texttt{emphasis} tag](image)

Figure 9. An example illustrating the use of the \texttt{emphasis} tag. The levels \texttt{STRONG}, \texttt{MODERATE} and \texttt{REDUCED} are used to demonstrate different degrees of emphasis rendered within the same paragraph.

### 4. SAPI 5 XML tags

Speech Application Programming Interface (SAPI) [7] has also developed their own synthesis markups for prosody and emphasis adjustment, which is referred as the SAPI 5 XML tags. A recent version of CU VOCAL was developed to achieve SAPI-compliance and was named CU VOCAL SAPI engine, as described in [10]. In this paper, we report on further enhancement to CU VOCAL SAPI engine in order to support four SAPI 5 XML tags. This gives users the flexibility to customize the speech output with controls over pitch, speaking rate, volume, and emphasis in their speech-enabled application developed using the CU VOCAL SAPI engine.

The SAPI 5 XML tags for pitch, speaking rate, volume and emphasis control are \texttt{pitch}, \texttt{rate}, \texttt{volume} and \texttt{emph} respectively. CU VOCAL handles these tags by mapping them to the equivalent SSML tags to achieve the same effect in prosody and emphasis adjustment. The descriptions of the four SAPI 5 XML tags are listed in Figure 10 [3].

\begin{verbatim}
<pitch>

The \texttt{pitch} tag controls the pitch of a voice. The tag can be empty, in which case it applies to all subsequent text, or it can have content, in which case it only applies to that content.

Attributes: \texttt{middle}, \texttt{absmiddle}

One of the above attributes must be present. Both attributes control the relative pitch value of the voice. The difference between the two attributes is that when scoped, the \texttt{middle} attribute is relative whereas the \texttt{absmiddle} attribute is absolute.

The value of both of these attributes should be an integer between negative fifty and fifty, which corresponds to the values -50% and +50% in the \texttt{pitch} attribute of the SSML \texttt{prosody} tag. Values outside of this range will be truncated by CU VOCAL. A negative sign in the values means deduction and a plus sign means increment. E.g. A value of -10 decreases the default CU VOCAL pitch level by 10%, and 20 increases the default pitch level by 20%.

\end{verbatim}
Example:

```
<pitch middle="15">中文大学</pitch>
```

increases the pitch level of the phrase “中文大學” by 20% (when scoped, the middle attribute gives relative changes in pitch values).

```
<rate>
The rate tag controls the rate of a voice. The tag can be empty, in which case it applies to all subsequent text, or it can have content, in which case it only applies to that content.

Attributes: speed, absspeed
One of the above attributes must be present. Both attributes control the relative rate of the voice. The difference between the two attributes is that when scoped, the speed attribute is relative whereas the absspeed attribute is absolute.

The values of these attributes should be an integer between negative ten and ten, which corresponds to the values -10% and +10% in the rate attribute of the SSML prosody tag. Values outside of this range will be truncated by CU VOCAL. A negative sign in the values means deduction and a plus sign means increment. E.g. A value of -10 decreases the default CU VOCAL speed by 10%, and 5 increases the default speed by 5%.

Example:
```
<rate absspeed="8"><rate absspeed="5">中文大學</rate></rate>
```

increases the speaking rate of the phrase “中文大學” by 5% (when scoped, the absspeech attribute gives absolute change in rate).

```
<volume>
The volume tag controls the volume of a voice. The tag can be empty, in which case it applies to all subsequent text, or it can have content, in which case it only applies to that content.

Attribute: level
The value of this attribute should be an integer between zero and one hundred, which corresponds to 0 and 50 in the volume attribute of the SSML prosody tag. Values outside of this range will be truncated by CU VOCAL.

One hundred represents the default volume of the speech engine (in our case, CU VOCAL) as suggested by SAPI specification. Lower values represent percentages of this default. That is, 50 corresponds to 50% of the original volume.

Example:
```
<volume level="50">中文大學</volume>
```

reduces the volume of the phrase “中文大學” by half.

```
<emph>
The emph tag places emphasis on the text contained by this element.

No attributes.
```

The effect of the emph tag is equivalent to the moderate level of the SSML emphasis tag.

Example:
```
<emph>中文大學</emph>
```

emphasizes the phrase “中文大學” with a moderate level.

Figure 10. The description of the pitch, rate, volume and emph SAPI 5 XML tags.

5. Summary and Conclusions

This paper describes our work in enhancing our Cantonese TTS synthesizer, CU VOCAL, to support user adjustments in prosody and emphasis using SSML 1.0 tags or SAPI 5 XML tags. We have implemented two tags of SSML 1.0 which are related to prosody and style, they are prosody (with pitch, rate and volume as attributes) and emphasis. We have also handled four SAPI 5 XML tags, pitch, rate, volume and emph, to allow similar prosody and style adjustments in CU VOCAL SAPI engine.

6. Acknowledgments

This work is substantially supported by the Innovation and Technology Fund from the Hong Kong SAR Government (project code CUHK ITS/117/01) and also by a grant from Microsoft Research. We also wish to acknowledge Mr. Chi-Kin Keung for his contributions to speed control in CU VOCAL, and Dr. Wai-Kit. Lo for his valuable comments.

7. References

4. CSS3 Speech Module working draft: http://www.w3.org/TR/2004/WD-css3-speech-20040727/
7. SAPI 5.1 http://www.microsoft.com/speech/