Chinese Romanization for Chinese Voice Browsing

Zhiwei Shuang, Fanping Meng, Qin Shi, Liqin Shen

IBM China Research Lab

Abstract:

This document focuses on extending the Speech Synthesis Markup Language for better support of Chinese. We propose to use Chinese Romanization to annotate Chinese pronunciation in "phoneme" element. We also propose SSML to use diverse predefined and widely used pronunciation annotation standards for different languages, at least as a complement to the created general standard. Thus SSML can be more easily accepted and used around the world.

1. Introduction to Chinese Romanization

Chinese is quite different from most western languages. First of all, a Chinese character represents a meaning more than a pronunciation. The homograph phenomenon is very common for Chinese characters. The same Chinese character can have many different meanings and different pronunciations. It is very difficult to be 100% sure of its pronunciation without understanding its meaning. That is one of the most difficult problems for Chinese TTS systems. So it will be very helpful if the pronunciation can be given explicitly.

The International Phonetic Alphabet [IPA] is a set of phonetic characters generally used for describing the pronunciation. It tries to collect an exhaustive set of pronunciations for all kinds of languages. Due to its exhaustive coverage of languages, it has become very complicated and difficult to learn. Thus, even a well educated Chinese adult can not annotate Chinese Pronunciation in IPA without special training. Besides this, there are special linguistic phenomena in Chinese that can not be conveniently described by IPA, such as "tone", "neutral tone" and "retroflex".

In comparison with IPA, Chinese Romanization is much more convenient for annotation of Chinese Pronunciation.

First, Chinese Romanization is specially designed only for Chinese instead of all languages. Thus, a small and simple alphabet set is enough to distinguish different pronunciations in Chinese. It is much easier to remember and edit. Meanwhile, some special rules can be designed to describe the special phenomena of Chinese. For example, adding 'r' in the end to describe a "retroflex" syllable.

Secondly, Chinese Romanization is widely used and learnt. Every student in primary school of China Mainland learns Chinese Characters together with their Chinese Romanization. Many foreigners also begin to learn Chinese by Chinese Romanization. For many computer users, they use Chinese Romanization to input Chinese Characters on computer.

Meanwhile, Chinese government has brought into effect a standard for Chinese Romanization since 1996 [1]. This standard defines the transliterating rules and writing format using "Hanyu Pinyin Scheme". It is in effect for education, publishing, information processing and other related industries in China.

Because of all the above reasons, Chinese Romanization is fit to annotate the pronunciation of Chinese.

2. Use Chinese Romanization in phoneme element

The writing rules of Chinese Romanization conform to P.R.C state standard “Basic rules for Hanyu Pinyin Orthography” [1] published by *China State Bureau of Quality and Technical Supervision* (CSBQTS) in 1996. According to the naming method of alphabet, we propose to use “x-CSBQTS-96” to represent Chinese Romanization alphabet.

```
<?xml version="1.0"?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
       xml:lang="zh-CH">
  <phoneme alphabet=" x-CSBQTS-96" ph="duìbuqǐ"> 对不起 </phoneme>
  <!-- This is an example of Chinese Romanization -->
</speak>
```

As to [1], the tone annotation can be replaced by number or letter for technical convenience. We propose to use 1, 2, 3 and 4 to annotate the 4 normal tones of Chinese while using 0 to annotate the “neutral tone” phenomenon.

```
        <?xml version="1.0"?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
       xml:lang="zh-CH">
  <phoneme alphabet="x-CSBQTS-96" ph="dui4bu0qi3"> 对不起 </phoneme>
  <!-- This is an example of Chinese Romanization
       using number to describe tone -->
</speak>
```

3．Extension for other language or dialects

The goal of SSML is to “provide a rich, XML-based markup language for assisting the generation of synthetic speech in Web and other applications”. To reach the goal, we need more and more users of SSML, such as ordinary Web applications developers, to learn and use the SSML easily. So, we need to define the SSML based on ordinary people’s knowledge and skill rather than professional linguistics’ knowledge. Otherwise, it will be a long way for SSML be widely accepted and used.

Many languages and dialects have formed up their own standards to annotate the pronunciation. For example: the Linguistic society of Hong Kong has published a simple, easy-to-learn and easy-to-use “LSHK Cantonese Romanization Scheme” in 1993 [2]. The characters used in this scheme are basic Latin characters and numbers, therefore can be input

using an ordinary computer keyboard. Thus, this scheme is widely adopted in various areas: education, Cantonese information process and computer input method, etc. So we also propose to use “The LSHK Cantonese Romanization Scheme” to annotate Cantonese pronunciation.

It is possible to form up a general standard to annotate all languages’ pronunciation, however such a standard may become very difficult to use. Another way is to use the predefined and widely accepted pronunciation annotation standards for different language. At least, these diverse standards should be an important complement to the general standard.

Reference:

1. Basic rules for Hanyu Pinyin Orthography, National Standard of P.R.C, GB/T 16159—1996, published in 1996-01-22, implemented from 1996-07-01, published by *China State Bureau of Quality and Technical Supervision*.
2. “The LSHK Cantonese Romanization Scheme”, published in 1993, published by *Linguistic Society of Hong Kong*, Website: http://www.iso10646hk.net/jp/index.jsp.