The Concept of Speech Synthesis Markup Language

Lixing Huang Jianhua Tao
National Laboratory of Pattern Recognition (NLPR)
Institute of automation, Chinese Academy of Sciences
{lxhuang, jhtao}@nlpr.ia.ac.cn

1 Introduction

Synthetic speech close to natural sounding can be heard now a day. Recent advancement of multimedia interfaces between man and machine largely increased interests on Speech Synthesis Markup Language which could be used to control the speech synthesis system to generate more expressive speech and extremely extends its functions in human machine interaction.

As we know, the speech synthesis system tries to interpret the inputted text and gives the naturally synthesized speech. Although many technologies and models have been used to generate the synthesized speech in reading style, there are still lots of problems in pronunciation and prosody prediction. Some markup languages, for instance, SSML, SABLE, VoiceXML, and so on, were proposed to solve these problems. Existing markup languages provide the users kinds of interfaces to control the speaking style. The earliest markup language called SSML (different from the W3C SSML) contains four elements, and has the ability to mark the phrase boundaries, indicate the emphasis word, specify pronunciation and embed other sound files. After SSML, other markup languages appear one after the other. In 1998，the SABLE specification was coauthored mainly by researchers from University of Edinburgh and Bell Labs. Like other markup languages, the SABLE has elements to describe the text structure and acoustic parameters. The VoiceXML is designed mainly for creating audio dialogs. It has two types of dialogues use different ways to gather input from users. VoiceXML also has a self-contained execution environment for interpreting markups at run-time. It is mainly used for controlling the dialogue flow. The current W3C Speech Synthesis Markup Language is developed mainly based on the JSML and SABLE. It actually provides comparatively rich elements to control and guide the generation of synthetic speech. The functions of these elements cover the text structure, pronunciation, text processing, and prosody and so on.

Different markup languages have different emphases. Although one markup language cannot cover all aspects, it should contain at least three functions, which are speaking styles control, module analysis and debugging and the extensions in multimedia and communication. With the development of speech synthesis, the multilinguality becomes more and more important, markup languages should also take it into account.

2 Basic functions of Markup Language

Different markup languages have different elements, emphasize particularly on different aspects. No matter how many elements they have, three basic functions should be contained.

2.1 Generating different speaking styles

There are many acoustical parameters which can affect the result of synthetic speech. And we think it should be able to control most of these parameters through the markup language. A powerful markup language must have the ability to describe acoustical characteristics very well. As we all know, the prosody is the most important aspect of all the acoustical parameters. Whether a markup language characterizes the prosody well or not determines its performance. So there should be some elements in the markup language to describe the pitch range, the pitch contour, and the duration and so on. It is our aim that we can describe the characteristics of the prosody accurately from these elements. For example, in order to specify the prosody contour, we may markup certain key points as follows:

<prosody contour= "(0%,default)(30%,+20%)(70%,+50%)">小溪流</prosody>

Also, there are other parameters will affect the quality of the synthetic speech. For example, a pause always exists in the place where punctuation occurs, or between prosodic phrases, and the length of the pauses will be different at different places. So it is useful to set the length of pauses according to the context. For instance, we can set the length of a pause in the following way:

无数的小溪流<break strength="weak"/>最终汇成了大海 Other valuable parameters include accent, tone and so on.

Except all the above, there is an important problem we should pay attention to, which is some words can be pronounced in different ways. For example, "1984", we can pronounce it as one thousand nine hundred and eighty four, and also can pronounce as nineteen forty eight. A markup language should contain certain element to distinguish the differences. For example,

他出生于<sayas type="date">1984</sayas>年

2.2 Module analysis, debugging and evaluation

Generally, there may be three modules in a speech synthesis system, text analysis module, prosody analysis module and acoustic module. Information between different modules is transferred by proper data structure, and finally sound wave generated by last module. All that we can get from the system explicitly is the sound, but there are not any medial results output. If we want to get the medial analysis result, we have to debug our programme and place break points in the appropriate places and watch the values of corresponding variables. But with the SSML and XML synthesizer, we can not only hear the synthetic speech, but also get medial analysis results explicitly. We analyze medial results to examine whether each module performs well or not. This makes debugging easier to a great extent.

Furthermore, we can utilize some tagged text to test each module. Without markup languages, plain text is the input of the system. If we want to observe the performance of each module in all possible conditions, lots of plain text need to be picked up manually. It will be a tough work. But if the input is the markup language, what we need do is just to modify the values of proper attributions of elements. Different conditions will be simulated easily. For example, if we want to know the effect when speakers pronounce rapidly, all we need to do is to set the value of an attribute, such as, *<prosody rate="x-fast">丛林小溪水</prosody>*. Obviously, it is more efficient than picking up plain text for test.

The reason why a speech synthesis system sounds natural is good performance of each module. Different systems may have different advantages. Some are good at text analysis, some are good at prosody analysis, but the medial results produced by each module cannot be shared because different systems have different data structures. It is no doubt that markup language is an efficient way to solve this problem. The medial analysis results represented in the form of certain markup language based on XML, such as SSML, is independent from system. This makes sharing more convenient.

2.3 The extensions in multimedia and communication

In order to make synthetic speech more expressive, some special parameters should be considered. For example, if we want to synthesize a talk several persons attending, a parameter to indicate the talking one is necessary, for example,
<voice name=”张”>现在几点了</voice>
<voice name=”李”>十点半</voice>
If we want a talk with a piece of background music, a parameter should be used to indicate the source of the music. For example,
<audio src=”D:\Music\歌唱祖国.mp3”, type=”background”/>中华民族终于站起来了
Many other kinds of examples will be taken.

An excellent human computer interface will consist of kinds of multimedia, not just speech. So the markup language should have the ability to describe not only speech but also other forms of multimedia. For example, the combination of visual and audio will make a synthesis system more expressive such as talking head. If a markup language has interface to describe facial animation, we can control not only synthetic speech but also the expression of a face when talking. For instance, if we want to express a feeling of joy, we may set higher speaking rate and other prosody characteristics, and also we can control the facial expression by setting the facial animation parameters. This will perform better than traditional synthesis systems obviously.

3 Multilingual Language Extension

As speech synthesis systems used more and more widely in different areas, multilingual comes to a significant topic today. The markup language should be developed not only for English language but also for non-English languages such as Chinese Mandarin. The non-English languages all have their own characteristics different from others. So the markup language is supposed to take account of the characteristics of other languages. Take Chinese Mandarin for example.

Text structure plays an important role in determining the length of pause time in speech synthesis. For instance, the length of pause between paragraphs is generally longer than that of sentences. In traditional English languages, we have four kinds of text structures, that is, paragraph, sentence, phrase and word. But in Chinese Mandarin, four kinds of text structures are not enough. In Chinese, one word can contain some characters and every character is a syllable, but in English languages, one word is just a syllable. For example,
<s>

```
    <TOKEN token="各位朋友">
        <WORD word="各位">
            <POS>
                <PRO />
            </POS>
            <SYLLABLE syl="各">
                <PHONETIC>ge4</PHONETIC>
            </SYLLABLE>
            <SYLLABLE syl="位">
                <PHONETIC>wei4</PHONETIC>
            </SYLLABLE>
        </WORD>
        <WORD word="朋友">
            …
        </WORD>
    </TOKEN>
</s>
```

Different from English languages, Chinese, a lexical tone language, is said to have an inventory of five lexical tones, including the neutral one. Chinese assigns tonal targets on a lexical as well as phrasal level, while English languages only assign an intonation tune on a phrasal level. It is more complicated than English languages.

Additions and extensions to the SSML should consider such differences mentioned above.

4 Conclusion

At the standpoint of users, the markup language is just an interface to manipulate the speech synthesis. For experts, it is not only an input source but also a tool they can use to debug their systems; for non-experts, although they are lack of the knowledge about synthesis, they can obtain the proper output by marking the raw text and modify the values of attributes of elements. In a word, with the markup language and the parser, the core of speech synthesis can be considered as a black box to the users, they needn’t know the details about functions of each module.

Some of the speech synthesis systems have already contained the XML parser to make use of the SSML or other markup languages based on XML. But most of them are only for English languages, and designed only for speech synthesis. For non-English languages, the recent version of SSML may not be suitable enough to describe all the information to assist the generation of synthetic speech. Under the development of web service and telecommunication, the need of a multimodal interface which is more expressive is also exigent. So the additions and extensions to SSML should try to make it more powerful to satisfy the need of non-English languages and multimedia

Reference

[1] Burnett, D., Walker, M., Hunt, A., Speech Synthesis Markup Language (SSML) Version 1.0, W3C Recommendation, 7 Sep. 2004, http://www.w3.org/TR/speech-synthesis/

[2] Andrew Lampert, 2004, Text-to-Speech Markup Languages

[3] Axelsson, J., Cross, C., Ferrans, J., McCobb, G., Raman, T., Wilson, L., 2004, XHTML+Voice Profile 1.2, VoiceXML Forum Specification, http://www.voicexml.org/specs/multimodal/x+v/12/spec.html

[4] Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., Edington, M., SABLE:A Standard for TTS markup, ICSLP98