Overview of Chinese Speech Synthesis Markup Language

Yan Jun, Yin Bo, Wu xiaoru, Wang Ren-hua, Liu Qingfeng
Anhui USTC iFLYTEK Co., Ltd
University of Science & Tech of China

Since the appearance of KD2000 which was the most important Chinese speech synthesis system in 1998 in Chinese mainland markets, the application scope and integrated business of speech synthesis keep on increasing. The deepening of its application brings about a higher demand for the speech synthesis markup language. iFLYTEK set up relevant enterprise standards in 2001, regulating how to use markup language in TTS system. After the coming up of SSML draft, iFLYTEK keeps on closely tracking and focusing on it and finds out SSML is a well-defined text markup language system which can be used in resource share and module exchange and realizes the fine control over speech synthesis so that the effect of speech synthesis can be improved efficiently. However, at the same time, it is found that SSML markup language doesn't pay enough attention to the technical and application problem in Chinese speech synthesis, and it needs to be extended and improved to fit the situations of Chinese speech synthesis. The necessity to improve SSML lies on the following two aspects: on the one hand, Chinese is a language distinctively different from western languages, and the features of Chinese character and speech ask for some special control elements; on the other hand, the actual business environment has new control demand for Chinese speech synthesis during the process of application of speech synthesis in China.

In the aspect of language, the distinct differences between Chinese and western languages is that the basic grammatical unit of western languages (such as English) is word and words are separated by blank or punctuation, so there does not exist difficulty in separating words when doing grammatical analysis; while the basic grammatical unit of Chinese is character and several characters together constitute a word, so the boundaries between words are not separated by blank or punctuation. The difficulties in Chinese grammatical analysis rest with the variety of words composition, i.e. the same sentence may have several results of separating words, and these results may be grammatical and semantic correct. Therefore, three problems will probably appear during word separation: the different meanings aroused by the meaning duality of natural language, the auto-mechanical word separation, and the size of word-separating dictionary. The probability of the three mentioned errors is little, but once word separation goes wrong, the effect of synthesis will be very much influenced or even destroyed. At the same time, they are unsolvable problems in natural language understanding and auto-word-separation technology at present. Hence we need to enable customers to control the word-separation result through adding word-separation elements

in order to reach a correct and clear meaning and improve the synthesis effect. Another phenomenon in Chinese is polyphone, i.e. the same character may have different pronunciations in different language environment, especially in different words, so the word-separation result will affect the pronunciation of Chinese character and the understanding of sentence meaning directly. Therefore, the application of word-separation element will be advantageous to the analysis and disposal of polyphone and avoidance of pronunciation errors.

In the aspect of speech, the pronunciation of Chinese is also distinctively different from that of western languages. The pronunciation unit of Chinese is character, each of which stands for a syllable. Each syllable has four tones, or no tone to express unstressed syllables. Chinese is a language with tones and different tones are used to distinguish different characters and meanings. As there does not exist the concept of tone in International Phonetic Alphabet, it cannot accurately notate the character pronunciation during the phonetic notation of Chinese characters. The People's Republic of China uses the Scheme for the Chinese Phonetic Alphabet which was officially passed by CPPC in 1958 to spell Chinese. And in 1982 the International Organization for Standardization acknowledged the Scheme for the Chinese Phonetic Alphabet as the international standard to spell Chinese. Therefore, it is necessary to extend the label of phonetic notation when applying SSML to dispose Chinese, in order to support the widely used Scheme for the Chinese Phonetic Alphabet in China.

There exists another problem when disposing words composed of English letters in Chinese speech syntheses system. Because Chinese Roman alphabet orthography and Scheme for the Chinese Phonetic Alphabet are widely used in China, there are a lot of words composed of English letters, but actually they are Chinese characters written in Chinese Pin Yin instead of English words, especially the names for person and place, such as "Jiang Zeming". If these words are pronounced in a way of English word pronunciation, it is not consistent with the pronunciation customs of Chinese people and is difficult to understand. Therefore, a system is needed to decide whether a word of English letters should be read according to the way of English word pronunciation or Chinese Pin Yin.

The actual application puts forward new requirements to SSML system. In telephone speech service system, customers always hope that the synthesized speech can be played together with background music in order to upgrade their experience. When synthesizing a text in actual application, the background music may be added in a given position, or the background sound is switched during the synthesis process. Therefore, a specific background sound element will facilitate the customers to append and control the play of background sound.

Based on the analyses of the above mentioned problems, iFLYTEK extends and modifies the SSML according to Chinese disposal environment on the basis of full

research and understanding of the design idea of SSML, then brings forward the CSSML system, i.e. Chinese Speech Synthesis Markup Language. iFLYTEK always focuses on the latest development of markup language both in China and abroad, and keeps on modifying and improving CSSML according to the demands collected in actual application.

As for the above mentioned problem of Chinese word-separation, iFLYTEK preserves the *word* and *phrase* elements in the early version of SSML. *Word* element is used to define the boundary between Chinese words, which is advantageous to ensuring the accuracy of word-separation, and further ensuring the accuracy of disposing polyphony. *Phrase* element includes several *word* elements and *phrase* elements. It is used to define the boundary between phrases at different levels. Synthesis system can better express the pause and connection between different grammatical levels by using the level information of phrases, so that the synthesis effect can be more cadent, natural and fluent. On the other hand, the combination of elements of *word*, *phrase*, *sentence* and *paragraph* supplies a complete grammatical level information. iFLYTEK applies these elements into the data exchange of text analysis result, which realizes the separation of front-end text analysis and back-end synthesis algorithm.

As for the mentioned scheme for the Chinese Phonetic Alphabet and Pin Yin words, iFLYTEK extends the attribute supported by the *phoneme* element in CSSML. Besides maintaining the support for the *ph* attribute of IPA phonetic alphabet sequence, CSSML also adds the *py* attribute, supporting the Pin Yin sequence in accordance with GF 3006-2001 scheme for the Chinese Phonetic Alphabet. In addition, phoneme element still adds the *lang* attribute, supporting the speech label in accordance with RFC1766 and GF 3006-2001 regulating dialect codes, such as en, zh-cn, zh-hk, zh-tw, and indicating which language to be applied in the pronunciation of the contents in the scope of effect. For instance:

他姓<*phoneme py*=”zeng1”>曾</*phoneme*>

国家主席<*phoneme lang*=”en”>Jiang Zeming</*phoneme*>

As for the background sound, iFLYTEK adds the *environment* element in CSSML, demonstrating the sound field environment in the pronunciation of text in the scope of effect. Through *src* attribute, the background sound file used in synthesis can be appointed. The *repeat* attribute is used to designate whether the background sound needs to be replayed when the synthesized speech is longer than background sound. After the introduction of *environment* element, customers can put any background music at any position of the synthesized text to realize different auditory experiences.

Considering the large amount of Chinese characters and the feature that different dictionaries are applied in different text context, CSSML revises the position restriction of *lexicon* element. In SSML, *lexicon* element can only be put after *speak* element and before other elements. It can only be a single element, and its scope of effect is a whole

text, i.e. a SSML synthesized text can use only one dictionary. While in CSSML, *lexicon* element is allowed to put in any position of the text and include other elements, and is used to designate the dictionary applied in different text domains. In this way, the switch of dictionary in a CSSML synthesized text can be realized. And it also can be realized that different dictionaries can be used in different text domains to achieve better text analysis effect.

iFLYTEK cooperated with University of Science and Technology of China and developed the first Chinese speech synthesis product in Chinese market in 1998. As the deepening of the application of speech synthesis technology in banking, touring and other fields, users complain about the method that the synthesized speech can only be adjusted through adjusting parameters when the speech synthesis engine is integrated. They think this method has too many limitations on application and cannot meet the requirement of adjusting the synthesis effect flexibly in different situations and different types of texts, and hope that there is a more convenient and easier method that can be used to control the speech synthesis effect more agilely. In consistency with the demand in application, iFLYTEK mapped out the enterprise technology standards Distributional Speech Synthesis Language Markup Specification, regulating the markup language used in the speech synthesis system of the dominant speech synthesis product of iFLYTEK --- InterPhonic series. The standards received positive market feedback and further simulated the development of speech synthesis application. In October 2003, CNSC approved the establishment of the national standards---Technical Standards of Chinese Speech Synthesis led by iFLYTEK and definitely required the standards to describe and define the method and type of markup language used in speech synthesis. In October 2004, after one year's repeating discussion, the Technical Standards of Chinese Speech Synthesis to the person-machine speech alternation standard work team of the Ministry of China Information Industry, in which the format of data exchange used in Chinese speech synthesis system is given a special description. CSSML defines a set of markup language system based on XML document structure. The goal of agile control over synthesized speech effect can be reached through simple elements on synthesized text in a distributional or independent system. The standards was handed in to collectivity workgroup in May 2005 after the subject technique group of person-machine speech alternation standard work team carried out repeating opinion solicitations and modifications from November 2004 to April 2005. In August of the same year, it was handed in to the office of standards administration of Ministry of China Information Industry after verified and approved by the workgroup. The standards are supposed to be formal national standards in China at the end of 2005.

iFLYTEK began to support CSSML since InterPhonic CE 2.1 speech synthesis system in 2003, and keeps on improving the support for CSSML standards in the following synthesis system version. In the latest InterPhonic 4.0 speech synthesis

products, most of CSSML features are sufficiently supported. In addition, regarding the problems found in the process of editing CSSML elements by customers, iFLYTEK offers visual CSSML editing tools to facilitate the customers. iFLYTEK receives the approbation of customers from every walk of life in China after offering support for CSSML in its formal products, which are widely used in telecom, banking, insurance, negotiable securities, education and so on. Now, iFLYTEK has considered CSSML as a significant method to offer optimization of speech synthesis effect and customization service, and has gained great achievement.

In 168 and 114 information service in telecom industry, clients frequently use CSSML to optimize the synthesis effect and offer better services. For instance, during Korea-Japan FIFA, 168 sound information services in 296 cities in China use iFLYTEK InterPhonic speech synthesis products to offer telephone inquiry service on sports news and features. It used a lot of CSSML when editing program contents, modified the inaccuracy in word-separation and polyphone pronunciation, and ensured the high quality of inquiry service. Shanghai 114 number inquiry service also applies CSSML in its calling number inquiry services and optimizes the information of synthesized enterprise names and relevant business.

Call Center in negotiable securities also uses a large amount of CSSML to optimize the effect, for example, Guoyuan paper uses CSSML to optimize its information service of 96888, such as today's special prompts, financial and economic information, dynamic stock comment, company introduction, sales department introduction, business introduction and so on.

In Haier's nationwide customer service telephone service also applies iFLYTEK speech synthesis system to report its product number. The pronunciations of number/numerical value and special symbol are regulated by *say-as* element in CSSML, which ensures the accuracy of information report of special format.

CSSML can also be widely used in desktop education products. Gowell is a company developing Mandarin speech education software in Hong Kong. Its product applies iFLYTEK speech synthesis products to synthesize standard Chinese pronunciations. As its product has to synthesis the pronunciation of characters and words without context, to avoid the problem of wrong pronunciation of polyphones, it also uses CSSML to control pronunciations.