TOWARDS SYNTHESIS OF FOCUS IN MANDARIN TEXT-TO-SPEECH SYSTEM

Dezhi Huang Xiaoliang Yuan and Yuan Dong

Speech and Natural Language Processing Laboratory
France Telecom, R&D Beijing Co., Ltd.
{dezhi.huang, xiaoliang.yuan and yuan.dong}@francetelecom.com.cn

ABSTRACT

This paper introduces the significance of synthesis of focus in Mandarin text-to-speech (TTS) system, as well as the key challenges in research on synthesis of focus. The proposal on the extension of Speech Synthesis Markup Language (SSML) is presented for the improvement of intelligibility of key words or phrases, and also demonstrated by an example finally.

I. INTRODUCTION

A focus is defined as the semantic centre of a sentence [1]. In daily communication, focus is generally assigned to be prominent in order to help the listener to understand what is conveyed clearly. For instance, “是李教授得了奖金” (means that Prof. Li wins the prize) emphasizes who wins the prize. It also connotes that Prof. Li is more successful than the other professors. “李教授是得了奖金” emphasizes what Prof. Li wins finally. This method to manifest the focus is called the syntactical method [2]. Additionally, we also represent the focus with the phonetic method. For example, “楼房都震塌 了” means that buildings are destroyed by the earthquake. If the syllable “都” is the stress of utterance, it means that there is no building left in this earthquake; If the word “楼房” is the stress of utterance, it connotes that the magnitude scale of earthquake is very large.

Today TTS technologies have been wide applied in many fields, such as voice-enable information services, multimedia and entertainment, etc. Among these technologies, parametric synthesis and concatenative synthesis are the most prevailing method of TTS. Especially, many commercial systems have employed corpus-based concatenative synthesis which makes it possible to synthesize the speech with high intelligibility and articulation [3]. However, most current corpus-based systems are based on the neutral reading-style utterances. Therefore, the lack of expressivity of focus is a common problem for these systems.

Synthesizing the focus naturally has become an important aspect in current research on speech synthesis. It is helpful for:

- ● Analyzing the syntactic of sentence
- ● Understanding the meaning of utterance
- ● Capturing the turn-taking
- ● Comprehending the attempt and emotion of speaker

Moreover, Synthesis of focus is important to improve the intelligibility of synthesized speech in a noisy environment. In summary, the technique of synthesis of focus can improve the expressivity of TTS system.

II. KEY CHALLENGES IN SYNTHESIS OF FOCUS

There are two key challenges currently exist in the research on synthesis of focus. First of all, nowadays, neither natural language processing nor understanding is syntactically an accurate method to estimate the exact position where a focus appears [4]. As we know, the location of focus is context-sensitive. For instance:

Sentence1: 昨天谁去了长城？(Who went to the Great Wall yesterday?)

Sentence2: 昨天小李去了长城。(Xiao Li went to the Great Wall yesterday.)

When processing Sentence 2 separately, it is difficult to estimate the position of focus only depending on its syntactic feature. If we analyze the above two sentences as a whole, the string "小李" is the focus in Sentence 2. Therefore, the practical application of focus location requires technical evolution of natural language processing and understanding. Second of all, the lack of appropriate acoustic model also holds the progress of synthesis of focus. There are many phonetic manifestations to realize a focus in Mandarin. The relationship between focus and its acoustic features needs to be further explored.

III. STATE OF THE ART

From the viewpoint of linguist, a focus can be categorized into four types: informational focus, contrastive focus, semantic focus and topical focus according to the theory of pragmatics. The manifestation method of focus should match the type of focus. It also means that the method of synthesis of focus needs to be adapted to the type of focus.

In the research on synthesis of focus some progresses have been obtained. Pitrelli and Eide applied ToBI prosodic template to associate the questioning and contrastive emphasis [5]. ToBI labels and text features are used in turn to generate F0 contour and segment duration. Kiriyama, Hirose and Minematsu developed a method to generate the rules of prosodic focus control for a spoken dialogue system [6]. These rules were then revised and validated by listening tests many times. Chen employed the Fujisaki model to analyze and synthesize the focus in Mandarin, and acquired some rules of tone commands affected by focus [7]. Research on phonetic realization of focus in English declarative intonation, has revealed that the pitch range of a narrow focus syllable is expanded, while the pitch range of the postfocus syllable is suppressed [8]. Although some results have been obtained, we are still so far away from high quality synthesis of focus.

IV. THE PROPOSAL FOR SSML

Different from English, a focus in Mandarin is not one-to-one corresponding with an emphasis. Emphasizing a syllable can manifest a focus obviously. But this is not the unique method. In Mandarin, the weakening of adjacent syllables or change of manner of pronouncing syllable also makes the prominence of focus clear.

The emphasis element has been defined to markup the contained text be spoken with emphasis in SSML [9]. The level of emphasis element includes strong, moderate, none, and reduced. It is believed that the emphasis tag is not completely competent for synthesis of focus in Mandarin. We hope that a new element called <focus> could be added in SSML. The following examples demonstrate the usage of <focus>.

EXAMPLE 1:

明天最高气温多少？(How many degrees are tomorrow’s high?)

明天最高<focus type = “contrastive” position = “primary”>30 度</focus>。(Tomorrow’s high will be 20 degrees centigrade)

The string “30” should be an emphasis, while the syllable “度” should be weakened.

EXAMPLE 2:

你买了什么？ (What have you bought?)

我买了<focus type = “informative” status = “primary”>苹果 </focus> 和 <focus type = “informative” status = “secondary”>李子</focus>。(I bought apples and plums.)

The syllable “苹” and “李” are emphases. The emphasis level of syllable “苹” is higher than that of syllable “李”.

REFERENCES

[1] Guessenhoven, C., Focus, mode, and nucleus, Journal of Linguistics, Vol. 19, 377-417, 1983

[2] Xu L. J., Concept of focus and its manifestation in Chinese, Contemporary Research in Modern Chinese, Vol. 3, 1-22, 2001

[3] Chu M., Peng H., Chang E., A concatenative Mandarin TTS system without prosody model and prosody modification, Proceedings of 4th ISCA workshop on speech synthesis, Scotland, 2001

[4] Wang Y. J., Chu M., He L., Location of Sentence Stresses within Disyllabic Words in Mandarin, proc. of the 15th International Congress of Phonetic Sciences, Barcelona, 2003

[5] Pitreli J. F., Eide E. M., Expressive speech synthesis using American English ToBI: questions and contrastive emphasis, Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 694-699, 2003

[6] Kiriyama, S., Hirose, K. , Minematsu, N., Prosodic focus control in reply speech generation for a spoken dialogue system of information retrieval, Proceedings of IEEE Workshop on Speech Synthesis, 139-142, 2002

[7] Chen G. P., etc., Quantitative Analysis and Synthesis of Focus in Mandarin, International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages Beijing, China, March 28-31, 2004

[8] Xu Y., Xu C. X., Phonetic realization of focus in English declarative intonation, Journal of Phonetics, 33, 159-197, 2005

[9] W3C, Speech Synthesis Markup Language (SSML) Version 1.0, http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/