Toward Synthesizing Expressive Mandarin Speech

Hongwu Yang, Shuang Li, and Lianhong Cai
Department of computer science and technology
Tsinghua University, Beijing 100084, China

I. INTRODUCTION

Research efforts in the field of TTS have placed emphasis on the naturalness in synthesized speech to facilitate various applications in Human-Computer Interaction (HCI). The ideal synthetic speech for HCI should not only have proper pronunciations, but also convey the appropriate semantics within the context of use. “Context” refers to the textual context of the document, the identity of the interlocutors in the interactive conversation, the application scenarios, etc. For example, synthetic speech for news reports may adopt lucid and smooth characters while sports commentaries may call for a more animated character.

This paper focuses on expressive text-to-speech synthesis. Expressions in speech encompass many elements. Our work focuses on emotional and stylized synthetic speech in synthesizing speech. Emotion originates from the speakers’ psychological and physical states and is realized through spectral and prosodic parameters. Style is dependent on the semantics of the spoken message and the conversation scenarios so that it can be realized with global prosodic features. Emotion and style are also interdependent. In general, emotion has relatively local effects and its acoustic parameters are more dynamic while style has relatively global effects and its acoustic parameters are more stable in the speech signals. Emotion and style thus jointly modify the acoustic features of the speech signal for more affective and effective conveyance of the underlying message. Thus a TTS system that can simulate different emotions and styles will make HCI more natural and desirable.

II. EXPRESSION OF SPEAKING STYLE

Expression of speaking style in speech is studied in stylistic linguistics, which focuses on the characteristics of linguistic variables within discriminative contexts so that the principles of linguistic selections made by particular individuals and communities can be established. There are many definitions of speaking style. Halliday defines “style” as “the relationship between the

participants of linguistic activity, especially the level of formality (formal or oral) of the expression." Yuanren Zhao viewed style in terms of intonation, classified as Chinese tone, neutral intonation and emotional intonation. Blado et al. defined style in terms of the causal-formal dimension; Abe looked at style in terms of information genres, e.g. literary novel, advertisement and encyclopedia style; Higuchi et al. brought forward hurried, angry and gentle styles against the unmarked style; Gibbon described style as reading style and several kinds of spontaneous styles that have different speaking rates and pitches.

For the purpose of engineering speech synthesis systems that are expressive in style, we need to understand and characterize the *patterns* of speaking styles, the dependencies of such patterns on the *content* and *context* of the input text, as well as the *acoustic correlation* of these patterns in the speech signal.

Speech style is a kind of expression pattern, which is related with the content of speech, the activity of conversation, the emotion of speaking and the emphasized object. With different speech content, we get different expression pattern. For instance, we need to make the synthesized speech smooth and lucid if we want to make the listener understand clearly by stating some facts or principles (broadcasting ordinary news) with TTS; synthesized speech should be of passion if sports games are to be commented; in information broadcasting system in airport, smooth and cordial speech is required. Besides, speaking model also influences expression pattern. E.g. we use statement or authoritative modal to express speaking activities providing information, while requesting modal is adopted when it comes to information acquirer. What's more, there are relations between expression pattern and emotion state- as for one news report, good news is expressed with positive emotion, while bad news is expressed with negative emotion. Expression mode, finally, concerns with the emphasized object. E.g. we always emphasize the new-coming focus when changing the topic or when new information appears.

Speaking styles must accord to the context of the communication. A TTS system not only provides information efficiently, the styles of the TTS system are also need appropriate with the context. Generally, the lax style may be best for the information provide application while the active style may be best for the good news or entertainment application. Furthermore, the synthetic speech must also suit for the different information. For example, new information needs to be emphasized so that the user can focus on them. Hence, the appropriateness is the most

factors in the expressive speech. The style of the synthetic speech must fit the context of communication.

III. THE PROPOSAL FOR SSML

Based on the discussion above, we conclude that current SSML is lack of control on expressive speech. Sentences with the same text should have different expression pattern. According to this, we put forward tags to express modal, which include 4 properties- content type, conversation activity, emotion activity and emphasized object as in Table 1. Content type tags the content of text, which may be “news”, “reading”, “information broadcasting”, “conversation”, etc. Different text types bear different global prosodic features. Conversation activity describes the action of speaking at present. E.g. “information provision” and “information acquisition” have different conversation activity. “Emotion activity” delineates the emotion situation of current content, which can be positive, negative or even neutral. Apart from this, we can also evaluate it in different emotion ranges or with different dimension values. Emphasized object specifies the content that is emphasized presently, such as changing of topic or the new-coming content.

TABLE 1 PROPOSED TAGS

Proposed tag	Status	Numerical value	Descriptive value
Content type	Proposed		News Reading Information broadcasting Conversation ……
Conversation activity	proposed		Confirm Request Give Greet Apology
Emotion activity	Proposed	-1,0,1	Happy Sad Angry Calm Despair ……
Emphasis	exist		Strong Moderate None reduced

An example on expression pattern is as follows:

<pattern content type=’information broadcasting’, dialog activity=’give’, emotion activity=’-1’>Flight 1121 has delayed for 1 hour. Please wait in <emphasis=’strong’> xxx <emphasis/> <pattern/>

REFERENCES

[1] David Crystal, A first dictionary of linguistics and phonetics. Boulder, Colo.: Westview Pr., 1980.

[2] Yuanren Zhao, “Tone and Intonation in Chinese”, Bulletin national Institute of History and Philology Academia Sinica, vol. 4(2), pp121-134, 1933.

[3] Jacques Terken, Variability and Speaking Style in Speech Synthesis, In E. Keller, B. Bailly, A. Monaghan, J. Terken and M. Harkvale. Improvement in Speech Synthesis, pp 199-203, JOHN WELEY & SONS