Pronunciation of Nouns in Text to Speech systems

E.Veera Raghavendra, Lavanya Prahallad
IIIT Hyderabad, India
raghavendra@iiit.net, lavanyap@cmu.edu

At present most speech synthesis systems use raw text as their input which is understandable from a human point of view but problematic for the machines since the process of converting text to speech is very complex; in this paper we discuss the need for having a specific SSML tag for each "mention" (1st occurrence, 2nd occurrence) of a proper noun in the text or paragraph. We discuss that when a proper noun appears first time in the text, then it is spoken more prominently than its second or third or subsequent occurrence. We highlight the need for incorporating a specific tag in SSML to take care of this mention-case. The SSML format is a compromise between human and machine needs. SSML is often embedded in Voice-XML scripts to drive interactive telephony systems. However, it also may be used alone, such as for creating audio books. The advantage that SSML brings is that the designers of such language generation systems need only understand the basic SSML language and do not need specialist speech synthesis knowledge.

Introduction

Speech Synthesis Markup Language (SSML) is an XML-based markup language for speech synthesis applications. SSML directs all Text Analysis steps, providing a standard way to control aspects of speech such as pronunciation, acronym expansion, volume, pitch, rate, range, duration, pause, emphasis, etc., across different synthesis-capable platforms.

The intended use of SSML is to improve the quality of synthesized content. Different markup elements impact different stages of the synthesis process. The markup may be produced either automatically, for instance via XSLT or CSS3 from an XHTML document, or by human authoring. Markup may be present within a complete SSML document or as part of a fragment embedded in another language, although no interactions with other languages are specified as part of SSML itself. Most of the markup included in SSML is suitable for use by the majority of content developers. However, some advanced features like phoneme and prosody (*e.g.*, for speech contour design) may require specialized knowledge."

SSML provides a flexible interface which can be used as a module of any system whose end aim is the synthesis of speech. As it is a completely specified language it is possible to write a program which will transform any other form of input into SSML format. It can be used with any speech synthesizer and will therefore provide a standard interface which will carry out part of the transformation between a text of any format and a speech synthesizer of any design.

What is working today?

SSML would also handle all the problematic processes included in the building of the TTS systems: The pronunciation of a word may be included directly in the text using a pronunciation tag. A style sheet may be attached to SSML which includes information which is specific to a particular speech synthesis application For example: a list of common abbreviations can be included in a style sheet and every time one of these is encountered in the text it will be automatically expanded.

Issues of SSML:

Given that SSML has been chosen to be similar in scope to the annotation systems found in other TTS systems, the main problem is to define a set of annotations such that they are platform independent. It is in trying to solve this problem that we adopt the same strategy used in the field of document processing namely the separation of logical and physical aspects of the input.
Thus the problem of defining SSML comes down to specifying a set of tags which most systems will be able to understand and process and doing so in a manner such that the resultant synthetic speech from different systems is judged to be roughly the same for a given input.

Use MENTION tag in SSML:

We specifically looked at the first and second occurrences (mentions) of the words in an utterance. Our assumption is that first mention of a word in a given conversation is relatively better articulated than their second mentions [1]. We would like to propose or introduce an additional tag <MENTION> that marks proper nouns with the mentions (i.e. first occurrence, second occurrence). Such tag would enable the synthesizers to utter the proper nouns according to their mention – the synthesizers could change the pronunciation according to the mention of the proper nouns.

Reference:

1) Kishore Prahallad, Alan W Black and Ravishankar Mosur, *"Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis"*, in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), France 2006.