Position paper for W3C Workshop on Internationalizing the Speech Synthesis Markup Language (SSML)

Subject: An extension to the `<say-as>` element for diacritics auto-completion.
Date: 23rd of September 2005
Authors:

Przemysław Zdroik przemyslaw.zdroik@telekomunikacja.pl
Krzysztof Majewski krzysztof.majewski@telekomunikacja.pl

Telekomunikacja Polska, Research and Development Centre, Poland

Abstract

The document describes a proposal of adding a new extension to `<say-as>` element. The extension covers Polish language specific issue concerned with omitting diacritics in content of messages sent via Internet or GSM network. The main role of this extension is to ease implementation of automatic message readers.

Table of Contents

1. Motivation

Polish set of characters includes 9 special characters that are not in standard ASCII codepage. Those characters are mostly regular letters with a diacritic: **ą, ć, ę, ń, ó, ś, ź, ż**, **ł**. Because those letters are typed with computer keyboard as special (inconvenient for beginners) key combinations, the diacritics signs are often omitted. The worse situation occurs while writing Short Messages – most of embedded T9 dictionaries or even terminals do not support those characters at all. This results that in applications like e-mails, SMS, Instant Messaging or Internet news, people use "jargon" language, without diacritics. While viewed as text this might be correctly understood by a human, however the same sentence spoken according to Polish pronunciation rules sounds unnaturally and might have a different meaning.

The auto-completion of diacritics can be done at least in the two following ways:

1. Static substitution vocabulary, based on statistical language model.
2. Advanced algorithms based on grammar rules that analyses whole sentences.

While first method is quite simple and can be implemented as extra pronunciation rules, it does not cover many cases, when there are two words that differ just with a diacritic. For example Polish word **paczek** (what means *packages* – plural, Genitive) can be left as is, or corrected to **pączek** (what means *a donut* or *a but* – singular, Nominative). The second method utilizes Polish language specific issues, especially inflections based grammar. In that case, above example could

be handled with rules, like e.g.: “add diacritic to character ‘*a’* if word *paczek* follows a singular pronoun”.

To avoid possible mistakes, the auto-completion feature should be applied only for text, which is supposed to contain words with missing diacritics, like Short Messages contents or Instant Messaging conversations. This is why SSML specification should include a parameter to determine the type of pronunciation.

We are convinced that this case applies also to other Slavic languages, like Czech or Slovak. We do not know specificity of other languages, but we feel that the situation described above might apply as well.

2. Proposed extension

As a solution of described problem, we suggest to introduce new value for the `interpret-as` attribute: `jargon` (value name ‘jargon’ is only exemplary – it could be defined with other word). For this value, we propose the following values for the `format` attribute, which can be used by Synthesis Processor to apply diacritics auto-completion in way, which is most adequate for the particular kind of text:

`format` value	**used for**
no_diacritics	general case (default)
sms	content of Short Message
email	content of e-mail message
im	content of message of an Instant Messenger
news	content of an Internet news post

As an alternative, introduction to the SSML the new tag `<jargon>` with the attribute `type={no_diacritics|sms|email|im|news}` could be considered.

3. Examples of missing diacritics

Expression without diacritics	**UNIPA Transcription**	**Proper Polish writing (with diacritics)**	**UNIPA Transcription**	**English translation**
paczek rozy	`paCek rozi0`	pączek róży (*or* paczek rózy)	`pO%~Cek ruZi0 (or paCek ruZi0)`	a rosebud (*or* rose's packages)
smiac sie	`s`mJac c~e`	smiać się	`c~`mjat+c~ c~E%~`	to laugh
lodz	`lod&z`	łódź	`wud&z~`	a boat
blogosc	`blo`gos`t+s`	błogość	`bwo'goc~t+c~`	bliss