Proposals for Extending the Speech Synthesis Markup Language (SSML) 1.0 from the Point-of-View of Hungarian TTS Developers

Géza Németh, Géza Kiss, Bálint Tóth

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
{nemeth,kgeza,toth.b}@tmit.bme.hu

Abstract

In this paper we give proposals for extending SSML 1.0 so that it can be used to properly mark-up a broader range of languages, including the Hungarian language. The proposal includes extensions for supporting text structure description, text-to-phoneme conversion, setting the speaking style, prosody.

1. Introduction

Following the description of the speech synthesis process given in SSML 1.0 recommendation [1] Section 1.2, we propose extensions to the recommendation at various steps of the process that could contribute to getting better speech output by a synthesis processor for marked-up text in Hungarian and other similar languages (synthetic, agglutinative, using diacritics).

We will refer to some specific properties of the Hungarian language that need special attention in the course of the document where it seems necessary.

2. Describing text structure

2.1. Reasons for allowing more detailed text structuring

The recommendation already contains two elements that help in describing the structure of the document, making it possible to explicitly specify the paragraph and sentence boundaries, instead of letting the synthesis processor attempt to determine it.

However in certain cases further structuring of the text can help the more exact operation of later phases of the synthesis process, e.g. text-to-phoneme conversion and prosody analysis. Although it is possible to give the exact pronunciation and prosody using the `phoneme`, `prosody`, `break` and `emphasis` elements, if it is possible to give higher level information that helps rendering the appropriate speech output, it is advisable for several reasons. First, it is easier for a human editor to specify e.g. the syllable structure or the part-of-speech of a word than to look up and code the IPA phoneme string for the word and to describe the exact pitch contour for the sentence. Second, replacing the synthesis processor may necessitate the rewriting of the phoneme specification (if a processor-specific phonemic/phonetic alphabet is used) and the prosody description (since the prosodic attribute values are only indicative and different processors may render them differently, so the result might not convey the desired meaning).

In sections 3.1, 5.2 and 5.3 we give a few examples for Hungarian where such higher level information is desirable. In section 6.1 we give a proposal for the SSML elements to be used for describing it.

3. Assisting text-to-phoneme conversion

3.1. Using syllable structure

Hungarian is a highly agglutinative language, therefore the text-to-phoneme conversion approach that is successful for English, namely listing most words with their proper pronunciation in a lexicon, cannot be used for Hungarian because the number of word forms is practically infinite; instead, a set of pronunciation inference rules are used that convert Hungarian letter-combinations into the usually used sounds. This conveys the risk that in some word combinations and words (which are themselves often combinations of a base word and different kinds of suffixes in Hungarian) the inference rules fail if the syllable (or morpheme) structure is not known and the word has not been put into some kind of exception dictionary.

E.g. the word “egészség” (“health”) may seem to contain the letter combinations “s” = [ʃ] and “zs” = [ʒ], but in fact it contains “sz” = [s] and “s” = [ʃ]. There are many similar examples with this and other letter combinations, e.g. “halászsas” (“osprey”) and their suffixed forms (e.g. “halászsasokat” – plural, accusative; etc.).

If we can specify the syllable structure of these words in SSML as described in 6.1, the pronunciation inference rules can produce the correct pronunciations without having to describe them explicitly:

```
<w syllables="e-gész-ség"> egészség </w>
```

instead of

```
<phoneme alphabet="ipa"
ph="&#x25B;ge&#x2D0;#x283;#x283;e&#x2D0;g">
egészség </phoneme>.
```

3.2. Using language of included foreign text

The recommendation [1] contains an example in Appendix F where an Italian movie title is included in an American English sentence. The suggested handling of this case is that the same English voice is used for the whole sentence, and either the text is read using the English letter-to-sound rules or the pronunciation is given using the `phoneme` element.

A certain alternative solution seems more desirable when the synthesis processor “knows” the Italian pronunciation inference rules (e.g. a `lexicon` element for Italian is present in the document). In this case it would be sufficient that we indicate the language of the embedded text. We cannot use the `xml:lang` element because that may also change the voice; to

this purpose we introduce a new language attribute into the phoneme element so that it can be used to label an embedded foreign language expression for changing only the letter-to-sound rules to be used.

Using the same example, in this case we can write the simpler form (as described in 6.2)

```
…<speak … xml:lang="en-US">
The title of the movie is:
<phoneme lang="it"> La vita è bella
</phoneme> (Life is beautiful)
```

instead of

```
…<speak … xml:lang="en-US">
The title of the movie is:
<phoneme alphabet="ipa"
ph="&#x2C8;l&#x251;
&#x2C8;vi&#x2D0;&#x27E;&#x259;
&#x2C8;&#x294;e&#x26A;
&#x2C8;b&#x25B;l&#x259;"> La vita è bella
</phoneme> (Life is beautiful).
```

Alternatively, instead of replacing the alphabet and ph attributes, we can add the lang attribute to them, to indicate that we want the letter-to-sound rules for the specified language to be used if they are available, otherwise the pronunciation in the ph attribute is used:

```
<phoneme lang="it" alphabet="ipa" ph="…">
La vita è bella </phoneme>
```

4. Assisting text normalization

Text normalization can be effectively assisted by the use of the say-as element.

The set of constructs that we found appropriate to use in our practice include: date, time (including time intervals like opening hours), number, currency, name, address. Additionally it seems reasonable to include: acronym/abbreviation, web, e-mail, phone, program-code, table, equation.

5. Assisting prosody prediction and description

5.1. Using speaking styles (prosody profiles)

One can easily see that people use quite different kinds of prosody in different situations or when reading different kinds of text. We speak differently when speaking with friends, when giving a talk at a conference, when reading news or reading stories to our children. (For a quantification of some properties for news reading, story telling, novel reading and advertisements in Hungarian see [2].) We can call this important aspect of speech the "speaking style" (or "prosody profile") used with the text in question.

Spelling of names or reading them syllable by syllable can also be considered as a special reading style. This has been used for Hungarian e.g. in a reverse directory application created for a Hungarian mobile company as described in [3].

Modern TTS systems are likely to be able to imitate these styles to some extent, therefore this should be included in the specification, as described in 6.3. The synthesis of expressive and emotional speech is also in the focus of research at present, therefore we propose the inclusion of this aspect of speech into the markup language also.

5.2. Using syllable-level prosody prescription

While word in an analytic language (like English or Chinese) are often quite short and only have one meaning, words in a synthetic and highly agglutinative language (like Hungarian or Korean) are often quite long, made up of several morphemes and have very complex meanings.

Therefore in a contrastive sentence, wherein the speaker wants to call attention to a difference between two statements, the speakers of agglutinative languages sometimes have to stress one morpheme (which is often one syllable) in a word, instead of stressing a separate word, e.g. (here bold type indicates the place of the emphasis):

"Nem a doboz**on**, hanem a doboz**ban** van a könyv."
("The book is not **in** the box, but **on** the box.")

Stress in Hungarian is mostly indicated by a rise in the pitch; in the case of some speakers, this stress may be aided by a short pause before the stressed syllable, a decrease in the speaking rate and/or an increase in the volume on the syllable.

There are also other situations where pitch fall and rise on certain syllables gives the desired meaning. E.g. "Elmentek" means "They are gone" if the pitch is continuously falling, but means "Are they gone?" if the pitch rises exactly at the beginning of the second syllable and then falls down on the third syllable.

For this reason, it should be possible to prescribe emphasis and explicit prosody (pitch contour, speaking rate, durations, breaks, volume) on a per-syllable basis in SSML.

In SSML 1.0 sub-word units cannot be labeled, so emphasizing syllables cannot be described properly, without modifying the text to an incorrect but well-sounding form. (The use of the sub element is not possible either, as the parts of the alias value cannot be labeled.) SSML 1.0 gives a way to prescribe the pitch contour with percentage position values as pivot points, but this should also be extended to other aspects of prosody and syllable positions, as explained in 6.4. This, accompanied by the syllable-structure description in 6.1 can be used to facilitate precise and proper use of prosody for agglutinative languages also.

5.3. Using part-of-speech information

Just as in any other language, there are word forms in Hungarian that have quite different meanings and even part-of-speech category depending on the context. (These are in fact different words that happen to have the same word form.) E.g. the Hungarian word "hogy" can be an interrogative adverb (the question word "how") or a conjunction ("that"):

- "Mondd, hogy vagy?" ("Tell me, how are you?")
- "Igaz, hogy jól vagy?" ("Is it true that you are alright?")

Here "hogy" has quite different emphases depending on its part of speech: it will have strong (or focus) emphasis as an interrogative adverb, and reduced emphasis as a conjunction.

It is not always easy to automatically determine the part-of-speech in a specific sentence, but it can affect emphasis and other aspects of prosody and even pronunciation (although the latter is not characteristic of Hungarian, but it is so in other languages, e.g. English). Therefore a way to indicate part-of-speech when not obvious should be present in SSML as described in 6.1, as a higher level classification of text for reaching the desired pronunciation and prosody, instead of explicitly prescribing these.

6. Proposal for new SSML elements

6.1. A new text structure element

The `w` element represents a word. The `w` element has two optional main attributes: `syllables` and `part-of-speech` (or `POS`).

The `syllables` attribute, if given, contains the syllable structure of the word, namely the syllables connected by hyphens. E.g.:

```
<w syllables="ha-lász-sas"> halászsas </w>
```

It is an error if the value of the syllables attribute and the content of the `w` element are not equal, not regarding the added hyphens ("-"). If the word already contains a hyphen at a certain position, no additional hyphen is to be added in the syllabification.

The `part-of-speech` (or `POS`) attribute, if given, contains the part-of-speech of the given word. This comes from a language-specific list, but probably contains some or all of the following: `adjective`, `adverb`, `cardinal number`, `conjunction`, `determiner`, `interjection`, `noun`, `ordinal number`, `preposition`, `pronoun`, `proper noun`, `verb`.

This attribute can be accompanied by other attributes that further specify the given part-of-speech category. E.g. the Hungarian language is a synthetic one, which means among other things that one word can have many properties that would be expressed by word order or separate word items in an analytic language like Chinese or English. E.g.

```
<w POS="noun" number="plural"
case="accusative"> halászsasokat </w>
```

Obviously these additional attributes can be given only if the `part-of-speech` attribute is given.

The `w` element can be used without any attributes if its use is desirable, e.g. for languages where space is not used to separate words.

6.2. Language attribute for text-to-phoneme conversion

One can use the `lang` attribute with the `phoneme` element to name the letter-to-sound rules to be used on the contained text.

The `lang` attribute can occur alone in the `phoneme` element, or together with the `ph` (and optionally the `alphabet`) attribute. If both `lang` and `ph` are given, and the synthesis processor can create the pronunciation of the content using the language given in `lang`, it will do so; if not, it will use the phonemes given in `ph`.

The `lang` attribute can have the same values as the `xml:lang` attribute. Additionally, we propose the possible use of the `"x-unknown"` value: if the synthesis processor is capable of identifying the language of the text using some LID (language identification) technique, it must process the text as if the identified language was specified in the `lang` attribute; if it is not capable of doing so, nothing is changed. The same use of the `"x-unknown"` value would be desirable with the `xml:lang` attribute, since in certain applications the language of the text to be rendered may be unknown (e.g. embedded foreign language sentences may occur in a longer piece of literature).

6.3. Speaking-style attribute

The `speaking-style` attribute can be used in the same elements where the `xml:lang`, i.e. the `voice`, `speak`, `p` and `s` elements.

Synthesis processors can define their own set of supported speaking-styles. They should support the `"spelling"` value, and may support e.g. `"syllabification"`, `"causal"`, `"news reading"`, `"story telling"`.

The emotional content of the speech may be specified with a separate `emotion` attribute because the two aspects often complement each other. Possible values are `happiness`, `sadness`, `anger`, `surprise`, `disgust`, `fear`.

6.4. Prescribing prosody on a per-syllable basis

For prescribing the exact prosody on a per-syllable basis, the `contour` attribute of the `prosody` element can be used with syllable positions as time position values. The time position can be defined as `syl1`, `syl2`, …, `syl_end` Syllable position `n` means the start of the n^{th} syllable, the special `syl_end` value refers to the end of the expression. Syllable positions greater than the number of syllables are considered equal to the `syl_end` value.

The syllable structure of the content can be given using the `w` element, as described in 6.1, or if not given, it must be determined by the synthesis processor.

7. Conclusions

We presented several real-life examples where the current state of the SSML 1.0 specification may fall short of giving sufficient help to the synthesis processor when dealing with text in the Hungarian language and other agglutinative languages. We proposed four minor extensions to the recommendation so as to enable it to precisely handle these and other cases, namely the word element for describing text structure, a language attribute for the phoneme element, a speaking-style and emotion attribute, and a way of describing prosody on a per-syllable basis.

8. References

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

[2] Olaszy, G., "Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella- és a reklámok felolvasásában" (Characterizing Prosodic Structures in News-, Fairy Tale-, Novel- and Advertisement-Reading, in Hungarian) , *Beszédkutatás 2005*, Budapest, 2005.

[3] Németh, G., Zainkó, Cs., Kiss, G., Fék, M., Olaszy G., Gordos, G.: "Language Processing for Name and Address Reading in Hungarian", *Proc. of IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2003)*, Beijing, China, 2003, pp 238-243.