Position Paper for SSML Workshop In Crete

Roger Tucker & Ksenia Shalonova, Outside Echo Ltd
Dafydd Gibbon, University of Bielefeld
roger,ksenia@outsideecho.com; gibbon@uni-bielefeld.de

28th April 2006

The Authors

Roger & Ksenia run the Local Language Speech Technology Initiative (see www.llsti.org), which is producing TTS in the languages of (mainly) the developing world. As yet, none of the languages have been commercialized.

Prof. Dafydd Gibbon is a partner in the LLSTI project, an expert in African languages, and experienced in building TTS systems.

From our experience with the languages LLSTI has covered so far, there are four areas where we could have some input into the SSML standard.

Agglutinating Languages

Words in agglutinating languages (Turkish, Finish, isiZulu, Ibibio) can be easily broken down into their constituent parts by native speakers, as each of the inflectional prefixes or suffixes in the words carries a separate grammatical meaning. This decomposition is required for proper g2p and tone assignment in Bantu languages, and is therefore essential for producing natural and intelligible TTS.

To allow the precise definition of a morphological breakdown in SSML, we propose a <morph> tag with a “decompose_as” attribute. For instance in Ibibio the word “deppe” which means “not buying” would be specified as:
<morph decompose_as=”dep + pe” > deppe</morph>.

African Tonal Languages (Dafydd Gibbon)

Some orthographies for African tonal languages (e.g. Yoruba) contain tone marks, but others (e.g. Ibibio, isiZulu) do not. In contrast to East Asian languages, there are three steps for tone assignment:

1. lexical tone: tones function as phonemes, as in East Asian languages;
2. morphemic and morphosyntactic: grammatical morphemes may have characteristic tones, and tones may have independent grammatical meanings;
3. terraced vs. discrete level tone: the tones in African languages tend to be level tones (high, low, etc.) but may be combined into contour tones (high-low, low-high, etc.), and sequences of tones are realised in terraces (the pitch difference on the high-low transition is greater than that on the low-high transition), though languages with more than 2 tones may have discrete levels, i.e. withouth the usual downtrend.

For speech synthesis, the lexical and morphemic tones will need to be recovered from the lexicon and the grammar and marked, if the orthography has no tone marking. The

terraced tone relation can in general be recovered and marked automatically from the tone sequence with a finite state model.

To support these in SSML, W3C may well want to support tone markup at all three levels, where:

1. the first level requires simple phonemic tone markup like the proposal for Mandarin (usually only two are needed, **h** and **l**, though up to 4 have been found in a few languages);
2. the second level would be an attribute of the <morph> tag e.g. referring to the previous example (where occasionally the segmental component may be zero):

   <morph decompose_as="dep + pe" tone="h+l"> deppe</morph>
3. the third level requires position-dependent allotones in terraced tone sequences to be specified (particularly important for unit-selection techniques). This could be covered with the existing attributes of the <prosody> tag, but may need an attribute terrace_pos with numerical values ranging over sequential positions 1, ..., *n* in the tonal domain, e.g. a phrase or sentence.

It may also be necessary at levels 1 and 2 to mark *floating tones*, i.e. tones which have a grammatical meaning but which attach to whatever the nearest syllable is, perhaps with an attribute "floating" with boolean values "yes", "no".

Specifying dialect and speaking style

It is not straightforward to define what is normative speech for a language with various dialects (where no one dialect is considered to be the normative pronunciation). The standard pronunciation might be determined in several ways:

- The speech of broadcast readers of central TV/Radio stations can be considered as standard.
- Socio-linguistic study can be carried out. This type of research requires a lot of effort – plenty of recordings and their analysis. Nevertheless, it is the most reliable method as it allows verifying changes in speech culture and thus defining the normative speech (pronunciation standard typically changes significantly over a 20-30 year period).
- "Compulsory" appointment – the speech of a particular person (professor, writer, actor…) can be defined as standard.

In our experience, these steps haven't taken place and in fact, people in the same city will have different ideas on what "Normative" speech is. In this case, the ability to specify a dialect in SSML would be very important (though that assumes that a TTS system in that dialect exists). For instance you might have some generic values:

<voice region="North">

but also some specific values:

<voice region="Newcastle">

It is also interesting to notice that for European languages the speaker should normally have a loud and distinctive voice, whereas in Ibibio culture, for example, it is very insulting to speak loudly, so the synthesis should be able to support a gentle voice with the corresponding voice quality. This would be an expansion of the voice tag:

<voice type=”gentle”>
But what should the values of the type attribute be?

Say as

Many languages have exceptions which require the use of the “say as” tag, but are these exceptions always easily encodable using other words from the language?

For instance, Hindi has got almost direct g2p rules except for the phenomenon of schwa deletion. In the LLSTI Hindi system (from HP Labs India), this was partly solved by using extended g2p rules, but to completely predict schwa deletion, morphological decomposition would be required, and that is not yet fully available.

For a user to specify schwa deletion without resorting to phonemic transcription, it may not always be possible to use standard Say-as functionality (except to resort to the phoneme level description which requires expert knowledge), as there is no markup for the schwa, or indeed for the absence of schwa. Is there a case for extending the alphabet to allow this?

This assumes of course that the W3C has already a standard for non-Roman scripts like Devanagari.