Adaptable speech-based interfaces

Pavel Žikovský, Pavel Slavík

Department of Computer Science and Engineering
Czech Technical University in Prague
Czech republic

Email: xzikovsk@sun.felk.cvut.cz, slavik@cs.felk.cvut.cz

Abstract. As there are many voice systems, which can stand for user interface for visually impaired people, there is still a problem with explaining more complicated structures such as trees, etc. through speech. The output of such an interface is then crippled into a text reader. The problem is, that we have lost the (graphical) information about the structure. This paper will describe a solution to this problem, based on using more than one voice color to represent the described structure and position within it.

1. READING STRUCTURED INFORMATION

Typical examples, in which the reading can be problematic, are mathematical and logical expressions. As the complexity of the expression increases, its vocal form becomes quite hard to understand. This is not a problem of the speech synthesis, it's simply the fact that an utterance like „x times left bracket b plus left bracket c minus sinus left bracket d times left bracket … „ is quite complicated and hard to follow. From this example we can generalize the classes of information, which are hard to say: These are informations, which tend to have a clear structure, such as programming languages, expressions, or more clearly any information, which is hierarchically ordered. It's obvious, that hypertext documents also belong to this class.

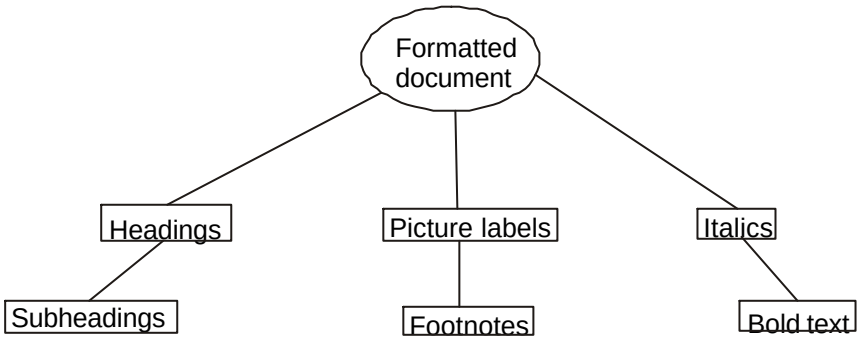

Figure 1 - Example of well-formatted document structure. Each branch has a different voice; voice pitch increases with node depth

Because trees can best represent those structures, we can isolate a subgroup of structure nodes, which are at the same level in tree structure or lie on the same branch. If we assign each such subgroup a different human voice, the intelligibility of the spoken form increases. As an example think about a system, which reads mathematical expressions using the following rules: The content of the brackets is read by a different voice for each bracket pair. Also, function names are being read using another voice than common bracket voices. There is another important problem: What to do if

brackets are nested? Due to the impossibility of mixing several voice colors together, the resolution is one of those from a systematic kind - lets give each level some priority.

As already said, similar structures as in the fore mentioned mathematical expression can be found all over the computer-human interaction field. A well-structured web page with headings, subheadings, italics and bold text is a obvious example of such a structure, as is a common structured text.

Some more applications:

- reading programming language source (different libraries – different voices, begin-end, brackets, etc.)
- reading mathematical (logical) expression – each level of brackets is being represented by one voice, also the same with operations priority (operands, which belong to one operator are read by one voice color)
- providing multilevel information about a system (car, street/rail traffic, etc.)
- changing voice timbre in speech corpus for concatenative speech synthesis
- personalizing any speech output
- assigning each avatar (inhabitant of virtual world) a different voice in virtual worlds

As the project is in the stage of laboratory development, we have performed some tests on intelligibility of multicolored utterance. We asked 3 people to listen to 10 synthesized texts, in which each part of one text was said by a different voice and there were different numbers of voice colors in these texts. Listeners didn't know how many voice colors would be used and they were asked to put a mark in the text, at the spot where they think the voice color has changed. If at least 95% of changes were recognized by listener, we assumed this change was recognizable. The table presents the maximum number of perceived voice colors for each listener. It's clear from the results that we have a large possibility of varying voices within an utterance.

Listener No.	**1**	**2**	**3**
Voices perceived	14	12	15

REFERENCES

[Pierce 92] J. Pierce, *Human-Computer Interaction*. Addison-Wesley, 1992.

[KaMa 99] A. Kain, M.W. Macon, *Spectral voice conversions for text-to-speech synthesis*. Oregon Institute of technology press, 1999.

[ZiSl 98] P. Žikovský, P. Slavík, *Fast Fourier Transformation in Voice Synthesis*, TSD'98 conference paper.

[AnonTut] Author Unknown, *Tutorials on Speech Technology*, http://murray.newcastle.edu.au/users/staff/speech/home_pages/tutorial.html