Using meta-text to improve intelligibility of speech-synthesized e-texts

Date: *July 2011*

Student: *B. Versteegh*
Utrecht University
3248909

Supervisors: *G. Bloothooft*
Utrecht University

D. Binnenpoorte
Het ConsultancyHuis

Abstract

Current Text-to-Speech software such as *Vocalizer* is able to produce fairly natural speech from texts that do not contain meta-text. Meta-text that is part of most modern electronic text-formats is ignored by *Vocalizer*, resulting in unnatural output, or loss of structural information. The present research designs and tests a method to preserve meta-text information in Text-to-Speech conversion. Preservation was done by mapping various structural elements in the e-text to speech, non-speech audio and pauses. A listening experiment, using 23 participants, was performed to measure this method's effectiveness in improving three aspects: listening comfort; perceived speech intelligibility and perceived synthesis quality. In the case of list-structures, significant improvements between 18% and 30% were measured in all three aspects. Omission of a large data-table resulted in significant improvements between 21% and 61% in all three aspects as well. Mappings for headings, images, page-breaks did not result in significant improvements.

Introduction

Progress in the development of Text-to-Speech (***TTS***) technology, over the past decade, has opened new opportunities for practical applications. For many years, synthesized speech was hard to follow, unnatural and unpleasant to hear, causing frustration for its users. Commercial *TTS-engine*s such as *Vocalizer* can now produce an intelligible, pleasant and natural sounding voice that allows for exposure to synthesized speech of longer texts, such as text-documents.

Modern electronic texts (***e-texts***) do not only contain natural language, but also text meta-data (***meta-text***) that defines the structure and mode of presentation of the text (*Xydas and Kouroupetroglou,* 2001) . Meta-text causes a piece of text to be underlined, or to be shown as a *header*, for example. HTML and MS-Word documents are examples of popular e-text formats that contain meta-text. In some formats, the meta-text is part of the document, but hidden to the user (MS-Word), and in some cases, authors can directly insert the meta-text (HTML). E-texts without meta-text are called ***plaint-text*** files. Any text in a plaint-text file is displayed verbatim, with the exception of automatic word wrapping.

Today, most e-texts contain meta-text of some kind, but some *TTS-engine*s such as *Vocalizer*, ignore the meta-text in e-texts. This leads to two kinds of problems. First, structural and visual information about the text is lost, so the listener will have an inferior understanding of the contents, compared to the original document. Secondly, certain meta-text acts as a boundary marker for phrases. If such meta-text is lost, *TTS-engine*s cannot determine phrase boundaries, and will string multiple phrases together. These two issues reduce intelligibility of the document, so it's clear that meta-text should not be ignored. *Xydas and Kouroupetroglou (*2001*)* have developed a system for converting e-text to Speech and Audio (***e-TSA composer***) that maps meta-text to speech or non-speech audio. Their work does not investigate to what extend the e-TSA composer improves intelligibility, or which meta-texts reduce intelligibility the most, when ignored.

The present research aims to measure the effectiveness of a document preprocessing system that maps meta-text to speech and audio, with respect to enhancing intelligibility of the audio output. Our hypothesis is that such a system will indeed improve intelligibility.

Pilot Experiment

Introduction

The purpose of the first experiment is to discover and categorize the variety of current errors that occur when meta-text containing documents are converted to speech by a modern *TTS-engine*. The results of this experiment will determine what kind of meta-text is problematic, and which meta-text will be considered by our preprocessor.

Method

The *TTS-software* we used for this pilot experiment was Nuance *RealSpeak*, as we didn't have Nuance *Vocalizer* available initially. Though *Vocalizer* is a newer version of *RealSpeak*, the two are still very similar products, and we knew that there would not be big differences in meta-text processing between the two. As test material, six e-texts were selected to represent the variety of e-texts used today, in terms of form and content. Here is a list and short description of the e-texts that were chosen for this experiment.

Table 1: Overview of e-texts used for the Pilot Experiment

Kind of document	Filetype	Characteristics
Meeting agenda	Doc	No meta-text, whitespace and dashes for formatting
Meeting notes	Doc	Little meta-text, only bold text for headings.
Bachelor thesis in Medicine	PDF	Running text, page headers and footers.
Material for Logic course	PDF	Formulas, figures, running text.
Material for Philosophy course	HTML	Running text, some images, lists.
Law text	HTML	Many deeply nested lists, list-items are long phrases.

Each of these e-texts were converted to speech by copy-pasting the text (including meta-text) into Nuance *RealSpeak*. The text was copied using the default software for such files, on a Windows computer. The experimental procedure was to listen to the speech produced by *RealSpeak*, while taking notes of anomalies, especially those that were clearly meta-text related. The experiment was performed by two internship supervisors and myself.[1]

Results

Errors related to meta-text comprised about one third of our findings. The others were linguistic problems that were either mispronunciations at the word-level, inappropriate prosody at the phrase-level, or misinterpretation of symbols and numbers at the semantic level. Although our focus was meta-text related issues, the most apparent class of problems was still linguistic or semantic.

1 Although not acceptable for a regular experiment, this experiment serves only as a Pilot to determine our plan for the actual research. Results of high precision or reliability are not a requirement.

Table 2: Examples of linguistic errors

Text input	Speech output	Notes
filosofie	<fil'osofie>	Stress on wrong syllable
de term mens	<de termens>	Assimilation across word boundaries
Gottlob Frege	/ɣɔtlɔp ɾreɪːɣə/	Mispronunciation of foreign words

Numbers were often not read in a useful way. The reading style depends much on the context, and the Speech Engine did not take this into account.

Table 3: Number errors due to different contexts

Text input	Speech output	Notes
(John Doe, 1996).	<john doe .. one thousand nine hundred and ninety six>	The year is read as an ordinal number.
[1-10] [1:10] [1.10]	<one .. ten>	Numbers are a range; ratio or chapter-index, but are read as sequences.

The errors that were related to meta-text, were grouped by type of meta-text:

Table 4: Anomalies in Synthesized Speech, grouped by meta-text type

Meta-text type	Number of errors
Lists	8
Headings	7
Captions	3
Images	3
Headers and Footers	3
Tables	3

Although *lists* were mentioned most often in the notes, they are also a common structure. In most cases *lists* were recognizable, because of appropriate intonation and pauses in the speech. However, in several cases the items on the list were read as a single phrase, without any pauses.

Table 5: Example of failed list rendering

Text input	Speech output
• first item • another item • final item	<first item another item final item >

Most of the *heading*-errors were caused by *chapters*' index numbers. Commonly, a *heading* is written with a numeric prefix, followed by the chapter's title. *RealSpeak* did not pause after the numbers, causing confusion. Roman numerals and *chapter-titles* with sub-indexes (using periods) were not rendered intelligibly. In other cases, there was no pause between the heading and the following or preceding sentence. Sometimes it was not clear that a particular phrase was a *heading*, rather than part of the running text.

Table 6: Examples of failed heading rendering

Text input	Speech output
1 Introduction	<one introduction >
5.1.2 Conclusion	<five .. one .. two conclusion >
IV Appendix	<i v appendix >
6. Discussion Our results indicate [...]	<six .. discussion our results indicate >

In places where the document contained a *graphic* with a caption, the *caption* was read as part of the running text. As listeners are not aware of the existence of the related graphic, captions caused some confusion.

The *tables* in the sample documents were read as a one dimensional string of words. The two-dimensional structure that makes tables so useful, was not conveyed by the speech output. Phrase or word boundaries were sometimes lost and content from different cells or even rows were read without any pauses.

One of the texts contained a *header* and *footer* text on every page. To hear them on every page, as if they were part of the running text, was distracting and unpleasant.

Discussion

From our results, we have found several specific meta-text elements that - if ignored by the speech engine - will reduce the quality of the generated speech. For each of these elements: *headings, lists, tables, figures, headers and footers*, we will design and implement a method that maps the meta-text to speech and audio.

Optimization

Meta-text

In the Pilot Experiment, we have discovered various kinds of meta-data that should somehow be converted to speech and audio, so that their structural and functional information will not be lost. How to best represent this information using audio is not a trivial thing. For some elements, such as complex tables or figures, there may not exist any auditory representation that can retain all information. Difficulties in converting e-texts to audio have several causes.

- The visual and auditory domain are very different.
- The meaning of meta-text is sometimes ambiguous.
- Limitations of the *TTS-engine*.

The parameters that could be used to express an idea *visually*, are very different from the parameters in the *auditory* domain. Conversion between the domains is difficult, because there is no intuitive way to map each visual parameter to an *auditory* one. If a word is marked **bold**, for example, how could we modify the spoken word so that it represents boldness? Using emphasis on those words seems appropriate, but what about phrases that have larger fonts; are in italics; are colored or underlined? Emphasis seems appropriate for all of these decorations, but mapping all of them to emphasis would make them indistinguishable. Perhaps different kinds of emphases could be used, but this could not be intuitively understandable to a listener. Mapping each to another *auditory* parameter will make it less intuitive as well. Besides the above visual variations of text, graphics such as lines and boxes can arguably not be intuitively converted to audio at all. Conversion of visual data to auditory data will either inherently result in information loss or in an visual-auditory mapping that is not intuitive, and therefor not understandable without training.

Table 7: Different parameters of the Visual and Auditory domain

Visual domain (print)		Auditory domain
• text size • text boldness • underlining • italics • color • lines • spacing • symbols • images • tables	*How to map* **? ⇒**	• volume • voice • pitch • intonation / prosody • speed • pauses • non-speech sounds • whispering • effects (e.g. echo)

Ambiguity in meta-text makes it difficult to choose proper speech and audio output for a phrase. Ambiguity arises from the fact that there are different kinds of meta-text (*Coombs*, 1987) .

- Presentational
- Descriptive
- Procedural

Presentational meta-text changes the appearance or layout of a piece of text. Common examples are **boldness** or *italics*. ***Descriptive***, or ***semantic*** meta-text is used to specify the function or semantic category of a piece of text. The presentation of elements with descriptive meta-text is then defined on another level. For example, a phrase is marked to be a heading. The software that displays the text could then make the text somewhat larger, and perhaps in bold. ***Procedural meta-text***, used to define macro's and variables in a more programmatic style, is not relevant so will not be discussed.

When we convert e-texts to speech and audio, the structure and function of the various parts of the document are important. Using the descriptive meta-text, a converter can recognize which elements are *headings*, and adjust speech or audio-output accordingly. However, when only presentational information is present, such as that a phrase is *bold and large*, the phrase *might* be a heading, but we cannot be sure. The meaning or function of presentational meta-text depends on conventions and interpretation. It is therefor not easy to define rules that can convert them to intuitive auditory equivalents.

Despite the difficulty of designing adequate conversion rules, we eventually could still come up with a set of rules. A final limitation is posed by the *TTS-software*, because it may not expose control over all aspects. Consider some example rules from Table 8; the *TTS-software* may not allow insertion of non-speech audio, or does not allow the user to customize intonation.

Table 8: Examples of mappings from meta-text to speech and audio

Element	Speech or Audio output
Tables	• Not read at all; or • Read only the first row; or • Read only the column titles; or • Read all data sequentially, with intonation
Lists	• Rising pitch near the end of each item, and falling at the last • Short pauses between the items.
Headings	• Pauses before and after heading • Read chapter number
Figures	• Do not process • Say the word “figure ” • Play a sound
Bold text	• Change intonation for emphasis

In *Vocalizer*, the *TTS-software* that was used for the present research, intonation can only be controlled after switching to phonetic input. This means that the input bypasses *Vocalizer*'s internal lexicon and the grapheme-to-phoneme conversion must be performed before input is sent to the *TTS-software*. This is a very complicated process, that almost entails building another *TTS-engine*. So for *Vocalizer*, solutions that map meta-text to intonation patterns are not feasible.

Conversion Rules

As a consequence of the above outlined limitations, application of conversion rules is restricted to unambiguous, descriptive meta-text for this research. The limited control of speech parameters that *Vocalizer* offers, limits us to use only pauses, spoken text and non-speech audio to preserve the structural meta-text information. Within this range, we have implemented the following set of rules.

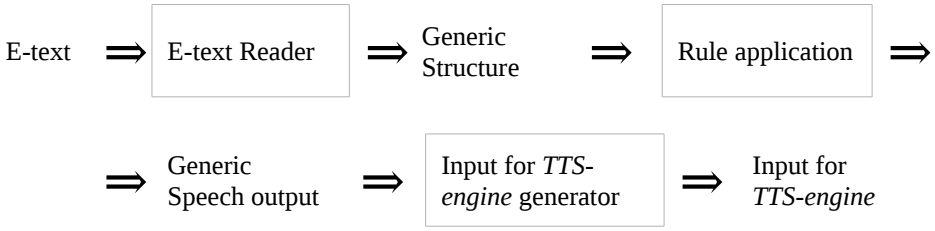
Table 9: Implementation of mappings from meta-text to speech and audio

Element	Speech or Audio output
Tables	Play a sound icon, then: <This is a table of y rows by x columns>
Lists	Play a sound icon before each list item.
Headings	Play a sound icon, read heading, pause briefly.
Figures	Play pencil sketch sound.
Headers and Footers	Do not read.
Page Breaks	Play sound of a page flipping.

The sound icons are single; chorded or consecutive glockenspiel notes. For the *table*, it would be preferable to somehow preserve the contents, as well as its structural information. Research shows it is possible to linearly present the contents, while using intonation to convey the table structure (*Spiliotopoulos et al*, 2005). Since phrase prosody cannot be controlled via *Vocalizer* meta-text, this method is unfortunately not possible. Without proper intonation, we presumably wouldn't be able to present tabular data intelligibly and pleasantly via speech or audio. Our last resort is therefor to omit the table's contents entirely, while informing the listener of its presence by means of spoken text and a sound icon.

The rules were implemented using the Python Programming language. The architecture is a pipe-line consisting of three components. The first component reads an e-text file and recognizes certain meta-text. For HTML and DocX an XML-based module was created. The e-text reader component abstracts different implementations of meta-text, such as *lists*, *tables* or *headings*. Then the second component contains the optimization rules. It receives the abstract components or fragments thereof as input, and then selects appropriate abstract speech output, such as "pause for 400ms". The abstract outputs are then translated to text and meta-text in a format that can be understood by *Vocalizer* (for example /?pause=400/). This way, a different output component could be written to serve a different *TTS-engine*.

Table 10: E-text preprocessor



Verification Experiment

Introduction

Based on the Pilot Experiment, we have built a text-preprocessor that contains rules for converting meta-text to speech and audio. We performed a verification experiment to determine whether this conversion step has lead to improved speech output; speech intelligibility in particular. The preprocessor inserts pauses at several structural boundaries, and auditory icons before structures such as headings and list-items. We expect that this leads to better intelligibility of at least those structures. The first experiment also revealed that inappropriate speech output and lack of pauses can cause some frustration or distress. We therefor also expect that listeners will better enjoy listening to the optimized speech output compared to the unmodified output. To help our interpretation of the results of this experiment, we will also ask participants to decide on the usefulness of the sound icons they will have heard in the samples (selecting from five options, ranging "*very useful*"–"*annoying*"), after the experiment is over.

Method

For this experiment, 23 participants listened to both an ***optimized*** and an ***unmodified*** version of six different speech samples, generated from six corresponding source e-texts. Each source e-texts contained one specific meta-text that we have optimized:

23 participants × **6 source e-texts**
- Unordered List
- Numbered List
- Chapter Heading
- Figure
- Table
- Page Break

× **2 Versions**
- Unmodified
- Optimized

The participants were instructed to rate each sample on three subjective qualities, using a five-point bipolar scale (*bad*, -, *neutral*, -, *good*). ***Speech Intelligibility*** was explained as how well the participant is able to hear the words and understand the meaning of what was said. For ***Synthesis Quality***, participants were instructed to judge their perception of the quality of speech. This consists among others of appropriate intonation, proper pronunciation of words and sentence prosody; and how natural the speech sounds. ***Listening Comfort*** is what the participants could use to express frustration or joy they felt while hearing the samples. If something bothered them, they could give a negative rating on Listening Comfort.

Table 11: Rating of three subjective qualities using a bipolar scale

Category	Bad		0		Good
Speech Intelligibility	◯	◯	◯	◯	◯
Synthesis Quality	◯	◯	◯	◯	◯
Listening Comfort	◯	◯	◯	◯	◯

Participants were divided in groups A and B, to counter order-effects: group A starting at sample 1, and B starting at 7, continuing with 1 after sample 12. Half of the Unmodified-Optimized pairs were presented in that order; the others vise-versa. Pair members were at least three apart.

Table 12: Listing of twelve samples and their order of presentation for participant-groups A and B

Index		Sample	Version	
A	B		Unmodified	Optimized
1	7	Chapter Heading	-	
2	8	Page Break		+
3	9	Figure		+
4	10	Unordered List	-	
5	11	Numbered List	-	
6	12	Table		+
7	1	Figure	-	
8	2	Chapter Heading		+
9	3	Page Break	-	
10	4	Numbered List		+
11	5	Table	-	
12	6	Unordered List		+

The participants were also informed that the audio samples may contain non-speech sounds, and that the sound of a pencil scratching means that the source document contained an image at that position.

Results

For each (unmodified, optimized) sample pair, a paired Student's t-test was performed, to measure whether there is a significant difference between the pairs in both groups. In Table 13, each row contains the mean *difference* in score between the optimized and the unmodified version (positive numbers suggesting improvement), for that category; the *standard deviation* of the score, the *interval* that has a 95% probability to contain the population's average, the *t-value* that is used to determine *P*, the significance value and the normalized *improvement* of score. P values below 0.01 indicate significant difference between the two samples.

Table 13: Results of verification experiment (22 degrees of freedom)

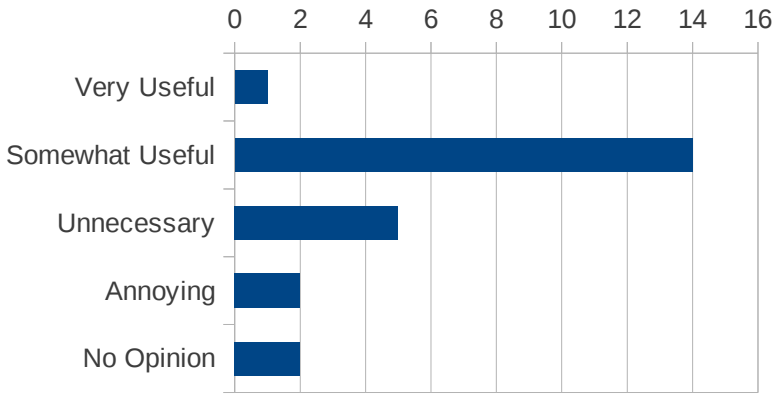
		Paired Differences						
Variables	**Category**	**Mean**	**Std. Dev.**	**95% Confidence Interval of the Difference**		**t**	**P (Sig. 2−tailed)**	**Improvement** *
				Lower	Upper			
Heading	Δ Intelligibility	-0.087	1.041	-0.537	0.363	-0.401	0.692	-2.18%
	Δ Comfort	0.043	1.522	-0.615	0.702	0.137	0.892	1.08%
	Δ Quality	-0.130	1.517	-0.786	0.525	-0.412	0.684	-3.25%
Figure	Δ Intelligibility	-0.043	0.706	-0.349	0.262	-0.295	0.770	-1.08%
	Δ Comfort	0.261	1.054	-0.195	0.717	1.187	0.248	6.53%
	Δ Quality	0.261	1.287	-0.296	0.817	0.972	0.342	6.53%
Unordered List	Δ Intelligibility	1.130	1.359	0.543	1.718	3.990	**0.001**	28.25%
	Δ Comfort	1.217	1.043	0.767	1.668	5.600	**0.000**	30.43%
	Δ Quality	1.130	1.014	0.692	1.569	5.348	**0.000**	28.25%
Numbered List	Δ Intelligibility	0.739	1.322	0.168	1.311	2.682	**0.014**	18.48%
	Δ Comfort	1.043	1.522	0.385	1.702	3.288	**0.003**	26.08%
	Δ Quality	1.130	1.359	0.543	1.718	3.990	**0.001**	28.25%
Page Break	Δ Intelligibility	0.087	1.041	-0.363	0.537	0.401	0.692	2.18%
	Δ Comfort	0.348	0.982	-0.077	0.772	1.699	0.103	8.70%
	Δ Quality	0.391	1.406	-0.217	0.999	1.335	0.196	9.78%
Table	Δ Intelligibility	2.435	1.080	1.968	2.902	10.814	**0.000**	60.88%
	Δ Comfort	1.565	1.308	0.999	2.131	5.738	**0.000**	39.13%
	Δ Quality	0.870	1.140	0.376	1.363	3.657	**0.001**	21.75%

* Scores were converted to a scale ranging [-2,2]. The improvement is the percentage of the maximum difference on this scale [4].

The e-texts containing an *unordered list*; a *numbered list* and a *table* respectively had a significant improvement ($p<=0.01$) in all three measured aspects: perceived intelligibility, synthesis quality and listening comfort. Improvement ranges between 18% to 30% for the *lists*, and intelligibility of the e-text containing a *table* was improved by 61%, while comfort increased by 39% and synthesis quality by 22%. User experience for the e-texts containing the *header*, a *figure* and a *page-break* elements respectively, was not significantly enhanced by the addition of sounds and pauses where these elements occurred.

Although the participants were not told about the function of the sound icons, except for the pencil scratch, the majority of the participants found the sounds at least somewhat useful (14 of 25[2]). Five participants found the sounds unnecessary, while two found them annoying.

Illustration 1: Usefulness of sound icons, according to participants



Discussion

The optimization rule for *tables* was to skip the table data, and to tell the user how many of rows and columns the table contained. User experience for the optimized version was significantly greater, but of course all information from the table was lost. The balance between presenting all information and creating a better user experience should depend on the application of the *TTS-technology*. Using different *TTS-software* that allows controlling prosody, using optimized intonation patterns, allows for retaining table contents, while simultaneously conveying the tabular structure.

2 Data for two participants was discarded for the listening part of the experiment, due to technical issues. However, their answers to the questionnaire were kept.

The improvement seen in e-texts containing *lists* probably had its biggest contribution from the preservation of the list-item boundaries. The sound icons might have been a contributing factor too. The experiment of the present research was not set up to determine the influence of different factors in the possible improvements. Other research should make clear whether audio icons; pauses or a combination of both is the best way to make list items more distinguishable and intelligible.

The page-break; heading and image meta-text did not see any improvement. There are several explanations for this. One is that the omission of this type of (structural) information did not cause any confusion, so the added sounds did not solve any actual problem. Another explanation is that there was in fact confusion in the unmodified version, but the provided solution in the form of audio icons was not adequate. Inadequacy could be because of insufficient instruction to the participants regarding the function of the sound icons, bad choice of sounds, or because sound icons cannot compensate for the missing information at all. In hindsight, we think page-breaks usually do not define document structure, except for a few cases, but then these page-breaks coincide with a transition to a new paragraph.

Interestingly, although the actual synthesis performed by *Vocalizer* was not directly modified, participants have reported higher synthesis quality for those structures where other improvements were measured. Surely, the insertion of phrase boundaries for list-items has influenced the prosody, so an improved prosody would then be regarded as better synthesis quality as well. This shows that the perceived quality of synthesized can be improved by detecting phrase boundaries, using meta-text.

Finally among those structures where significant improvement in intelligibility was measured, listening comfort had also increased. This is particularly important to those that wish to develop consumer products that use *TTS-technology*. Listening comfort, which will ultimately lead to consumer satisfaction, goes together with increased intelligibility. whether the increased intelligibility itself caused a better listening comfort, or vise-versa is a topic of further research.

In general, it should be clear that this was an experiment of limited scope. One requirement was to keep the experiment under 15 minutes, leaving insufficient time to present multiple sample pairs per meta-text. The results for one type of meta-text do not apply to all possible instances and uses of that meta-text. Fortunately, two of our significant findings were related to the list-structure, making our finding a bit more relevant. There are however countless of other types of lists (different length of item text, different levels of depth, mixed numbered and unnumbered, different semantic content), and to say that our optimizations will work well for most or all of them, would be naive.

Conclusion

Our results are based on an experiment with a small scope, so should not be taken to apply to e-text in general. However, we have found two instances where mapping meta-text to speech-and-audio leads to better perceived *intelligibility*, *listening comfort* and perceived *synthesis quality*. These instances are *numbered* and *unnumbered list-structures*. Using our method, improvements of least 18% were measured in all three aspects. Omitting *tables* from the running text also improves these three aspects, by 61%, 39% and 20% respectively. These improvements were achieved by the addition of sound icons, insertion of pauses, and for the table; a spoken description of the size of the table. Participants of the experiment generally (14 of 25) thought the sound icons to be useful.

The meta-text that represents *page-breaks*, *figures* and *headings* did not see any significant improvement by accompanying them with extra pauses and sound icons. The reasons for it are not clear, and could be due to the specific text-samples used, the length of the pauses, and whether the relationship between the sound icon and the represented structure was clear. More research in this area is desired.

Our hypothesis was that our previously described method would lead to improved intelligibility of synthesized speech. We partially accept our hypothesis for the case of list-structures. But for structural meta-text in general, our hypothesis should not be accepted based on the present research. There is possibly certain structural meta-text that would not benefit from any particular auditory mapping.

References

Coombs, J.H (1987), "*Markup systems and the future of scholarly text processing"*, in *CACM* Nov. 1987

Xydas, G. & Kouroupetroglou, G. (2001), *"Augmented Auditory Representation of e-Texts for Text-to-Speech Systems"*, University of Athens

Spiliotopoulos et al. (2005), *"Experimentation on Spoken Format of Tables in Auditory Interfaces"*, University of Thessaly