Explicative Document Reading Controlled by Non-speech Audio Gestures

Adam J. Sporka[1], Pavel Žikovský[2], and Pavel Slavík[1]

[1] Czech Technical University in Prague, Faculty of Electrical Engineering,
Department of Computer Science and Engineering,
Karlovo náměstí 13, 121 35 Praha 2, Czech Republic
{sporkaa, slavik}@fel.cvut.cz
[2] Musical Acoustics Research Centre, Music Faculty,
Academy of Performing Arts in Prague,
Malostranské nám. 13, 11800 Praha 1, Czech Republic
pavel.zikovsky@hamu.cz

Abstract. There are many situations in which listening to a text produced by a text-to-speech system is easier or safer than reading, for example when driving a car. Technical documents, such as conference articles, manuals etc., usually are comprised of relatively plain and unequivocal sentences. These documents usually contain words and terms unknown to the listener because they are full of domain specific terminology. In this paper, we propose a system that allows the users to interrupt the reading upon hearing an unknown or confusing term by a non-speech acoustic gesture (e.g. "uhm?"). Upon this interruption, the system provides a definition of the term, retrieved from *Wikipedia, the Free Encyclopedia*. The selection of the non-speech gestures has been made with a respect to the cross-cultural applicability and language independence. In this paper we present a set of novel tools enabling this kind of interaction.

1 Introduction

Text-to-speech sythesis (TTS) has become an established means of implementation of the output speech modality in various user interfaces. Its primary use is to mediate textual information in situations where the users may not or can not use their vision. Typical applications of TTS, besides various telephony applications, include systems presenting information to the visually impaired [8] or to the people in eyes-busy situations, such as drivers or pilots.

An important subset of these applications is reading of text documents. The TTS is frequently used for the low-cost rendering of audio books, such as in project Gutenberg [2]. Audio books are stored in a non-interactive compressed audio format (MP3, OGG, etc.) and accessed by means of common personal MP3 players. This use of TTS resembles radio drama broadcasts.

However, a similar approach could be used for accessing technical documents for the purpose of annotating unknown terms for the user so they may proceed through the material without switching to a dictionary or encyclopedia.

Technical texts, such as papers, manuals, technical reports, etc. are generally well-structured documents with the primary purpose to provide an objective and unequivocal

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 695–702, 2006.

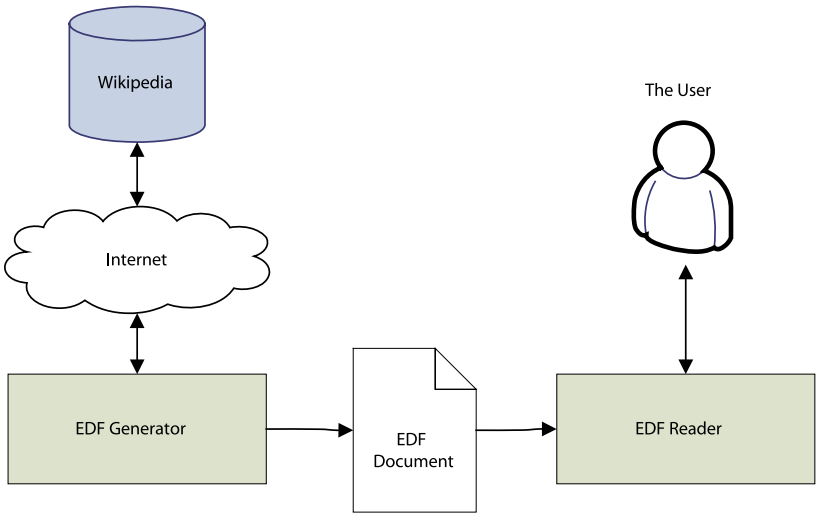

Fig. 1. Context of the system

description of a subject. However, this is different than in the case of books of fiction that many times use language that is not domain-specific or specialized. Technical documents contain numerous technical terms, that may be unknown to the reader, especially in the case when the reader's expertise does not entirely cover the subject of the document.

In our paper, we propose a system that combines the TTS with a non-speech input acoustic modality. The non-speech input is based on use of sounds other than speech, such as humming or whistling [7,5]. This interaction style has been investigated previously in [3] or [4]. In our system we use it for user-initiated switching between reading the document and explication of the terms contained in it.

Our system makes use of non-speech gestures which are short melodic patterns produced by humming. In general, each gesture may be assigned a system command which is executed upon the completion of the gesture. Our set of non-speech gestures is based on common non-verbal vocalizations that occur during general conversation.

Our system allows the user to interrupt the reading in the moment of miscomprehension of the text (see section 2.2), to which the system reacts by explaining the term in question. The user may stop the explanation of the term once they are sure they understood its definition. The system then resumes the reading, starting with the sentence that contained the explained term and thus allowing the user to regain the context of the text.

Another contribution of our paper is the method of automatic annotation of technical documents which fetches the explanations of all terms contained in the document.

The overall scheme of the system is shown in Figure 1. Our system is targeted to mobile autonomous systems working both on- and off-line. Therefore all explications have to be pre-fetched into a single document structure. We have defined such a structure and call it *Explicative Document Format*.

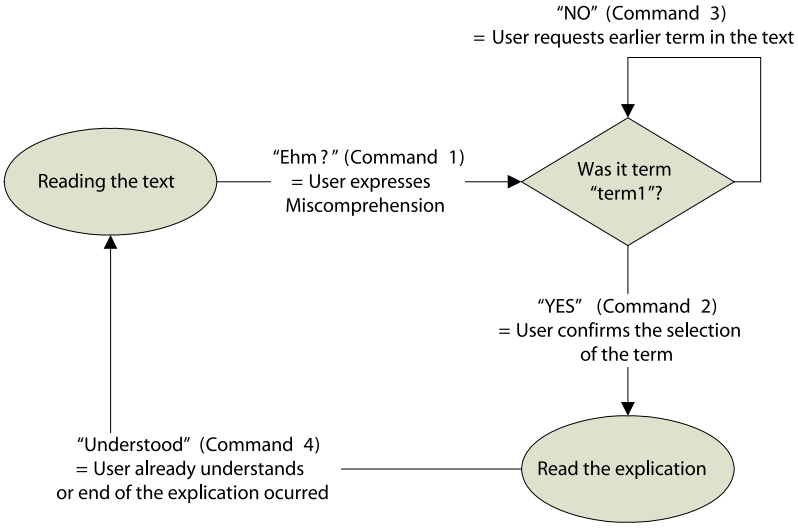

Fig. 2. User interface state diagram

2 System Description

Our system is based on an audio-only user interface, consisting of two modalities: A speech modality, used to transfer the text information to the user, and a non-speech gestures modality, used to control the reading process by the user.

The control mechanism has the following three states: reading the text, selection of the term for explanation, and explaining the term. The structure of control mechanism is explained in Figure 2. A female voice was used for reading the document whereas a male voice was employed to read the definitions.

2.1 Design of Non-speech Gestures

To determine a suitable set of non-speech audio gestures, we observed the occurence of non-speech sounds in common speech. This allowed us to identify four most frequent non-speech gestures, as shown in Fig. 3.

Subsequently, we have performed a semantic study of these gestures. The goal of the study was to determine their meaning as percieved by representatives of different cultures.

We have created a simple questionnaire whereby each participant of the study was asked what was the most obvious and natural interpretation of each of the four gestures in their own language and culture. The results of this study are summarized in Table 1.

The results show an agreement of meaning with most of the gestures as being understood by the participants from wide range of cultures. Gesture *a* was most frequently understood as an expression of a question. The common understanding of gesture *b* varied and therefore we decided not to use this gesture in our system. The gesture *c* was generally understood as agreement while the gesture *d* meant a disagreement for all participants. Considering these results, we decided for the assignment of the commands to the gestures as shown in Table 2.

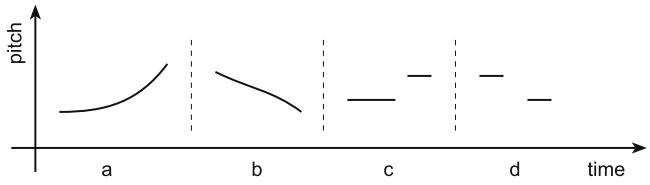

Fig. 3. Set of common non-speech gestures

Table 1. Interpretation of the acoustic gestures by the participants. The gesture labels corresponds to those in Figure 3.

Part.	Nationality	Gesture *a*	Gesture *b*	Gesture *c*	Gesture *d*
A	Austrian	"Question, agreement."	"Question."	"Yes."	"No."
B	Columbian	"Yes."	"Question."	"Yes."	"Negation."
C	Czech	"What?"	"Surprise."	"Agreement."	"Disagreement."
D	Czech	"Question."	"Negation."	"Yes."	"No."
E	Czech	"Question."	*N/A*	"Yes."	"No."
F	French	"What?"	"I don't think so."	"Yes."	"No."
G	German	"Question, agreement."	"Question."	"Yes."	"No."
I	Iranian	"What?"	"Interesting!"	"Yes."	"No."
J	Italian	"Surprise."	"Disagreement."	"Agreement."	"Disagreement."
K	Romanian	"Confusion."	"Part. satisfaction."	"Good."	"Bad."
L	Slovak	"What do you mean?"	"I don't think so."	"Agreement, willingness."	"Disagreement."
M	USA	"Don't know."	*N/A*	"yes"	"no"

2.2 Reaction Time upon Term Miscomprehension

From the user interface state diagram shown in Figure 2 it is clear that the usability of the system would be affected by the reaction time upon miscomprehension of a term, i.e. the time that passes between the user hears a term they do not understand and they express their miscomprehension.

We have carried out a brief experiment that was to answer the following question: "Upon a miscomprehension of a term, what is the common reaction time of the users after which they are likely to report the miscomprehension?"

There were seven participants of this experiment (1 female, 6 male). Only fluent English speakers were asked to participate.

Table 2. Assignment of gestures

Gesture *a* ("What?")	Command 1 (Explain!)
Gesture *c* ("Yes")	Command 2 (Yes)
Gesture *d* ("No")	Command 3 (No)
Gesture *a* ("Yes")	Command 4 (Understood)

Each participant was asked to listen to a technical text in English language, read by a TTS. The text was not presented in a visual form. Upon encountering a term the participant for some reason had not understood, he or she were asked to notify the experimenter by raising hand. After this notification, the playback of the text was interrupted. The participant was asked to state, what term was not understood. The experimenter highlighted the term in the text and made note in the text of when the user made the notification. Afterwards, the playback of the text was resumed.

For each miscomprehension, the reaction time in words has been recorded. 5 to 10 instances of miscomprehension were recorded for each participant. The Microsoft Mary voice, available in the standard installation of the Microsoft Windows XP Professional, was used as the text-to-speech engine [1].

The results are shown in table 3. The average reaction time was most usually between 2 and 4 words. In some extreme cases, the delay was as many as 8 words. Howewer, immediate responses and responses after a single word were noted in 22% of all records.

2.3 Explicative Document Format

In order to be able to operate in off-line conditions, our Explicative Document Format (EDF) has been designed to describe both main text and definitions of all terms in the document. Therefore, the EDF has two principal parts—the original text with tags around terms, and the explanatory parts, where all explications are stored.

Each definition has a unique identifier, which is used to link the terms in the text with their respective definitions. Figure 4 shows a short example of an EDF document. As the term which can be misunderstood can be longer than a word, the form of the tags has been designed in order to allow an overlapping structure. This fact disallowed us to use an XML-based format, as it does not allow the tag overlapping.

3 The Prototype Implementation

3.1 Creating The EDF Documents

To automate the creation of EDF documents we created a tool called EDF-WiKi, which scans the text for all possible terms, and attempts to fetch the explanations of the terms from the

Table 3. Reaction time upon a miscomprehension. D – reaction time (in words).

Subject	Gender	Age	EN Fluency	Culture	$\overline{D}$	SD(D)
B	M	50	near-native	Columbian	3.8	1.9
C	M	24	fluent	Czech	2.2	1.3
D	M	31	fluent	Czech	2.3	1.4
E	M	23	fluent	Czech	2.1	1.4
G	M	27	near-native	Greek	2.9	2.2
I	F	22	fluent	Iranian	3.0	2.4
M	M	40	native	US	1.9	1.5
Overall	—	—	—	—	2.8	1.9

Wikipedia [9] and creates the whole output document. From user point of view, the EDF-WiKi is straightforward: User opens a document and presses the button “EDF it!”. After all the information is retrieved from Wikipedia server, the program asks for a filename to save the EDF document.

From the programming point of view, the problem looks as follows: First, the document is scanned for all terms with length between certain number of characters (we used 4) and certain word count (we used 2; because of information retrieval time increases exponentially with the length of terms). For each term we remember it’s position in the text. Subsequently, the retrieval of all terms from Wikipedia is attempted. If Wikipedia answers that it does not know the particular term, the term is omitted, as it is probably common word. Finally, tags are inserted to the original text and explications are appended at the end of the document. The algorithm is more clearly shown in Figure 5.

3.2 EDF Documents Reader

The EDF reader is the run-time environment of our system. It is the implementation of the interaction paradigm, as described in section 2, allowing the user to access the documents in the EDF format.

Its prototype has been implemented in MS VC++ 6.0. Microsoft Speech API 5.1 [1] has been used to control the TTS. To detect the pitch of the humming, we have tested several techniques, such as zero-crossings, analysis of the FFT, and autocorrelation methods. For our purposes, we found the autocorrelation, as described in [6], to be the most efficient method.

4 The User Evaluation

In the user evaluation sessions, we asked four users (3 male, 1 female) to get acquainted with our system and use it for a while. We have used different *Wikipedia articles* that were not related to the participants’ fields of expertise. Each participant has been presented one article of about 600 words of length. When finished, we asked the following questions:

1. *Can you imagine the use of such a system while driving?* All users answered positively this question.

```
[1*Lorem [2*[3*ipsum*2]*1] dolor*3] sit amet, consectetuer
adipiscing elit. Morbi rutrum ...

|1*lorem ipsum. Sed eros leo, interdum sed, ullamcorper vel, porttitor a,
arcu. Aenean rhoncus ornare nisi. Nullam tempor nibh at est nonummy
placerat.
|2* ipsum. Proin venenatis placerat lectus. Proin porttitor quam sed
arcu. Sed laoreet lorem vitae felis. Maecenas elementum urna in magna.
|3* ipsum dolor. Cras nec pede. Sed mattis pellentesque metus. Cras
nisl. Vivamus a est nec nisi facilisis convallis. Vivamus luctus felis
sed augue. Proin pharetra ante sed justo. Quisque nec dui.
```

Fig. 4. Example of the EDF format. `[1*` ... start of a tag, `*1]` ... end of a tag, `|1*` ... start of a definition.

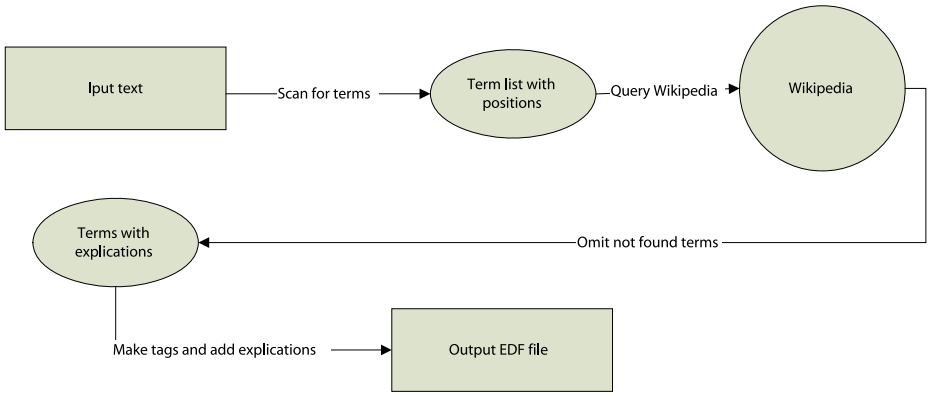

Fig. 5. EDF Generation Pipeline

2. *What other uses of this system you can think of?* Two users suggested the use of non-speech gestures for web navigation. One user suggested the adaptation of this system as a interactive manual for maintenance of machinery. Another user suggested a combination with the speech recognition.
3. *What benefits do you find in use of the non-speech gestures?* Three users answered that they found the non-speech gesture control intuitive. These users also reported that they found the set of gestures easy to use. One user expressed a concern that the non-speech gestures would interfere with the output speech modality.
4. *When using our system, did you find the explanation of the terms provided relevant?* Two users answered that in some cases the information presented was not that clear to them as the system gave an explanation that was not relevant to the use of the term in particular context. However, this is may be attributed to the data source [9] rather than to the interaction method itself.

The system was positively accepted by the users. All users were able to use the system only after a short explanation of its function. The non-speech gestures proved to be a viable modality for this particular application.

5 Conclusion

In this paper we have presented a system, which allows the user to listen to a technical text and ask for explanations of incomprehensible terms. The control of the system is performed by means of the non-speech acoustic gestures.

To control our system we have chosen a set of non-speech gestures according to the results of our study. The selected set of the non-speech gestures proved to be a natural way to input basic commands, such as agreement or disagreement. Our study has also proved that the selected gestures are cross-culturally applicable. From our experience, this control paradigm has a low cognitive load and can be used even during complex actions such as driving, etc.

The performed usability test proved a good usability of the whole system as well as the control by the sound gestures themselves. In future, we would like to track individual user

performance (such as the reaction time, etc.) and adapt the system responses accordingly in order to make the system even more user friendly, as well as to improve the fetching of the definitions from free on-line sources so that relevant definitions are extracted, depending on the context of the text.

Acknowledgment

Our sincere thanks belong to Catherine Forsman for proofreading of the text. The research was supported by the Ministry of Education, Czech Republic (MŠMT ČR), research program MSM 6840770014 and project No. 1M6138498401.

References

1. Microsoft Speech Application Program Interface (SAPI) Version 5.0. Online, retrieved 20 Mar 2006. `http://www.microsoft.com/speech`.
2. Project Gutenberg. Online, retrieved 20 Mar 2006. `http://www.gutenberg.org/`.
3. P. Hämäläinen, T. Mäki-Patola, V. Pulkki, and M. Airas. Musical computer games played by singing. In T. I. Evangelista G., editor, *Proc 7th Intl Conf on Digital Audio Effects, Naples, Italy*, pages 367–371, 2004.
4. T. Igarashi and J. F. Hughes. Voice as sound: using non-verbal voice input for interactive control. In *UIST '01: Proc 14th Annual ACM Symp on User Interface Software and Technology*, pages 155–156, New York, NY, USA, 2001. ACM Press.
5. A. J. Sporka, S. H. Kurniawan, and P. Slavík. Acoustic control of mouse pointer. *Universal Access in Information Society*, 4(3):237–245, 2006.
6. L. R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25(1):24–33, February 1977.
7. A. J. Sporka, S. H. Kurniawan, and P. Slavík. Non-speech operated emulation of keyboard. In *Designing Accessible Technology*. Springer-Verlag (London), 2006.
8. P. Žikovský, T. Pěšina, and P. Slavík. Processing of logical expressions for visually impaired users. In *Proceedings of TSD 2004, Lecture Notes in Artificial Intelligence LNCS/LNAI 3206*, pages 553–560. Springer-Verlag (Berlin), 2004.
9. Wikimedia Foundation, Inc. *Wikipedia, the free encyclopedia*. `http://www.wikipedia.org`.