Homer II—man–machine interface to internet for blind and visually impaired people

Nikola Pavešić, Jerneja Gros, Simon Dobrišek, France Mihelič[*]

Laboratory of Artificial Perception, Systems and Cybernetics, Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia

Received 12 May 2002; accepted 13 May 2002

Abstract

HOMER II is a voice-driven text-to-speech system developed for blind or visually impaired persons for reading Slovenian texts. Users can obtain texts from the Internet site of the Association of Slovenian Blind and Visually Impaired Persons Societies from their Electronic Information System where they can find daily newspapers, some novels and other information. The system consists of four main modules. The first module enables Internet communication, retrieves text to a local disc and converts it to a standard form. The input interface manages the keyboard entry and/or speaker independent speech recognition. The output interface performs speech synthesis of a given text and in addition prints the same text magnified to the screen. The user dialog is responsible for the user friendly communication and controls other tasks of the system. Homer II was ported from Linux to the MS Windows 9x/ME/NT/2000 operating systems. For the best performance it uses multi-threading and other advantages of the 32-bit environment. Further versions of the HOMER system with even more advanced dialogue modules and some basic World Wide Web browsing functionality will represent an important tool in the distance learning and teaching process for the impaired persons using academic networks.

Keywords: Internet interface; Blind and visually impaired users; Text-to-speech; Speech recognition; Dialogue management

1. Introduction

The problems in communication some ill, injured or disabled persons may have are well known. When they are unable to use their hands, speak or read, they are forced to use technical aids to overcome their problems [12]. For blind or visually impaired persons the Braille coding of texts is a common aid. This type of coding requests special editions of written corpora or special additional hardware components when used with computers. The solution is relatively costly and requires special users skills.

On the other hand, automatic text-to-speech and speech recognition systems can be used for this purpose. Such systems have many advantages. They are user friendly and encourage a natural way of communication. The communication goal can be achieved faster and they offer access to large text corpora via modern technical equipment (over the computer network, scanners, etc.) and have a relatively low price. However, these new speech technologies are strongly language dependent (especially speech synthesis) and general solutions for all languages cannot be applied directly. If speech technologies are to be used with the Slovenian language, language dependent parts of the systems must be developed for this purpose, using knowledge from the Slovenian language phonology, syntax and semantics.

The HOMER II voice-driven text-to-speech system developed for blind or visually impaired persons for accessing and reading Slovenian texts is presented in this paper. The HOMER II system demonstrates how the way of accessing daily news and other useful information can be improved for these disabled users.

The HOMER system can be integrated into Internet based applications enabling disabled persons to remotely access all the written information available in a given language. In this way the system will also represent an important tool in the distance learning process for impaired persons using academic networks.

[*] Corresponding author. Tel.: +386-1-4768-313; fax: +386-1-4768-316.
E-mail address: mihelicf@fe.uni-lj.si (F. Mihelič).

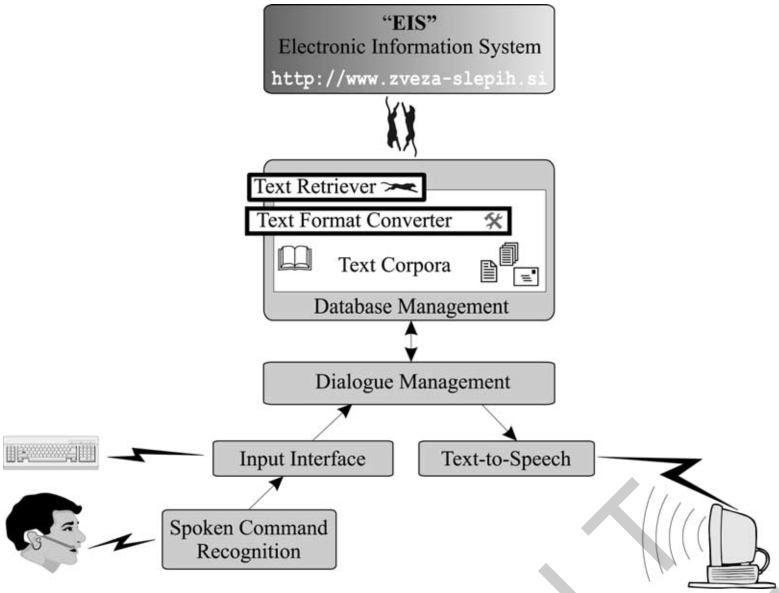

Fig. 1. The system structure.

2. System structure

The HOMER II system is a database interface, where the database are text corpora organised at the information source centre. The system incorporates a spoken language interface which is essential for blind and visually impaired users, especially the text-to-speech part. The voice control of the system additionally facilitates working with the system as there is no need for using mechanical interfaces such as a keyboard or mouse.

The system consists of four main modules (Fig. 1). The first module enables Internet communication, retrieves text to a local disc and converts it to a standard form. The second module is the one that manages dialogues with users and performs access to the text database. The next one is the text-to-speech output module which enables automatic generation of Slovenian speech from an arbitrary Slovenian text written in a non-structured text format using one of the standard character codings, like the Slovenian version of the 7-bit ASCII coding or the WIN-1250 and the ISO-8859-2 coding. Input to the system is performed via a keyboard, with some special selected keys or using the speaker-independent spoken command recognition module which runs in parallel with the other modules.

Table 1

"Da!"	*Yes!*
"Ja!"	*Yes!*
"Ne!"	*No!*
"Ponovi!"	*Repeat!*
"Naprej!"	*Forward!*
"Nazaj!"	*Backward!*
"Na začetek!"	*To the beginning!*
"Na konec!"	*To the end!*
"Prekini!"	*Break!*

The HOMER II system was designed to be implemented on a Pentium PC compatible computer with a minimum 32 Mb RAM with a built-in standard 16-bit sound card. The system was ported from Linux and DOS [8] to the MS Windows 9x/ME/NT/2000 operating systems. For the best performance it uses multi-threading and other advantages of the 32-bit environment. It requires approximately 10 Mb of disk space for the program code and for the text-to-speech and speech recogniser module inventory, and the additional storage facilities for Slovenian newspapers files or other texts arranged according to a predefined structure.

3. Internet communication

The strategy of the text file acquisition task, was switched from the specially designed modem communication protocol in the first version [8] to the standard Internet communication protocol.

The system was built for users, who are connecting to the Internet from their home, via a modem and their own computer using the Windows operating system. Normal connection is established via dialog windows using a mouse, special keys and entering passwords and some other keywords. Since we could not expect this procedure to be performed by blind or visually impaired people, a special software solution was designed to establish the connection and automatic text file acquisition. The Microsoft product RAS API (Remote Access Services Application Programming Interface) [10] was used to build the interface. The interface incorporates a feature which enables the establishment of the dial-up connection with the PPP (Point-to-Point Protocol) protocol with the Internet provider and is a part of the MS Windows 9x/ME/NT/2000 operating systems.

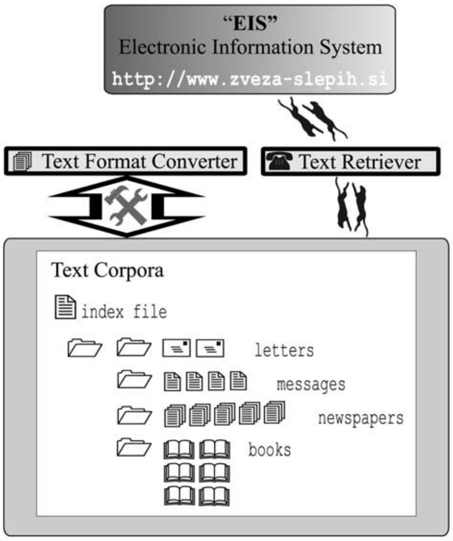

Fig. 2. The text corpora organisation.

4. Dialogue

The dialogue module manages dialogues with users, accesses to the text database and performs the system control function. It was designed with respect to the experiences we gained with the design of a similar dialogue module in another speech recognition system [11]. In order to enable robust speech recognition performance, we carefully selected a small number of nine easy-to-remember commands.

The following nine control commands (with English translations): Table 1 were chosen for the complete control of the system.

4.1. Text database organisation

The text database is organised in the form of a file system with simple partially structured text files (Fig. 2). The text database is automatically formed from the non-structured text files acquired from the Electronic Information System (EIS) of the Association of Slovenian Blind and Visually Impaired Persons Societies [4]. The available text files represent a few Slovenian daily newspapers, some Slovenian novels and daily messages from the centre's administration.

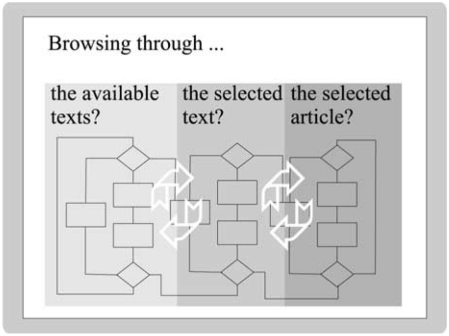

Fig. 3. Browsing through the text corpora.

The database is organised as a tree structure. The first level represents the index of all available texts (newspapers, instructions, etc.) acquired from the information centre. The second level consists of the table of contents (titles) for the available texts from the first level. The contents can be organised in two different modes, using topic descriptions or page numbers. The third level finally includes the texts, which are formatted in the form: title, subtitle, date, author, and body of the texts separated into paragraphs. The dialogue module enables transitions between these three levels (Fig. 3) at any stage of processing and navigating through the offered index.

4.2. Dialogue manager

The voice-driven dialogue manager enables the user to:

- start and shut down the HOMER II system;
- obtain read instructions for using the system;
- obtain read index of the available text;
- obtain read index of the articles in the text;
- navigate through the offered index of texts, topics, pages or articles;
- select the text and article to be read;
- obtain descriptive information about the selected article (subtitle, date, place, author);
- obtain quantitative information about the selected article (number of paragraphs, number of words);
- skip a paragraph of the selected article;
- start and stop the text-to-speech process.

5. Speech recognition

The speech recognition module performs two subsequent tasks: speech detection and isolated spoken command recognition. Speech recognition is presently performed as a server application and is connected to a user interface via the TCP/IP protocol. In this way exacting processing power for speech recognition could be reached by some other computer and the whole speech recognition module can be easily changed with another one, if necessary.

5.1. Speech detection

Speech detection is a separate part of the speech recognition module. It provides online information on whether a user is uttering a command. In this way the user can interrupt the text reading and switch to another article or newspaper or he/she can shut down the whole process at any time. We paid special attention to careful speech detection as it is very important for the accuracy of isolated spoken command recognition, where the feature string pattern from a silence/speech/silence acoustic event is requested.

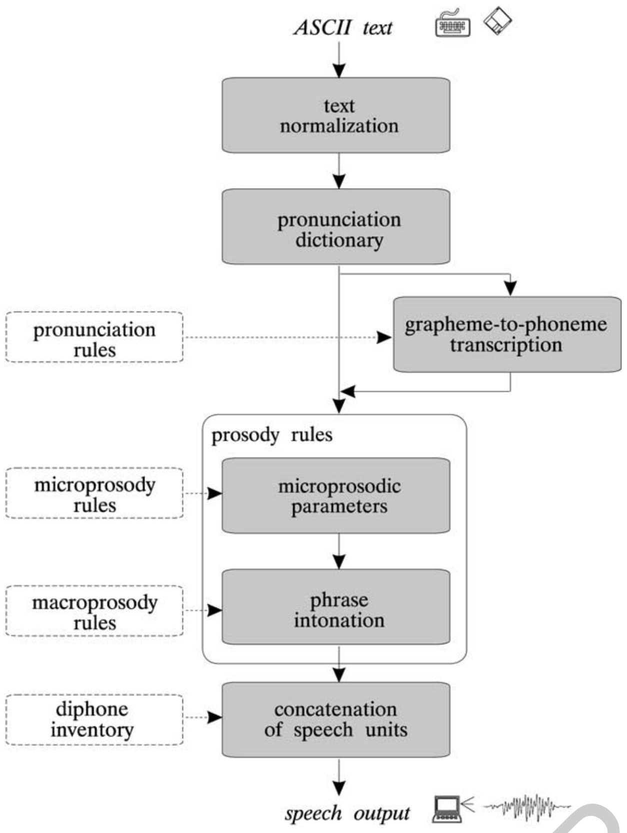

Fig. 4. The Slovenian text-to-speech system structure.

The speech detection algorithm is mainly based on the evaluation of log-energy and pitch detection on different time intervals of the incoming signal, searching for the mentioned silence/speech/silence acoustic event [3].

5.2. Spoken command recognition

Isolated spoken command recognition is based on the well-known Hidden Markov Modelling (HMM) of acoustic events [14]. In the system phone-like subword units are used as the fundamental units in silence/word/silence HMM models. Therefore, the number of states per word is not unique. Discrete HMM modelling is used, based on the prior vector quantisation of incoming feature vectors into 128 clusters using the k-means algorithm. Log-energy and the first 11 MEL-cepstrum coefficients determined on 10 ms overlapping signal frames are used to construct a 36-dimensional feature token [1]. The whole speech recognition module is designed in an open manner, enabling fast adaptation to different applications of isolated word recognition, also for larger vocabularies of up to several 100 words. Special computer programs for fast and user-friendly procedures of collection, segmentation and labelling of speech material and the training process can be used [1,2]. The recognition procedure also offers the unrecognised word category classification, which activates a request for repetition of the command.

5.3. Spoken command recognition accuracy

A preliminary off-line evaluation of the spoken command recognition accuracy, using a clean speech database of 20 training speakers and 6 testing speakers, yielded a recognition error rate lower than 2% [3]. However, the actual recognition rate is strongly dependent on the user's behaviour while interacting with the HOMER II system. In practise, the online recognition error rate increases, but remains below 5%.

6. Text-to-speech

For the automatic conversion of the output text into its spoken form the Slovenian text-to-speech system *S5* [5,7] based on diphone concatenation (Fig. 4) was applied. The input text is transformed into its spoken equivalent by several modules. A grapheme-to-phoneme or grapheme-to-allophone module produces strings of phonetic symbols based on information in the written text. A prosodic generator assigns pitch and duration values to individual phones. Final speech synthesis is based on diphone concatenation using TD-PSOLA [13]. The quality of the synthesised speech was assessed in terms of intelligibility and naturalness of pronunciation. Additionally, various aspects of the synthetic speech production process were tested.

6.1. Grapheme-to-allophone transcription

Input to the S5 system is the output text obtained from the HOMER II system. It is translated into a series of allophones in two consecutive steps. First, input text normalisation is performed. Abbreviations are expanded to form equivalent full words using a special list of lexical entries. The text normaliser converts further special formats, like numbers or dates, into standard graphemic strings. The rest of the text is segmented into individual words and basic punctuation marks. Next, word pronunciation is derived, based on a user-extensible pronunciation dictionary and letter-to-sound rules. The dictionary covers over 16,000 most frequent inflected word forms. In case where dictionary derivation fails, words are transcribed using automatic lexical stress assignment and 150 context dependent letter-to-sound rules.

6.2. Prosody generation

Prosody generation in S5 consists of four phases: intrinsic duration assignment, extrinsic duration assignment, modelling of the intra word F0 contour and assignment of a global intonation contour.

A two-level duration model first determines the words' intrinsic duration, taking into account factors relating to the phone segmental duration, such as: segmental identity, phone context, syllabic stress and syllable type (open or

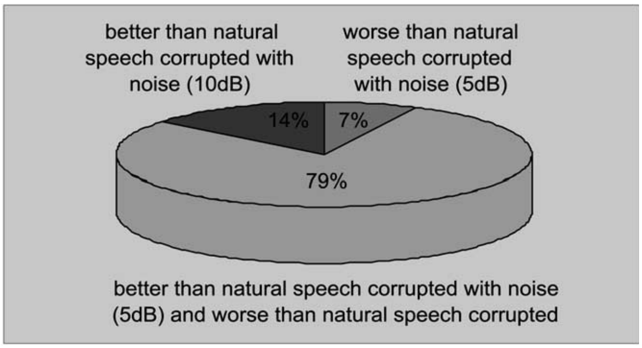

Fig. 5. TTS system performance.

closed syllable). Further, the extrinsic duration of a word is predicted, according to the higher-level rhythmic and structural constraints of a phrase, operating on a syllable level and above. The following factors are considered: the chosen speaking rate, the number of syllables within a word and the word's position within a phrase, which can be isolated, phrase initial, phrase final or nested within the phrase.

Finally, intrinsic segment durations are modified, so that the entire word acquires its predetermined extrinsic duration. It should be noted, that stretching and squeezing does not apply to all segments equally. A method for segment duration prediction was developed, which adapts a word with an intrinsic duration *ti* to the determined extrinsic duration *te*, taking into account how stretching and squeezing apply to the duration of individual segments [6].

Since the Slovenian language has been defined as a pitch accent language, special attention was paid to the prediction of tonemic accents for individual words. First initial vowel fundamental frequencies were determined according to previous measurements as suggested by Ref. [15], creating the F0 backbone. Each stressed word was assigned one of the two tonemic accents characteristic for the Slovenian language. Five typical F0 patterns were chosen from the variety of F0 patterns described in Ref. [15]. Finally a linear interpolation between the defined F0 values was performed.

We used a relatively simple approach for prosody parsing and the automatic prediction of Slovenian intonational prosody which makes no use of syntactic or semantic processing, but rather uses punctuation marks and searches for grammatical words, mainly conjunctions which introduce pauses [5].

6.3. Diphone concatenation

The final step within the TTS system is to produce audible speech by assembling elemental speech units. This is achieved by taking into account computed pitch and duration contours, and synthesising a speech waveform. A concatenative synthesis technique was used. The TD-PSOLA scheme permits pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications [15] without considerably affecting the quality of the synthesised speech. Diphones were chosen for concatenative speech units as a compromise between the size of the unit inventory, the complexity of the concatenation rules and the resulting speech quality.

6.4. Synthesis evaluation

The quality of the synthetic speech output was evaluated in terms of naturalness and intelligibility. The experiment was performed according to the ITU-T Recommendation P.85, which defines a testing method for evaluating the subjective quality of synthetic speech in real application voice servers available for Public Switched Telephone Network subscribers. The method takes into account both the performance and the attitudes of the users. The attitudes are assessed by the use of multiple scales.

The subjects were asked to fill in different templates in their response sheets related to the chosen application domain based on the information they heard. The application domain chosen was airline timetable information retrieval. Over 90% of the templates were filled in correctly. The wrongly understood items were mainly names of foreign airports, quite unknown to the audience and difficult to spell. About two thirds of the test subjects considered that the TTS system was appropriate to be implemented in an automatic information retrieval system. The remaining third often commented, that the synthetic speech quality was sufficiently high, however, they strongly opposed the process of machines taking over human work.

The second part of the test served to compare several features describing the synthetic voice quality to those describing the quality of natural speech distorted with different levels of gaussian noise. The experiment was carried out according to ITU-T Recommendation P.81, describing a method of comparing synthetic speech to natural speech distorted by a modulated noise reference unit. The synthetic speech received a mean opinion score, which was between distorted natural speech with a SNR ratio of 5 and 10 dB, as shown in Fig. 5.

7. Conclusions and future work

The voice-driven text-to-speech system for the Slovenian language is presented. The prototypes of the system are already in use and are undergoing evaluation tests. Improvements in the sense of more accurate and robust speech recognition and a user-friendly system to control high quality speech synthesis are planned for the future. Some work on a speech input which incorporates a larger dynamical list of permitted spoken control commands is already in progress. We are expecting further suggestions from the blind and visually impaired community, especially on the design of the strategy for communication with the system and, of course, remarks on the Slovenian speech synthesis quality. Many measurements and research in the field of micro and macro prosody modelling of Slovenian speech should be done as well as recordings of new diphone databases with different speakers. An extension of the HOMER system enabling arbitrary Slovenian texts over the Internet to be read is also planned for the future.

Blind and visually impaired people should be offered the facilities for usual World Wide Web browsing, however, there are many factors, which make this difficult. This challenge is covered by the Web Access Initiative (WAI) [16]. Further versions of the HOMER system will have some basic Web browsing functionality and the available text corpora will be reorganised and transferred from the existing EIS [4] to the new Web portal [17], exclusively dedicated to blind and visually impaired users. This latest research efforts were partially supported by a donation from the Hewlett Packard philanthropic project [9].

Acknowledgements

This work was partly supported with Grant No. 3411-00-22 2109 from the Slovenian Ministry of Science and Technology (MZT) and Association of Slovenian Blind and Visually Impaired Persons Societies (ZDSSS).

References

[1] S. Dobrišek, F. Mihelič, N. Pavešić, Merging of time delayed feature vectors into extended vector in order to improve phoneme recognition, Proceedings of the Fourth COST #229 Workshop on Adaptive Methods and Emergent Techniques for Signal Processing and Communications, Ljubljana, Slovenia (1994) 145–150.

[2] S. Dobrišek, J. Gros, F. Mihelič, N. Pavešić, Recording and labelling of the GOPOLIS Slovenian speech database, Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain II (1998) 1089–1096.

[3] S. Dobrišek, Analysis and Recognition of Phones in Speech Signal, PhD Thesis (In Slovenian), University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia, 2001.

[4] Electronic Information System ZDSSS, http://www.zveza.slepih.si/zdsss/eis/, July 2001.

[5] J. Gros, N. Pavešić, F. Mihelič, Text-to-speech synthesis: a complete system for the Slovenian language, Journal of Computing and Information Technology CIT-5 (1) (1997) 11–19.

[6] J. Gros, N. Pavešić, F. Mihelič, Speech timing in Slovenian TTS, Proceedings of EUROSPEECH’97, Rhodes, Greece (1997) 323–326.

[7] J. Gros, F. Mihelič, N. Pavešić, S5: an automatic reading system for Slovene, Proceedings of the International COST #254 Workshop on Intelligent Communications and Multimedia Terminals, Ljubljana, Slovenia (1998) 67–70.

[8] S. Dobrišek, J. Gros, F. Mihelič, N. Pavešić, HOMER: a voice-driven system for Slovenian text-to-speech synthesis, Proceedings of the International Workshop on Intelligent Communications and Multimedia Terminals, Ljubljana, Slovenia (1998) 71–74.

[9] HP Voice Web Initiative, http://webcenter.hp.com/grants, July 2001.

[10] A. Skonnard, Essential WinInet, Using the Windows Internet API with RAS, ISAPI, ASP, and COM, Addison-Wesley, Reading, MA, 1999.

[11] I. Ipšić, F. Mihelič, S. Dobrišek, J. Gros, N. Pavešić, Overview of the spoken queries in European languages project: the Slovenian spoken dialog system, Proceedings of the Scientific Conference on Artificial Intelligence in Industry, High Tatras, Slovakia (1998) 431–438.

[12] H. Levitt, Processing of speech signals for physical and sensory disabilities, Proceedings of the National Academy of Sciences of the United States of America 92 (22) (1995) 9999–10006.

[13] E. Moulines, F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication 9 (1990) 453–467.

[14] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, New Jersey, 1993, p. 507.

[15] T. Srebot-Rejec, Word Accent and Vowel Duration in Standard Slovene: An Acoustic and Linguistic Investigation, Slawistische Beitraege, Band 226, Verlag Otto Sagner, München, 1988.

[16] Web Access Initiative, www.w3.org/TR/WAI-WEBCONTENT, July 2001.

[17] Kalliope World Wide Web Portal, http://kalliope.fe.uni-lj.si, July 2001.