HOMER: a Voice-Driven Text-to-Speech System for the Blind

Simon Dobrišek, Jerneja Gros, France Mihelič and Nikola Pavešić

University of Ljubljana

Faculty of Electrical Engineering

Tržaška 25, SI-1000 Ljubljana, Slovenia

simond@fe.uni-lj.si

Abstract—HOMER is a voice-driven text-to-speech system developed for blind or visually impaired persons for reading the Slovenian texts. Users can obtain texts from the special corpora organised on the computer network server at the information centre of the association of the Slovenian blind and visually impaired persons. The system consists of three main modules. The text-to-speech module enables speech synthesis from an arbitrary Slovenian text input, the speech recognition module performs speaker independent isolated word recognition and the dialogue module controls the different tasks of the HOMER system and obtains texts from the source text corpora. Presently, the system runs under LINUX and requires a PENTINUM/133 PC with minimum 32 MB of RAM and an additional standard 16 bit sound card.

I. INTRODUCTION

The problems in communication some ill, injured or disabled persons may have are well known. When they can not use their hands, can not speak or can not read, they are forced to use technical aids to overcome their problems [8]. For blind or visually impaired persons the Braille coding of texts is an usual 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 requests special skills of the user.

On the other hand, automatic text-to-speech and speech recognition systems can also be used to for this purpose. Such systems have a lot of advantages. They are user friendly and encourage the natural way of communication. The communication goal can be faster achieved and they offer access to large text corpora via modern technical equipment (over the computer network, scanners etc) and they have a relatively low price. However, this new speech technology is strongly language dependent (especially speech synthesis) and general solutions for all languages can not be applied directly. If speech technology is 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 voice-driven text-to-speech system HOMER developed for the blind or visually impaired persons for accessing and reading Slovenian texts is presented in this paper. The HOMER system demonstrates how the way of accessing daily news and other useful information can be improved for these disabled users.

II. SYSTEM STRUCTURE

The HOMER system is a database interface, where the database are text corpora organised at the informa-

tion source centre. The system incorporates a spoken language interface which is essential for the blind and visually impaired users, especially the text-to-speech part. The voice control of the system additionally facilitates the work with the system as there is no need for using mechanical interfaces such as keyboard or mouse.

The system consists of three main modules (Figure 1). The first 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 non-structured text format using one of the standard character codings, like the Slovenian version of the 7 bit ASCII coding or even the WIN-1250 and the ISO-8859-2 coding. Input to the system is the speaker independent spoken command recognition module which runs in parallel with the other modules.

Fig. 1. The system structure.

The HOMER system was designed to be implemented on a Pentium PC compatible computer with minimum 32 Mb RAM with a built-in standard 16 bit sound card. The system runs under the LINUX operating system and requires approximately 10 Mb of disk space for 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.

III. DIALOGUE MODULE WITH THE DATABASE INTERFACE

The dialogue module manages dialogues with users, accesses to the text database and performs the control of the system. It was designed with respect to the experiences we gained with the design of a similar dialogue module in another speech recognition system [7]. In order to enable robust speech recognition performance we carefully selected a small number of nine easy-to-remember commands. These nine control commands ("Da!" - *Yes!*, "Ja!" - *Yes!*, "Ne!" - *No!*, "Ponovi!" - *Repeat!*, "Naprej!" - *Forward!*, "Nazaj!" - *Backward!*, "Na začetek!" - *To the beginning!*, "Na konec!" - *To the end!*, "Prekini!" - *Break!*) were chosen for the complete control of the system.

A. Text Database Organisation

The text database is organised in form of a file system with simple partially structured text files (Figure 2). The text database is automatically formed from the non-structured text files acquired from the information centre of the association of the Slovenian blind and visually impaired persons. For automatic text file acquisition we developed a special computer modem programme. Presently, we are changing the strategy of the text file acquisition task, switching from the specially designed modem communication protocol to the standard Internet communication protocol.

The available text files represent a few Slovenian daily newspapers, some Slovenian novels and daily messages from the centre's administration.

Fig. 2. The text corpora organisation.

The database is organised as a three level list of lists. The first level represents the list 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 (Figure 3) at any stage of processing and navigating through the offered list.

B. Dialogue Manager

The voice-driven dialogue manager enables to:

- start and shut down the HOMER system;
- obtain read instructions for the use of the system;
- obtain read lists of the available text;
- obtain read lists of the articles in the text;
- navigate through the offered list of texts, topics, pages or articles;
- select the text and select the article to be read;
- obtain descriptive information about selected article (subtitle, date, place, author);
- obtain the 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.

Fig. 3. Browsing through the text corpora.

IV. SPEECH RECOGNITION MODULE

The speech recognition module performs two subsequent tasks: speech detection and isolated spoken command recognition.

A. 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 she/he can shut down the whole process at any time. We put special attention on careful speech detection as it is very important for the accuracy of the isolated spoken commands recognition, where the feature string pattern from a silence - speech - silence acoustic event is requested. 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].

V. SPOKEN COMMAND RECOGNITION

Isolated spoken command recognition is based on the well-known Hidden Markov Modelling (HMM) of acoustic events [10]. 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 quantization of incoming feature vectors into 128 clusters using the k-means algorithm. Log-energy and the first 11 MEL-cepstrum coefficients determined on 10 msec overlapping signal frames are used to construct a 36 dimensional feature token [1]. The whole speech recognition module is designed in a open manner, enabling fast adaptation to different applications of isolated word recognition, also for larger (up to a few 100 words) vocabularies. Special computer programmes for fast and user friendly procedures of collection, segmentation and labelling of speech material and training process can be used [1], [2]. The recognition procedure also offers the unrecognised word category classification, which activates the request for the repetition of the command.

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

VI. TEXT-TO-SPEECH MODULE

For the automatic conversion of the output text into its spoken form the Slovenian text-to-speech system S5 was applied [4], [6] based on diphone concatenation (Figure 4). 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 [9]. 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.

A. Grapheme-to-Allophone Transcription

Input to the S5 system is the output text obtained from the HOMER 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.

B. 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.

Fig. 4. The Slovenian Text-to-Speech system structure.

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 closed syllable. Further, the extrinsic duration of a word is predicted, according to 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 is to 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 t_i to the determined extrinsic duration t_e , taking into account how stretching and squeezing apply to the duration of individual segments [5].

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 [11], 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 [11]. 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 international 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 [4].

C. Diphone Concatenation

The final step within S5 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 enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications [11] 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.

VII. 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, a user friendly system control high quality speech synthesis are planned for the future. Some work on a speech input which incorporates a larger dynamical list of allowed 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 the 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 data bases with different speakers. An extension of the HOMER system to enable reading arbitrary Slovenian texts over the Internet is also planned for the future.

ACKNOWLEDGEMENTS

The authors would like to thank the Slovenian Ministry of Science for granting part of the conference fees.

REFERENCES

- [1] S. Dobrišek, F. Mihelič, N. Pavešić, "Merging of Time Delayed Feature Vectors into Extended Vector in Order to Improve Phoneme Recognition", in *Proceedings of the 4th COST*

- #229 Workshop on Adaptive Methods and Emergent Techniques for Signal Processing and Communications, Ljubljana, Slovenia: 1994, pp. 145-150.
- [2] S. Dobrišek, J. Gros, F. Mihelič, N. Pavešić, "Recording and Labelling of the GOPOLIS Slovenian Speech Database", in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain: 1998, Vol. II, pp. 1089-1096.
- [3] S. Dobrišek, *Analysis and Recognition of Phones in Speech Signal*, Ph.D. Thesis in preparation (In Slovenian). University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia: 1999.
- [4] J. Gros, N. Pavešić, F. Mihelič, "Text-to-speech synthesis: A complete system for the Slovenian language", *Journal of Computing and Information Technology*, Vol. CIT-5, no. 1, 1997, pp. 11-19.
- [5] J. Gros, N. Pavešić, F. Mihelič, "Speech timing in Slovenian TTS", in *Proceedings of the EUROSPEECH'97*, Rhodes, Greece: 1997, pp. 323-326.
- [6] J. Gros, F. Mihelič, N. Pavešić, "S5: an Automatic Reading System for Slovene", in *Proceedings of the International COST #254 Workshop on Intelligent Communications and Multimedia Terminals*, Ljubljana, Slovenia: 1998 pp. 67-70.
- [7] I. Ipšić, F. Mihelič, S. Dobrišek, J. Gros, N. Pavešić, "Overview of the Spoken Queries in European Languages Project: The Slovenian Spoken Dialog System", in *Proceedings of the Scientific Conference on Artificial Intelligence in Industry*, High Tatras, Slovakia: 1998, pp. 431-438.
- [8] H. Levitt, "Processing of speech signals for physical and sensory disabilities", in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 92, No. 22, 1995, pp. 9999-10006.
- [9] E. Moulines, F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communication*, Vol. 9, 1990, pp. 453-467.
- [10] L. Rabiner, B. H. Juang, *Fundamentals of speech recognition*, PTR Prentice Hall, NJ: 1993. p. 507
- [11] T. Srebot-Rejec, "Word Accent and Vowel Duration in Standard Slovene: An Acoustic and Linguistic Investigation", *Slavistische Beitrage*, Band 226, Verlag Otto Sagner, München: 1988.