Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

A multi-language text-to-speech module

Carlson, R. and Granström, B. and Hunnicutt, S.

journal: STL-QPSR
volume: 22
number: 4
year: 1981
pages: 018-028

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

II. SPEECH SYNTHESIS

A. A MULTI-LANGUAGE TEXT-TO-SPEECH MODULE*

Rolf Carlson, Björn Granström, and Sheri Hunnicutt

Abstract

Recent advances in microprocessor, memory and signal processor technology have made it feasible to put complete speech processing equipment into a portable form.

At our laboratory a higher-level programming language has been developed that is especially suitable for rule description of linguistic processes. Text-to-speech systems for several languages have been written in this framework. Now also a cross compiler has been designed for the 16-bit microprocessor MC 68000, which makes transportation of programs from our research computer trivial. The language-independent parts of the program for the microprocessor are written in efficient assembler code.

Special hardware development has resulted in a portable, battery-operated unit that is capable of transforming text-to-speech at a speaking rate of 250 wpm (words per minute). This opens up the possibility of speech options on computer terminals, portable, large vocabulary talking language translators, etc.

The module has been tried in several applications for people with communication handicaps.

Introduction

Advances in speech research and electronics have made available a new class of devices based on different kinds of artificial speech. The devices range from those with a small fixed vocabulary of coded natural speech such as talking toys to unrestricted text-to-speech systems. We are here concerned exclusively with devices of the latter type that transform any text to synthetic speech. Since text-to-speech systems offer new possibilities for information display, the functional specification of the devices are in many cases still to be made.

At our laboratory a minicomputer-based text-to-speech system was developed three years ago and has since been used in different experimental applications, mostly as aids for persons with communication handicaps (1,2).

---

*This paper will be presented at the 1982 IEEE-ICASSP, Paris, France.

The present paper describes a new development based on state-of-the-art hardware, such as a 16-bit microprocesor, dense memory chips and a signal processor. In the resulting portable battery-operated device (Fig. II-A-1), we have tried to accommodate as many of the suggestions for improvement as possible based on the experience from the previous prototype.

We will also present some applications of the text-to-speech module in aids for the handicapped.

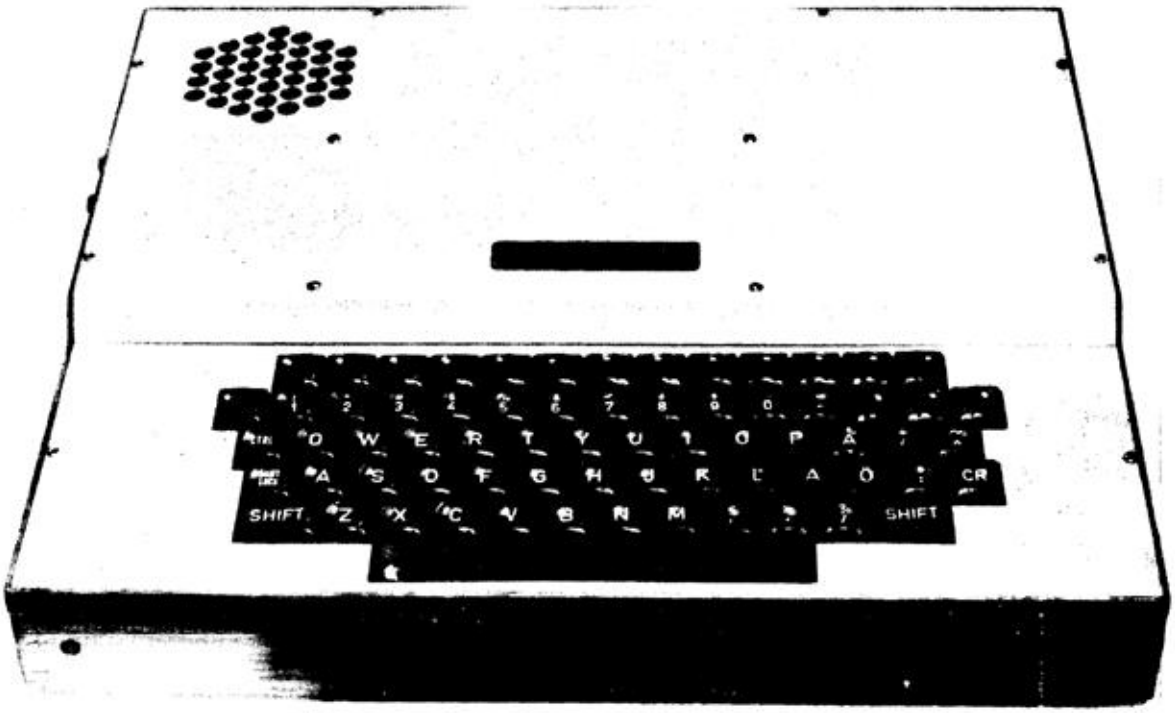

Fig. II-A-1. The text-to-speech module with key-board

Stucture of the Text-to-Speech Module

Our ambition has been to comprehensively deal with the text-to-speech conversion process. We want to do this for any kind of text , for several different languages in a way that makes the developed algorithms easy to implement in a portable system with state-of-the-art technology.

We have chosen to base the system mainly on rules, rather than on an extensive pronunciation dictionary. The reason for doing so is that the languages tend to contain a prohibitively large amount of low frequency words. In a Swedish one-million-word newspaper sample, over

100,000 different words were found (3). Another reason to try to formulate the text-to-speech process in rules is that the explicit formulation of these rules has an independent interest for linguistic research.

The rules of our system are formulated in a special higher level programming language. This notation parallels closely the notation used in generative phonology. It expands on this notation in that it is possible to operate on both variables and linguistic features associated with string elements (4).

Different parts of the text-to-speech process have been written as separate rule packages that can be connected in an appropriate way by a supervising program. The configuraton of a basic system can be seen in Fig. II-A-2.

Lexicon

The input text first meets the "lexicon" which typically contains a few hundred words. The lexicon is not formulated in the rule notation, but rather has a simple table structure. Besides the phonetic representation of the words, it can also contain syntactic information, e.g., word class , declension category, etc. The main function of this component in the present system is to identify frequent words that are usually unstressed, the so-called function words. Exceptions to the pronunciation rules can, of course, be included and frequent content words can also be added for speed reasons. Strings that are found in the lexicon are changed accordingly and passed on to the "phonetic rules" component.

Number Rules

In this component, digits and certain other non-letter characters are expanded to pronounceable phonetic representations and passed to the "phonetic component". An expression such as $4.50 will be pronounced as "four dollars and fifty cents". To some extent, this component is application-dependent. If, for example, a six-digit number is a telephone number, it should be pronounced according to other rules than an amount.

Text-to-Phonetic Rules

This component processes the greatest part of a normal input text. The size of this rule system depends very much on the language. For Swedish and English, it amounts to several hundred rules. For Spanish and Finnish it suffices with about 50 rules due to the close relation between spelling and pronunciation for these languages. This component

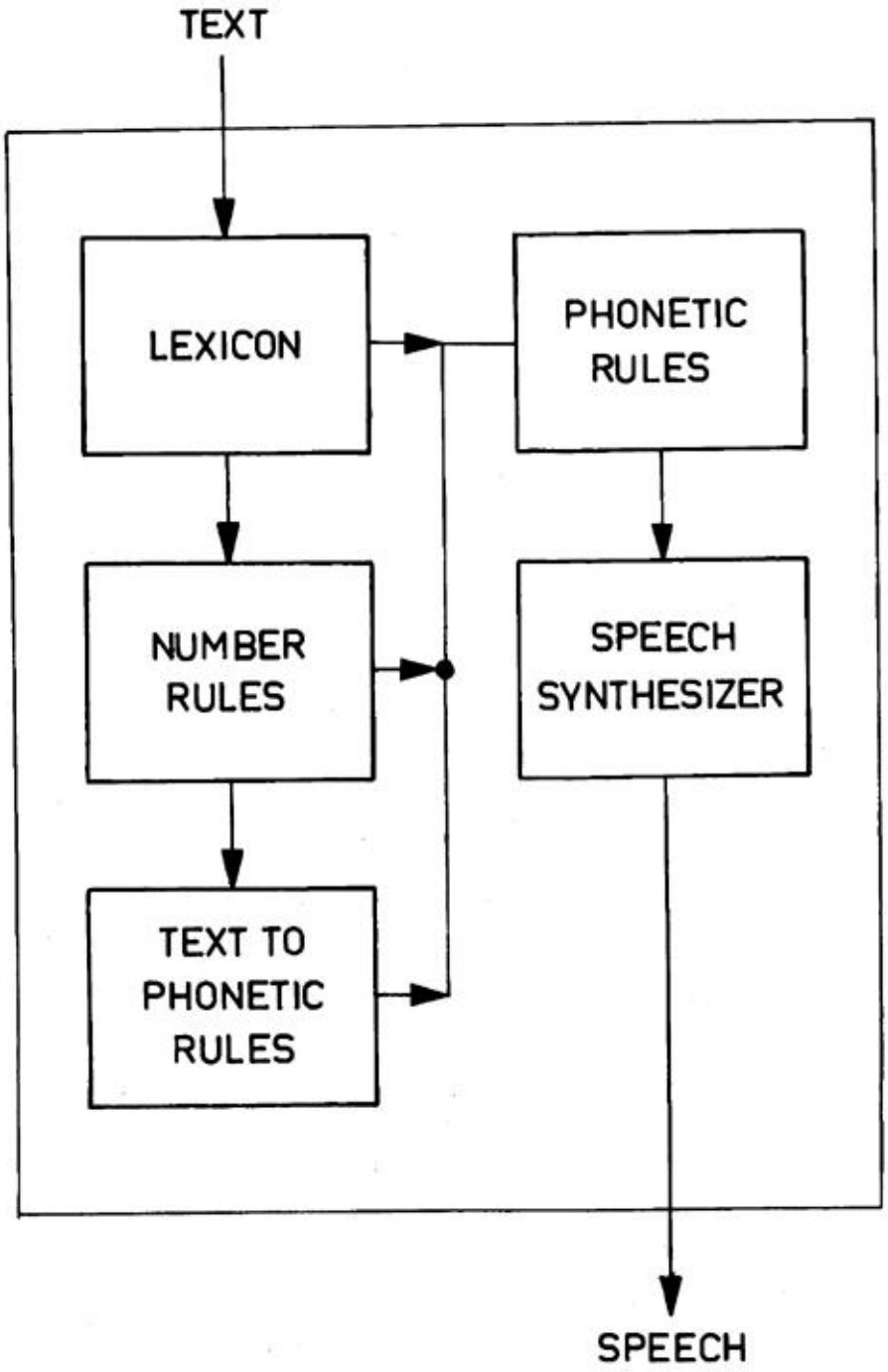

Fig. II-A-2. Structure of the text-to-speech system

operates on the "word" level, as do the earlier two. The output contains a full phonetic representation of the word, including information on syllable stress and accent. Phonological rules that operate across word boundaries are included in the next component.

Phonetic Rules

At this stage in the process, the results of the three earlier components are merged and processed on a sentence basis. These rules make heavy use of the parameter manipulation capability of the rule notation. Phonetic realizations of the segments are adjusted dependent on context both within words and across word boundaries. Prosodic realizations including the fundamental frequency contour and the durational structure of the sentence are also calculated.

Speech Synthesizer

Typically each 10 msec a frame of synthesis parameters are sent to the digital speech synthesizer. The synthesizer is a version of the OVE III synthesizers (5). It is a combined parallel/ cascaded filter structure with the possibility of mixed voiced/unvoiced excitation. A new feature is the dynamically variable "higher-pole correction", which makes it possible to model speech sounds produced with vocal tracts of different lengths.

System Development

Rule systems have been developed for several languages. The rule development has been done on our laboratory computer, a Data General Eclipse mini-computer. To support the rule development, there are programs for speech analysis and programs that collect statistics on the application of rules on different text material.

Computerized pronunciation dictionaries of frequency-ordered words have proved to be very valuable in rule development. Programs have been written that compare the result of the rules to this dictionary. The errors are listed along with statistics on error frequency and error type.

Rules in the phonetic component can be interactively developed and optimized by connecting a "joy-stick" to a variable within any rule. This variable need not be an output parameter but could correspond to another variable such as amount of reduction.

The developed rules and lexicons can be compiled into a form that

is directly accessible by the special text-to-speech hardware that is described below. The data defining a particular language are programmed into EPROM at the development system and moved to the text-to-speech module. A language typically occupies about six 64K memory chips.

Text-to-Speech Hardware

The present text-to-speech system consists of a formant speech synthesizer implemented on a signal processing chip, a powerful microcomputer based on the MC68000 16-bit microprocessor and a variety of text input equipment. In the basic configuration, the speech module consists of four euro-cards: the processor card with two serial I/O ports , two memory cards, which accommodate byte-wide memory chips of different types and sizes, and a syntheziser card with the signal processor, timer, audio filters and amplifier. The microcomputer box also contains rechargeable batteries, a loudspeaker, and different kinds of controls.

This box, connected to any kind of text source such as a conventional computer terminal, operates as a general text-to-speech system. To make the system fully portable, we have constructed a special "lid" to the computer box that contains a low profile keyboard and a 16-character liquid crystal display. The text-to-speech system attachments include enlarged keyboards and a 500-symbol Bliss board which will be described in another paper presented to this conference (6).

General Software Features

The actual structure implemented in the microcomputer differs somewhat from the base structure described, as can be seen in fig. II-A-3. By implementing an early branching possibility the text is now allowed to contain phonetic strings that are routed directly to the phonetic rules. These strings are delimited by special characters, chosen not to interfere with the ordinary text. This feature offers a way to correct mispronunciations that are especially disturbing in certain applications.

A user-programmable lexicon is also included. The lexical entries can contain both ordinary text and phonetic spellings or a mixture of the two. The text found in the user lexicon is handled in the same way as input text, i.e., recursions in the lexicon are possible. This makes

it easy to create very efficient individual abbreviation systems. Non-vocal persons can use this for fast message generation. The lexicon information is kept in CMOS RAM and is protected during power-down by a small on-board battery.

In Fig. II-A-3, an optional rule component preceding the phonetic rules can be seen. This Bliss rule component handles grammatical adjustments when the input is a sequence of codes for picture-like symbols. This special application is described in another contribution to this conference. It illustrates a case when the input information is something else than ordinary text. A related situation is when the text-to-speech system is connected to message-generating programs such as for information retrieval or question answering. In this case the program "knows" about sentence syntax, contrastive stress etc. This information can be used to improve speech quality if passed on to the text-to-speech module.

The text-to-speech module also contains programs that makes it easy to adapt to different applications. It operates in three basic modes: spelling, word by word, and sentence mode. An additional mode allows a combination of Bliss symbol and ordinary alphanumeric input.

The talking speed is adjustable both by a knob and by keyboard commands. In sentence mode, continuous reading at above 250 wpm (words per minute) is possible.

Besides the ordinary keyboard connection facility, there is the possibility of connecting to a host computer. This is necessary in the talking terminal application, but is also useful in other contexts.

Keyboard commands to the system consist of a command prefix plus a command character. Some of the implemented commands are:

- Change to spell mode
- Change to word mode
- Change to sentence mode
- Stop output from synthesizer
- Continue output from synthesizer
- Slow down speech
- Speed up speech
- Save last sentence
- Retrieve a saved sentence
- Reinitiate the program
- Write to a host computer
- Enter the user-lexicon editor
- Select an alternative language

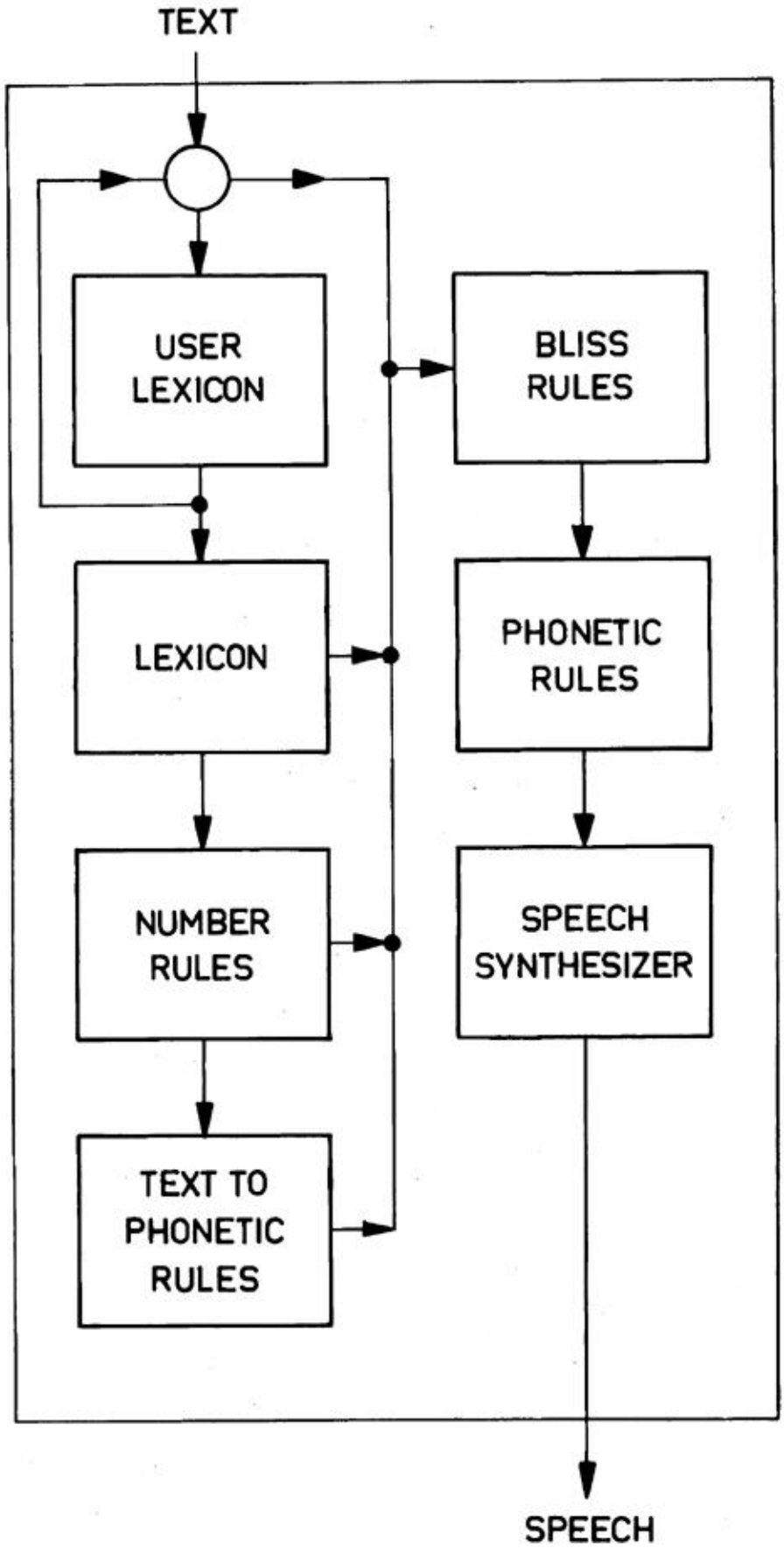

Fig. II-A-3. An expanded version of the text-to-speech system implemented on the microcomputer.

The text-to-speech system is programmed for several languages. Versions now exist for Swedish, English, Spanish, French, German, Danish, Finnish and Chinese. The rules for some of the languages are still rather preliminary. The quality of all the systems will increase as more knowledge becomes available.

Applications for the Blind

The generality of the text-to-speech system module makes it possible to use it in different kinds of aids for the blind. Some applications have already been explored.

Automatic "Talking Book" Production

In cooperation with the Swedish Federation of the Visually Handicapped (SRF), we use the system to record different kinds of text material on audio tapes. The text is supplied from ordinary casette tapes which are used for text editing and printing of Braille books at the SRF printing office. In the future, the blind person will be able to choose which kind of media he wants the text to be presented in: Braille or synthetic speech.

Talking Terminal

In cooperation with the Swedish Institute for the Handicapped, a special working place for a blind professional has been developed. It is equipped with a personal computer, a talking terminal, a Braille display and printer. The goal is to gain experience with how such equipment should look and what facilities are needed. The computer is used for both programming and editing of text.

Talking Braille Recorders

The recent development of Braille recorders with editing and search capabilities has created a possibility for blind persons to interactively work with great volumes of text. Speech is in many cases a more convenient and faster mode of presentation than Braille. Giving this kind of equipment a speech output option has greatly enhanced its usefulness.

New kinds of information systems, such as TV-text, Viewdata and electronic mail, will be directly accessible by blind users through an appropriate connection to a text-to-speech system.

Applications for the Non-Vocal

The previous prototype was used extensively by non-vocal persons in both communication and language training. The small size and the possibility to run on batteries has made the present module useful as a personal speech prosthesis.

The main problem now seems to be typing of text, since most non-vocal persons have additional motor disabilities. Different keyboards, such as a Canon Communicator and enlarged keyboards, have already been interfaced to the system.

Communication speed can be increased by storing whole phrases as 2-3 letter codes in the user programmable lexicon.

For the non-vocal that do not easily handle written language, an extension of the program has been developed. By indicating concepts on a board with Bliss symbols it is possible to produce well-formed spoken and/or written sentences (6).

Final Remarks

The text-to-speech module presented here is still not in regular production. We feel, however, that practice with the system in many different applications is of primary importance. The experiences gained are necessary for a good specification of the final product. Some of the features of the present system are worth mentioning. The modularity has made it easy to adjust to different needs. Using state-of-the-art technology makes the system small and relatively inexpensive. The choice of a powerful processor makes future expansions of the text-to-speech program easy to accommodate without sacrificing speech rate.

Text-to-speech technology will, with improved quality, spread to more applications in everyday life. This will promote greater production volumes and lower costs, and increase the availability also for the disabled.

Acknowledgements

Björn Larsson and Lennart Neovius are gratefully acknowledged for competently realizing the hardware part of the text-to-speech module.

References

(1) Carlson, R. & Granström, B. (1978): "Experimental text-to-speech system for the handicapped", J.Acoust.Soc.Am. 64, S163.

(2) Carlson, R., Galyas, K., Granström, B., Pettersson, M. & Zachrisson, G. (1980): "Speech synthesis for the non-vocal in training and communication", STL-QPSR 4/1980.

(3) Allén, S. (1970): Nusvensk frekvensordbok, 1 (Frequency Dictionary of Present-Day Swedish), Almqvist & Wiksell, Stockholm, Sweden.

(4) Carlson, R. & Granström, B. (1976): "A text-to-speech system based entirely on rules", Conference Record, 1976 IEEE-ICASSP, Philadelphia, PA, USA.

(5) Liljencrants, J. (1968): "The OVE III Speech Synthesizer", IEEE Trans. on Audio and Electroacoustics, AU-16, March.

(6) Carlson, R., Granström, B., & Hunnicutt, S. (1982): "Bliss communication with speech or text output", to be publ. in the Conf. Record 1982 IEEE-ICASSP, Paris, France.