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

Speech synthesis and recognition in technical aids

Blomberg, M. and Carlson, R. and Elenius, K. O. E. and Galyas, K. and Granström, B. and Hunnicutt, S. and Neovius, L.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 27
number: 4
year: 1986
pages: 045-056

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

B. SPEECH SYNTHESIS AND RECOGNITION IN TECHNICAL AIDS

Mats Blomberg, Rolf Carlson, Kjell Elenius, Karoly Galyas, Björn Granström, Sheri Hunnicutt, & Lennart Neovius

Summarized by Sheri Hunnicutt

Abstract

A number of speech-producing technical aids are now available for use by disabled individuals. One system which produces synthetic speech is described and its application in technical aids discussed. These applications include a communication aid, a symbol-to-speech system, talking terminals and a daily newspaper. A pattern-matching speech recognition system is also described and its future in the area of technical aids discussed.

Introduction

Because speech is the most natural way for people to communicate with each other, there has been a great deal of interest in the use of synthetic speech and stored or concatenated speech in technical aids for blind and nonvocal individuals. A number of devices are now on the market, and their use as technical aids is becoming more widespread. Aids involving speech recognition are only beginning to be available.

Speech synthesis and concatenated speech systems have unlimited vocabulary, and are more expensive than stored speech systems. Users find the higher quality of good synthetic speech less demanding to listen to than concatenated speech, particularly for long periods of time. A new language can, however, require a great deal of expertise and time to produce. Much research has been going on over a period of years to produce better female and child speech, but this problem is not yet solved. An important part of the solution lies in results of research in modelling the glottal source.

A Multi-Lingual Text-to-Speech System

The synthetic speech system developed at the Royal Institute of Technology in Stockholm is a text-to-speech system, accepting any text input (Carlson, Granström, & Hunnicutt, 1982). It presently exists in nine languages (or dialects): Swedish, American English, British English, Spanish, German, Norwegian, Danish, French and Italian. Sections of the system exist as separate components that are connected in the desired manner by a supervisory program. Fig. 1 shows the basic configuration.

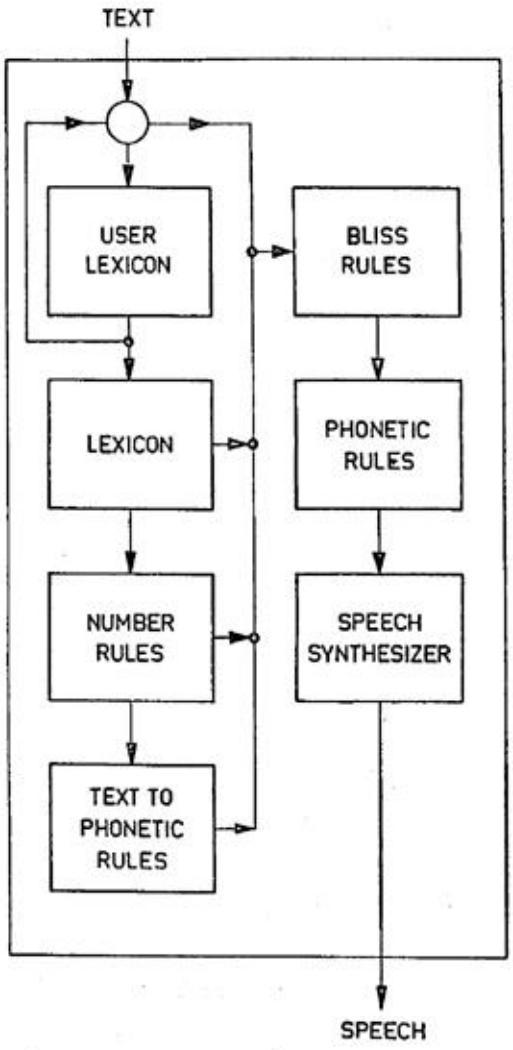

Fig. 1. Text-to-speech: System Diagram

User Lexicon

Input text is first examined for phonetic strings which are identified by special delimiters and sent directly to the phonetic rules. The text is then examined for the occurrence of words in the User Lexicon which contains special pronunciations desired by the individual user. These can be words or names which would otherwise be pronounced incorrectly or inappropriately, and can also be user-defined abbreviations. Since recursion is allowed in this lexicon, efficient abbreviation systems can be created by the user.

Main Lexicon

The text is then inspected for occurrence of words appearing in the main lexicon. This lexicon serves two purposes. Firstly, it contains the pronunciation for irregular words -- words that would be incorrectly pronounced by the grapheme-to-phoneme rules. The lexicon's size is

language-dependent, and can range from a few hundred up to several thousand words. Secondly, it serves a grammatical purpose. It identifies "function" words, frequent words such as determiners and prepositions which usually are not stressed in speech. This identification permits a grouping of words into phrase-like units that make the resulting speech more natural.

Number Rules

If numbers or other non-alphabetic characters occur in the text, they are processed by the number rule component. In this component, numbers are converted to words, and special attention is paid to correct pronunciation for amounts of money and years.

Grapheme to Phoneme Rules

All text, with the exceptions noted above, is then converted to a string of phonemes and stress marks using a set of rules. These rules make use of language-dependent phoneme definitions which include binary feature specifications. The size of this set is language-dependent. For English and Swedish, several hundred rules are needed. For Spanish, about fifty rules will suffice due to the close relationship between spelling and pronunciation.

Phonetic Rules

The resulting stress-marked phoneme strings from the processes already discussed are adjusted allophonically in the phonetic rules. These rules make use of the phoneme definitions which include specifications of formant frequencies, amount of aspiration, and transition timing. The influence of context, both within words and across word boundaries, is accounted for in the resulting phonetic realization of segments. It is this component which gives synthetic speech its superior quality of continuity since segment coarticulation can be specified in detail, both as to formant frequencies and timing of transitions.

Speech Synthesizer

When the above processing is completed for a sentence, a frame of synthesis parameters is sent to the digital speech synthesizer each 10 msec. This synthesizer has a combined parallel/cascade filter structure and can use either voiced (pulses) and/or unvoiced (noise) excitation. A special feature is the dynamically variable "higher-pole correction," which makes it possible to model speech sounds produced with vocal tracts of different lengths.

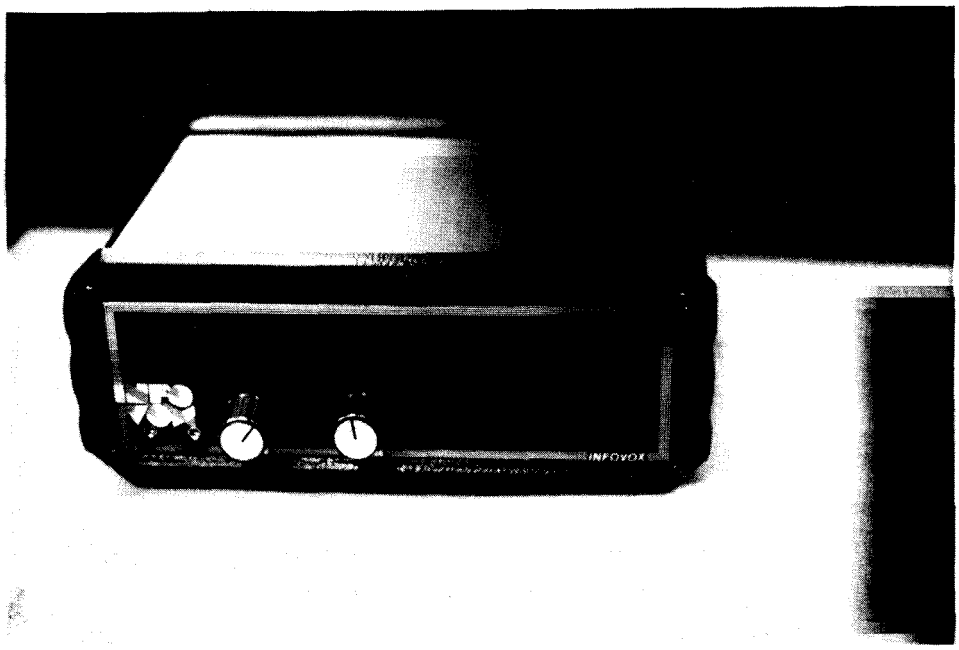

Photo 1: VoxBox: Stand-Alone Text-to-Speech System (Courtesy Infovox AB, Lennart Neovius)

Hardware Implementation

The synthesis hardware is based on a Motorola 68000 and a NEC7720 signal processing chip. Several versions are produced and marketed by a Swedish company, Infovox AB. One version is the VoxCard which is a double euro-card with two RS232C (V24) connections. VoxCard is also packaged in a stand-alone system with power supply, loudspeaker and function controls. This system is called the VoxBox and is shown in Photo 1. It can be connected to any conventional terminal or computer output. Another version of this system is designed for an IBM-compatible PC (personal computer). It is delivered with a loudspeaker and software for the desired languages. (See Photo 2.) The software includes separate library routines for easy interface to Pascal and C programs.

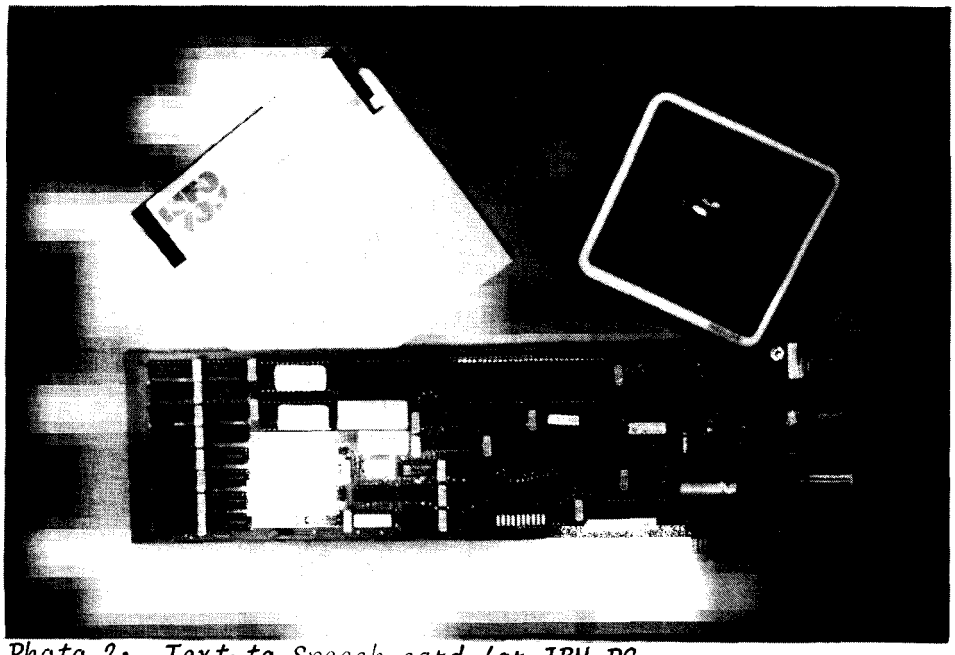

Photo 2: Text-to-Speech card for IBM-PC (Courtesy Infovox AB, Lennart Neovius)

Functions of the Text-to-Speech Device

The system contains programs that allow it to be adapted to different applications, e.g., to be used as a talking terminal. It operates in three modes: normal sentence mode, word by word mode, and spelled word mode. Reading speed is variable, and this speed is adjustable at any time. It is possible to connect to either a terminal or a host computer. Keyboard Commands are given to the system via a command prefix (ESC) plus a command character. Some of the implemented commands are:

- Change reading mode: spelling, word, line, sentence
- Change language: English, Spanish,French, German, etc
- Stop/continue output from synthesizer
- Change voice parameters: pitch level, dynamic range, breathiness, etc
- Store/retrieve voice type
- Change speech tempo
- Save/retrieve sentence
- Enter the user lexicon editor
- Index text
- Generate a tone
- Inspect phonetic text

Word Predictor

An existing function which has not yet been added to the commercial device is a word predictor, a program which predicts a word based on its first letter or letters (Hunnicutt,1986a). The predicted word may be accepted, or it may be rejected simply by typing the next letter of the word. A further prediction is then made from the letters that have been typed. Predictions are based on the frequency of the word in the language according to a large accompanying lexicon, and on their recency of usage. Words the user types in are stored and used to update the lexicon in content and word frequency. Such a function could usefully aid persons who type very slowly.

Application in Technical Aids

Text-to-Speech in Voice Prosthesis

The first use of the Swedish text-to-speech system as a communication aid was in 1978. This early system was based on a minicomputer and could be moved around on wheels. A teenager, diagnosed as suffering from cerebral palsy, was the system's first user, and typed with a mouthstick. The results during this last year he was in school were promising, and inspired a continuation of this type of application (Carlson & al, 1980).

In the following years, when the new systems containing microcomputers with signal processing chips became available, a number of other persons with various needs began to use them for verbal communication (see Photo 3). Some examples follow (Schildt & Sterner, 1986):

Photo 3: The VoxBox as part of a Communication System (Courtesy Swedish Institute for the Handicapped, Nestor Peixota Noya)

One man, who had both a laryngectomy and a glossectomy used the system quite flexibly. He typed quickly, and, particularly with the abbreviations possible in the user lexicon, could carry on a conversation in near normal time. He found speech a much more natural way to communicate than writing, and used it in conversations with his family. Telephone conversations also became a possibility once again. This seems to be a very important use for others, as well, since it gives a great deal more independence and privacy. Another man, left without speech after a stroke, began to use the system while still in a hospital. He later moved to a service apartment where he lives independently, having the help of an electric wheelchair, a personal computer and his speech synthesis device.

The device has also been used rather much for training reading and writing, particularly for practicing spelling. One non-speaking boy who seemed not to be able to learn to write more than a few short words, was able to learn better with the speech synthesis device. Using a "sounding out" mode, he listened to the sound of each letter until he could begin to write words. Another child, who has a brain injury and autism, was able to transfer his experience with a communicator to the synthesizer. He was further motivated by the device and greatly enjoyed hearing words with particular meaning for him. A group of non-speaking students has used it for reading a theatre play together -- the play was typed in

a file in the computer. When a particular actor's turn came, only his/her keyboard would trigger the reading of that part.

Multi-Talk

The speech synthesis device has recently been packaged in an attaché case as a special-purpose communication aid called Multi-Talk (Galyas, 1986). It is being produced by another Swedish firm, Fonema AB. To use it, one simply lifts the attaché case lid, turns on the device, and begins to type on the Epson keyboard. It runs on either rechargeable batteries or mains power. Multi-Talk comes equipped with up to four of the available languages; a language can be chosen by simply pushing a function key (see Photo 4).

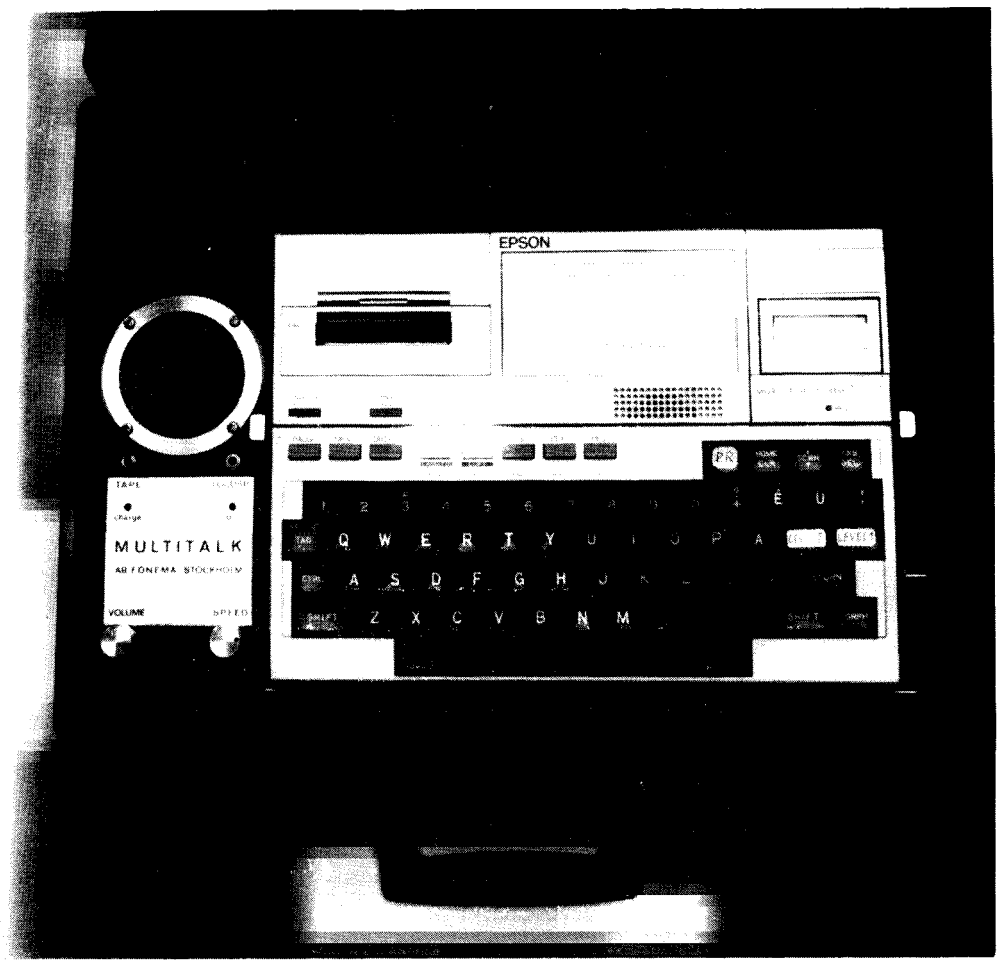

Photo 4: Multi-Talk (Karoly Galyas)

In addition to the usual text-to-speech features such as abbreviation and control of voice quality, Multi-Talk includes several features specially designed to aid communication. There is a contrast-adjustable screen to see what is being written, and a function key to hear what has already been written in the current sentence. The last sentence can be repeated, even in the middle of the following sentence, and can be

repeated word by word if desired. Any word can even be spelled out if it has not been understood. The printer included in the Epson can always be used to print out the text which is visible on the screen. In the near future, a word prediction algorithm will also be included.

The speed of communication can be increased very much by the availability of two "higher levels." One higher level allows the user to access stored messages with any single key. This level can be accessed for saying a message without disturbing work in progress at the base level. The other higher level also allows the user single-key access to messages. These messages, however, can be completed with further typed text, or can be copied into the sentence in progress on the base level.

Blisstalk

Another device containing speech synthesis, which was first built and tested about five years ago, is an electronic communication board called Blisstalk (Hunnicutt, 1986b). It is now produced by a Swedish company, Rehabmodul AB. (See Photo 5). On it are up to 500 Blissymbols which are selected by a magnet or by scanning. The board can be reprogrammed with any of 1400 available symbols: a few large symbols may be chosen for a beginner, and more can be added as the user progresses. Each symbol is represented in a lexicon by one or two corresponding words, their pronunciation and grammatical category. Some symbols have grammatical functions themselves, e.g., plural, verb tenses, possessive. Blisstalk also contains a special phrase structure grammar which takes account of word order, phrase order and grammatical information in the lexicon to produce well-formed sentences.

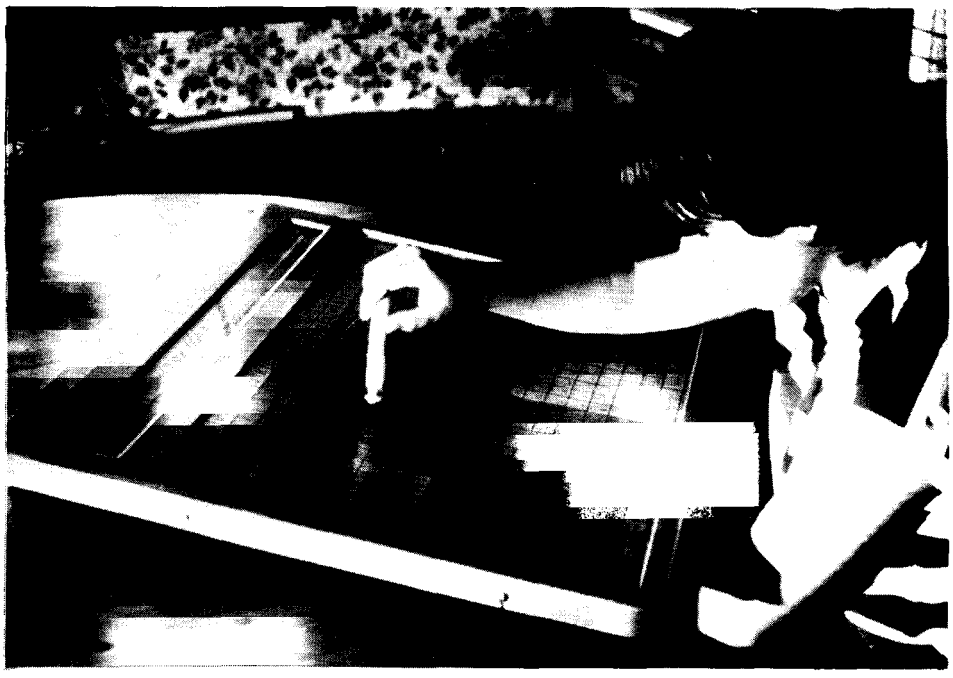

Photo 5: Blisstalk (Courtesy Swedish Institute for the Handicapped)

Besides this sentence mode are also a word-by-word mode, which does not access the grammar, and a character mode for pronouncing numeral and letter names. A sentence or other completed expression may be repeated, and up to 10 sentences may also be temporarily stored and quickly retrieved. The letters may be used to supplement the symbols by spelling out words, just as in the usual text-to-speech system.

Talking Terminals

The most widespread use of the speech synthesis system as a technical aid at this time is as a "talking terminal." In this application, information which is printed on a computer screen is read by the device. About 300 systems have already been installed as talking terminals. This technique is most used by blind persons and persons with low vision, and has also been implemented in a number of work stations in Sweden. These work stations are typically built around a personal computer, and include a Braille display and printer as well as a speech synthesis device. The results have been quite promising, particularly in several office applications such as word processing, register handling and local switchboard operation.

There are also several distributors of screen-reading programs, for such terminals as the IBM-PC and VT100. These programs detect certain commands from normal text input such as "Read current line," "Give cursor position" or "Read word by word" which are interpreted in special routines. These routines access the appropriate text and send it to the synthesis device to be read.

Daily Newspapers

Another application of speech synthesis for the visually handicapped is in the area of reading text which has been typeset by computer -- a common practice in printing offices nowadays. A project which has continued for several years in Sweden is to make daily newspapers available to persons with visual disabilities (Rubinstein, 1984; Carlson & Granström, 1986). At present, newspaper text is broadcast digitally to the homes of about 30 blind subscribers where it is stored on a magnetic disk during the night. The user can then, at his leisure, search the material for sections, headlines, or particular words with the help of a small microcomputer. The text can then be presented as synthetic speech. (See Photo 6.)

Speech Recognition

Although it has not been the major emphasis of this discussion, an area which will be increasingly useful in technical aids in the future also deserves to be mentioned. This is speech recognition. The system

Photo 6: Daily Newspaper via Text-to-Speech (Courtesy Chalmers Institute of Technology, Henryk Rubinstein)

which was developed at the Dept. of Speech Communication and Music Acoustics is a pattern-matching system (Elenius & Blomberg, 1986). It is also available from Infovox AB in the IBM-PC compatible form shown in Photo 7. The system digitally implements a 16-channel filter bank. This filter bank covers frequencies from 200 to 5000 Hertz in bands spaced according to the critical band scale which represents the frequency characteristics of the human auditory system. Thirty-two sample points derived from this information are matched with the stored reference patterns by dynamic programming time alignment.

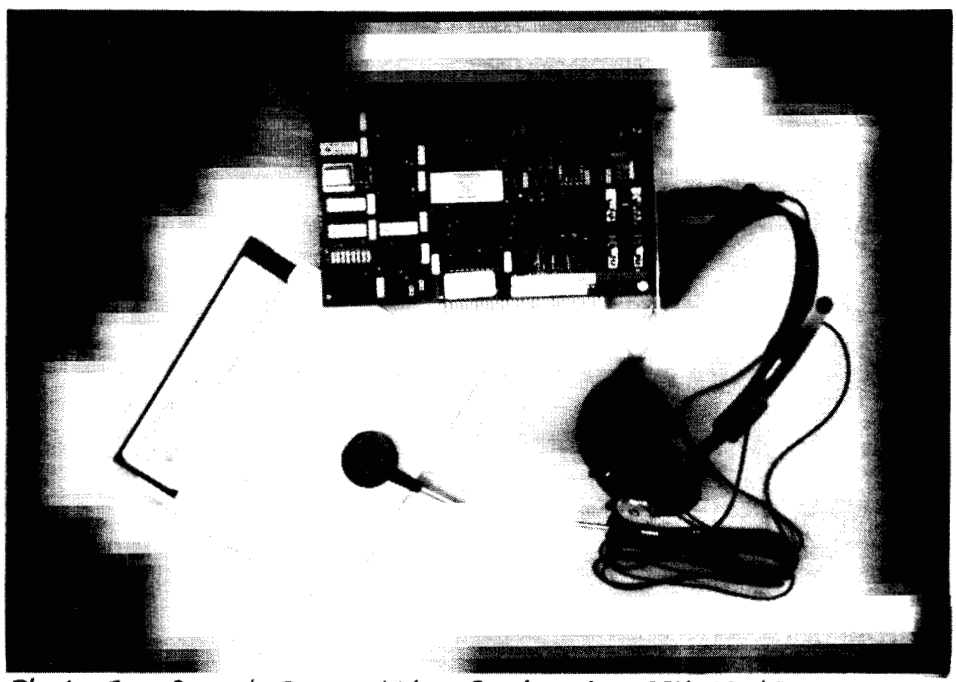

Photo 7: Speech Recognition Device for IBM-PC (Courtesy Infovox AB, Lennart Neovius)

Speech analysis and dynamic programming are accomplished using a NEC7720 signal-processing chip. Control of the recognition process and storage of the reference vocabulary are handled in the microprocessor and memory of the PC.

There are several ways in which this device could be used as a technical aid. It can, for example, be used to voice-control another unit connected to an I/O-port of the PC. Such a use would be voice control of a device for environmental control. It may also be used to add speech control to any already existing program. After speech input is initialized, the response strings of recognized utterances look like keyboard entries to the program to be run. A motorically disabled person capable of producing different (but consistent) utterances for each key on the keyboard would thereby have voice control of any user program. One particularly useful application is word processing in which utterances would access both keys and editing commands. Another use would be to integrate speech control into an applications program with calls to the recognizer's special functions. This could be an especially useful tool for a disabled programmer.

Conclusions

It is now possible for both speech synthesis and recognition to be used in technical aids for disabled persons. A text-to-speech system, and a pattern-matching speech recognition system, developed at the Dept. of Speech Communication and Music Acoustics, KTH, Stockholm were presented and the current use of the synthesizer in technical aids described. Included were Multi-Talk, a specialized communication aid, Blisstalk, a Blissymbol-to-speech system, talking terminals, and a daily newspaper source for the blind. The future of speech recognition in technical aids was discussed.

References

R. Carlson, K. Galyas, B. Granström, M. Pettersson & G.Zachrisson (1980): "Speech synthesis for the non-vocal in training and communication," STL-QPSR 4/1980, pp. 13-27.

R. Carlson, B. Granström & S. Hunnicutt (1982): "A multi-language text-to-speech module," pp. 1604-1607 in Proc. ICASSP 82, Vol. 3, Paris.

R. Carlson & B. Granström (1986): "Applications of a multi-lingual text-to-speech system for the visually impaired," pp. 7-96 in (P.L. Emiliani, Ed.): Development of Electronic Aids for the Visually Impaired, Martinus Nijhoff/Dr W. Junk Publishers, Dordrecht.

K. Elenius & M. Blomberg (1986): "Voice input for personal computers," pp. 361-372 in (G. Bristow, Ed.): Electronic Speech Recognition, Collins Professional and Technical Books, London.

K. Galyas (1986): "Talande hjälp för den talskadade," pp. 148-153 in (J. Allwood, Ed.): Mänsklig kommunikation, GULING 14, Dept. of Linguistics, University of Gothenburg.

S. Hunnicutt (1986a): "Lexical Prediction for a Text-to-Speech System," pp. 253-263 in (E. Hjelmquist & L-G. Nilsson, Eds.): Communication and Handicap: Aspects of Psychological Compensation and Technical Aids, Elsevier Science Publishers B.V., North-Holland.

S. Hunnicutt (1986b): "Bliss Symbol-to-Speech Conversion: 'Blisstalk'," J. of the Am. voice I/O Society, 3. pp. 19-38.

H. Rubinstein (1984): "FM Transmission of Digitalized Daily Newspaper for Blind People," Reprint from IEEE Communication Society Global Telecommunications Conference, Atlanta, GA, USA.

A. Schildt & M. Sterner (1986): Talsyntes som Talhjälpmedel, Handikappinstitutet, Bromma.