A SYNTHETIC SPEECH TERMINAL FOR VIATEL: DESIGN, IMPLEMENTATION AND PERFORMANCE

R. W. King and A.J. Hunt

School of Electrical Engineering and Computer Science
University of New South Wales

ABSTRACT - Speech synthesis technology has been incorporated successfully into several computer systems for use by blind people. There has, however, been relatively little attention paid to the specific problems of using visually-conceived information services such as videotex (of which Australia's *Viatel* service is an example) with synthesised speech output. In this implementation of a PC-based prototype 'talking-videotex' terminal, page layout processing and comprehensive user controls are provided to overcome the problems of page scanning. The terminal incorporates a low-cost SSI-263 synthesiser chip, with software to produce an Australianised accent, together with word-based prosody.

INTRODUCTION

Speech technology research has led to the commercial development of a number of low-cost, unlimited-vocabulary synthesiser devices and modules. Many of these have been incorporated in units aimed specifically at providing voice output for computer systems for use by blind people, and are available either as plug-in units for personal computers or as integral parts of special purpose terminals. Blind people are using these systems for word processing, record-keeping, and computer programming with great success and satisfaction .

It is evident however, that providing synthesised speech output for interactive and 'strongly visual' computer software may pose particular problems to blind users. If the information 'screens' have unfamiliar layouts, or include features which are not readily converted into speech, the effectiveness and efficiency of the user-system interaction are reduced. Such problems may increase with new computer packages, which often possess considerable sophistication and visual diversity in their screen layouts, by the use of graphics, windowing, and pop-up menus. Similarly, information held in videotex data-bases which may be accessed via the public telephone network, such as *Viatel*, is displayed in a variety of formats, and is invariably designed for a screen-keyboard interactive dialogue.

There are two strategies which may be employed to interface synthesised speech output to such screen-based systems. One is to provide an efficient linkage between a screen scanning and pointing mechanism and the synthesised speech output. A second strategy, expected to be more efficient, is to include an on-line layout analyser between the source screen and the synthesiser in order to present the information in a sensible order. This approach was used in a previous investigation of the design of non-visual (i.e. dynamic Braille and synthesised speech) videotex terminals (King, Cope & Omotayo, 1985). In that study it was found that the average time taken to accomplish information seeking tasks on the *U.K. Prestel* videotex system using synthesised speech output *(SC-01)* was between 4 and 6 times that taken using visual interaction.

A more effective talking-videotex system should be obtained through improvements in the three functional components of the earlier system: layout processing, speech synthesis, and user control of the synthesiser utterances and data-base interaction. This paper outlines the design considerations, implementation and performance of a new personal-computer based prototype terminal. This incorporates a low-cost speech synthesiser unit using the *SSI-263 (SC-02)* synthesiser chip and text-to-speech software designed to produce an 'Australianised' accent and simple word-based prosody. The paper includes a brief discussion of the user controls and the overall performance of the terminal.

SYSTEM DESIGN

The essence of this work is the transformation of information between display modalities. It is instructive, therefore, to review the technical characteristics and human factors aspects of the videotex data base and synthesised speech, before discussing details of the design.

Data-bases and Videotex

In general terms any information data-base may be regarded as having four dimensions, two for the layout of 'screens' or individual pages, the third being the character attributes, such as colour and size, and the fourth being the organisation of the data-base page structure. In these terms, therefore, the normal colour video display screen used for access to the data-base provides a three-dimensional window through which to view the data-base. The page layout and use of colour should help the user to perceive the required information readily, and a suitable display-keyboard interaction should facilitate the user's penetration of the data-base structure through a menu, index or keyboard dialogue.

The videotex system data-base is divided into pages of 24 rows of 40 characters per row for transmission and display. The normal viewing window corresponds to one page. Most of the information is presented as alphanumeric characters, but colours and graphics and attention to layout design are designed to make the pages visually attractive, and easier to read. Page access and reading strategy are under the complete control of the user. 'Interactive' or response pages support message services and transactions of various kinds. Illustrations of typical *Viatel* pages are given in Figures 1 and 2, and it should be observed that the material on these pages is <u>not</u> intended to be read line by line from the top to the bottom. The two pages illustrated also show the importance of the indexing information on each page. Although a page may be requested by its page number, as on the header row, it is common to move through the data-base to the required pages using the indexing data displayed on the pages themselves. The transmission uses a standard 1200/75 baud protocol, and hardware required for videotex access is available either as a plug-in card for many personal computers, or as a separate module to interface via RS232 to a computer generated display.

Synthesised Speech for Data-base Output

Speech output synthesised from unknown text may be considered to be a window with only 'one-and-a-bit' dimensions. The primary dimension of an utterance is its word content, the secondary dimension is its prosody. An effective synthesised output for a data-base must include provision for scanning the two dimensions of each page, and producing sensibly sized utterances. Page layout analysis and character attributes may facilitate prosody synthesis. The user must be able to access the fourth dimension of the data-base via an appropriate interaction dialogue, normally the keyboard.

The intelligibility of low-cost, unlimited vocabulary synthesised speech has improved considerably in recent years, but the speech quality is still far from natural. Nevertheless, the overall performance of low-cost devices and units is sufficient for applications such as voice prostheses and aids for blind people. It is known that intelligibility is reduced in the absence of correct prosody (Waterworth & Thomas, 1985), but automated techniques for accurate prosody insertion are relatively undeveloped (McAllister, Laver & McAllister, 1986). A second aspect of naturalness and intelligibility is that of accent. Most low-cost synthesis systems devices have relatively little accent control, although many offer pitch and speed control to the user. In the case of the *SSI-263* speech synthesiser used in many low-cost products, an approximately Australian accent may be produced by appropriate changes to text-to-speech software (Mannell, 1986).

Hardware and Software

The prototype PC-based talking videotex system uses a standard PC-modem card for videotex access, and an *SSI-263* speech synthesiser unit connected to the computer's parallel printer port. Computer software generates the required synthesiser register address and data codes in multiplexed form; the synthesiser unit therefore contains only demultiplexing hardware, the speech chip and an audio amplifier. This implementation of the hardware interface allows physical connection of the

synthesiser unit to any microcomputer with a Centronics printer port, and customising the port driver software.

The software is written in C, except for the assembler interrupt handler used for videotex data acquisition. The software design falls naturally into two independent parts: page data acquisition and processing, and speech synthesis. Clearly these two parts may be developed and tested evaluated separately. In addition, the control function provides the user with utterance control, and interaction with the layout processor and the videotex system. The following sections describe the main functional features of the software.

PAGE DATA ACQUISITION AND LAYOUT PROCESSING

Acquisition and layout processing software is required to identify the type of the received videotex data, and then process it using its layout features and character attributes. The page elements are ordered into a correct utterance sequence, and transferred to the output speech buffer.

Every request to videotex results in data being loaded into the page buffer. The contents of the received data stream, considered in conjunction with the current page status, defines the identity of the new data. It may be a new *normal* page, a new *interactive* page, the *echo* of a response to an interactive page prompt or a *system message* displayed on row 24.

The action of the layout processor depends on the current page status and the identity of the new data. System messages and response echos are passed directly to the output buffer. All new pages are stripped of non-utterable characters, such as graphics and other page-wide dividers formed from punctuation characters. The layout of new interactive pages are not processed further; it has been found to be satisfactory to output them on a row by row basis, up to the response prompt. Normal pages - which arise from most page requests - are analysed by locating the visual clues in the displayed page in order to divide up, classify and order the page elements. The problem is treated as a syntactic pattern recognition task.

The first stage is to obtain a *type* image of the page in which the utterable characters are classified as letters, numbers, and punctuation. Rarely is the loss of graphics characters significant, as 'symbolic' graphics, as opposed to 'layout-division' graphics, usually conveys information included elsewhere on the page. The page is then divided into *blocks*, first on the basis of background colour and secondly from any continuous graphics features which appear to separate one area of the page from another. The blocks are then considered in sequential order from the top of the page.

The aim is to order the utterable elements within each block. Two particular page layouts have to be distinguished. *Tabular* information is best uttered as a sequence of full width elements. In contrast, many pages contain *multiple columns*. Indexing structures are commonly in this form. Both tables and multiple columns contain similar layout syntax, making the process of distinguishing between the two layout forms somewhat hazardous. However, experimental investigation with a number of strategies has shown that rules can be found to make the distinction successfully. These require analysis of rows for their potential to be a 'seed' for table or multiple column layout. The other rows in the block are then checked for their similarity with the seed line. If a table is located, there is the expectation of a table header preceding the table. Multiple column indexing structures have a regular number - text syntax, sometimes reinforced by the use of different colours for the index number and its descriptors. The index numbers in a column are expected to be in a sensible range.

As a result of this analysis, the page elements are ordered and transferred to the output buffer. Each element is terminated with an utterance boundary, which is either a period, question mark or exclamation mark. The output buffer contents is shown in Figure 3 for the example page of Figure 1. It may be observed that text continuity between rows is checked and acted upon, and some abbreviations in the header are expanded. The control symbols in the output buffer are used by the output-user control to identify the start and type of each block, and the identity of index and table elements.

This form of layout processing has been found to work reasonably successfully on a sample of pages although the table and multiple column discrimination can be incorrect on some pages. This is likely to disorder the page to a confusing extent. The user, suspecting incorrect layout processing, can request the page to be output row by row. It may be observed that this automated page analysis is trying to imitate the human visual scanning-analysis-ordering processes involved in reading. Thus, conceptually, layout processing is related to the problem of good page design (Tullis, 1983), and it may be suggested that if this layout processor in unsuccessful then the page may also be hard to read.

SPEECH SYNTHESIS

Published rules (Elovitz, et al., 1976) formed the basis of the rules used in both the earlier *SC-01* based system and the new PC-based system using the more advanced *SSI-263* synthesiser chip. The greater phoneme flexibility, pitch and duration control of the latter has allowed a degree of 'Australianisation' of the synthesiser output, through modification of the text-to-phoneme rules, broadly along the lines reported elsewhere (Mannell, 1986). In developing this software for a 30000 word dictionary considerable attention was paid to suffix rules, and relatively little to the development of an exceptions dictionary (Hunt, 1987). Initial subjective evaluation with fourteen subjects revealed approximately 70% sentence comprehension. A significant improvement is likely with longer exposure to the voice, and further work on the text-to-phoneme rules and dictionary would also improve performance. While only one of the fourteen subjects described the accent as 'Australian', no-one described it as 'American'.

Simple stress, rhythm and intonation rules have been incorporated to provide a degree of natural prosody to the synthesised utterances. The prosody computations are derived from the syllables of each word computed from the phoneme string. This is a significant limitation, since stress in particular is semantically and syntactically determined. In the absence of sentence analysis the stress assignment target has to be 'neutral' rather than definitive.

The stress rules assign *full stress, secondary stress and unstress* to multisyllablic and monosyllabic words in different ways. Stress assignments for multisyllablic words based on linguistic rules (Wijk, 1966) have been implemented in a Prolog-like rule base, and have been found to be about 70% accurate. For monosyllabic words, a list of unstressed words (Ainsworth, 1973, Gimson, 1984) forms a basis for stress assignment.

The rhythm assignment locates the sentence *feet*, generally a stressed syllable, and calculates foot boundaries for the utterance (Witten, 1977) from stressed syllables, phrase and sentence boundaries. Two parameters in the program, minimum foot length and minimum syllable length, control the actual utterance feet and syllable durations within an utterance. The process attempts to maintain a natural rhythm which approximates isochrony. The principal problem of the rhythm assignment is its dependence on the prior stress assignments.

Only four pitch contours have been included in the program. Punctuation terminators enable identification of questions and exclamations. These invoke suitable pitch contours over the corresponding phrase. Within other phrases the normal decline in pitch is modified by including rises on stressed syllables of about 5 to 10 Hz.

USER CONTROL AND OVERALL EVALUATION

The new PC-based system gives the user four levels of control. *Utterance* controls allow any element to be repeated, or uttered as separated words or spelling. *Page* controls allow the user to move around the output speech buffer, enabling the user to 'scan' the contents both forward and backwards at both the element and block level. The start of the page may be accessed directly, and a single key may be used to request output of a page in un-processed form, as described above. Identification of a table heading, and its re-utterance, is also intended to speed up the access of the user to tabular data, which was one of the major problems in the early system.

Videotex requests may be made directly at any instant using the normal form of the dialogue. In addition, a request to a page identified by an index item which has just been uttered may be made by a single key, possibly a useful reduction on the required memory load of the user. All videotex responses are prompted to the user, as the keys are pressed. In addition the system offers a number of spoken messages for user assistance. The most important of these is the indication that new page data has been received. The average time for a new videotex page to be received and layout processed is only fractionally greater than the normal videotex access time, of about 6 seconds.

The fourth level of interaction is the use of function keys to change the voice to a number of pitches and rates, and in the experimental protoype, to invoke monotonous or prosodic output for comparison purposes.

Although the results of a full evaluation of the system are not yet available, it is clear that the improved layout processing and more comprehensive set of user controls have reduced the average access time on typical information seeking tasks, compared with the earlier system. The simpler videotex interaction dialogue has also improved the overall performance.

CONCLUSIONS

This paper has presented the main features of the design and implementation of a continuing development study of the a self-contained talking videotex terminal for use by blind people. It has been argued that to provide efficient 'one-dimensional' synthesised speech output of the data-base pages requires a significant amount of layout processing, and the provision of a comprehensive set of user-controls. The paper has also presented an outline of a low-cost speech synthesis unit, capable of producing a reasonable quality Australianised accent, and word-based prosody.

The underlying issue behind this work is the effective conversion of visually conceived information to synthesised speech. There are fundamental differences between the nature of information-seeking with speech using a human to human dialogue, and obtaining information via a screen-keyboard interaction. Nevertheless, this experimental study has, it is hoped, shed some light on the requirements of the supporting framework for practical applications of synthesised speech devices.

ACKNOWLEDGEMENTS

This study has been carried out with the assistance of financial support from the Australian Telecommunications and Electronics Research Board. In addition, the authors wish to acknowledge the interest of the New South Wales Royal Blind Society and Telecom Australia.

REFERENCES

Ainsworth, W.A. (1973), *A System for Converting English Text into Speech*, IEEE Trans. on Audio and Electroacoustics, AU-21, No 3, 288-290.

Gimson, A.C. (1984), *An Introduction to the Pronunciation of English*, pp. 221-280, (Edward Arnold: Australia)

Elovitz, H.S, et al. (1976), *Letter to Sound Rules for the Automatic Translation of English Text to Phonetics*, IEEE Trans. on Acoustics, Speech and Signal Proc., ASSP-24, No 6, 446-459.

Hunt, A.J. (1987), *Australian Accented Speech Synthesis*, B.E. Thesis, U.N.S.W.

King, R.W., Cope, N. and Omotayo, O.R. (1985), *Videotex for the Blind: Design and Evaluation of Braille and Synthetic Speech Terminals*, pp 803-808, Human-Computer Interaction (ed. Shackel), (North-Holland: Amsterdam).

McAllister, M.J., Laver, J. & McAllister, M.J. (1986), *A Preprocesor Algorithm for the CSTR Text to Speech System*, Proc. 1st Australian Conf. on Speech Science and Technology, 20-25.

Mannell, R.H. (1986), *Australian English and the Votrax SC-02 Chip*, Proc. 1st Australian Conf. on Speech Science and Technology, 260-265.

Tullis, T.S.(1983), *The Formatting of Alphanumeric Displays: a Review and Analysis*, Human Factors, 25, No 6, 657-682.

Waterworth, J.A. and Thomas, C. (1985), *Why is Synthetic Speech Harder to Remember than Natural Speech*, pp. 201-206, Proc. of Human Factors in Computing Systems II, (North-Holland: Amsterdam).

Wijk, A. (1966), *Rules of Pronunciation for the English Language*, (Oxford U.P.: London).

Witten, I.H. (1977), *A Flexible Scheme for Assigning Timing and Pitch to Synthetic Speech*, Language and Speech, 20, Part 3, 240-260.

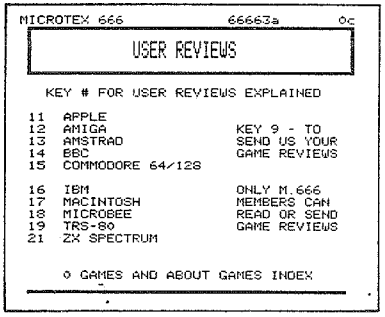

Figure 1. A typical *Viatel* page showing use of double-height characters, and a two column layout.

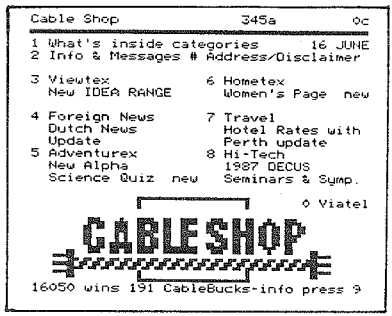

Figure 2. A typical *Viatel* page showing use of graphics to form a redundant title, and a two column index layout.

```
┐1└MICROTEX 666 page number 66663a 0 cents.┘
┐1└USER REVIEWS.┘
└0┬KEY # FOR USER REVIEWS EXPLAINED.┘
└1┬┤11├11 APPLE.
 ├12├12 AMIGA.
 ├13├13 AMSTRAD.
 ├14├14 BBC.
 ├15├15 COMMODORE 64/128.
 ├16├16 IBM.
 ├17├17 MACINTOSH.
 ├18├18 MICROBEE.
 ├19├19 TRS-80.
 ├21├21 ZX SPECTRUM.┘
└2┬KEY 9 - TO SEND US YOUR GAME REVIEWS.
ONLY M.666  MEMBERS CAN READ OR SEND GAME REVIEWS.┘
┐1└┤0├0 GAMES AND ABOUT GAMES INDEX.┘
```

```
key to control symbols:
     ┐ start of block          └ end of block descriptor
     └ start of sub-block      ┬ end of sub-block descriptor
     ├ start of index number   ┤ end of index number
     ┘ end of block
```

Figure 3. The contents of the output speech buffer for the page illustrated in Figure 1, showing control symbols, expansions of header line components and row continuity.