DEVELOPMENT OF A MICROPROCESSOR-BASED SPEECH RECOGNITION SYSTEM FOR REMOTELY OPERATED UNDERWATER VEHICLE

C.C. Fung, C. Romeo and A. Gregory

School of Electrical and Computer Engineering
Curtin University of Technology

ABSTRACT - Development of a microprocessor-based speech recognition system for a low cost remotely operated underwater vehicle (ROV) is reported in this paper. The system is designed to enhance the operator-controls for the TITAN range of ROV's. The experimental system is developed on a Motorola evaluation board with a M68HC16 16-bit microcontroller. Recognition is based on the minimum Euclidean distance between the energy functions of sampled templates. The algorithm is further enhanced with the incorporation of linear predictive coding and cepstral coefficients for distance measurement. It is implemented and tested on a PC-486 system with a SoundBlaster card. Experimental results indicated that the speech-based system will simplify the operator control console but performance of the HC16 microcontroller may not be desirable when all the tasks of control, communication and speech recognition are taking place. Further research is continued to improve the algorithms and system hardware so as to achieve acceptable accuracy and efficiency.

INTRODUCTION

Underwater Remotely Operated Vehicles (ROV's) have been playing increasingly important roles over the past decades in a broad area of applications. Most of these applications are subseas oil-related operations such as construction of jackets/platforms, pipeline installation and underwater welding (Barrett, 1988). Other applications are underwater video inspections such as detection of cracks in dams, salvage operations, fisheries monitoring and they are even used to enhance safety of divers (Amoux, 1986). The most recent incident in using ROV's is the investigation of the ferry Estonia on the seabed of Baltic Sea. The sinking of the vessel which claimed up to 900 lives is considered to be the worst maritime disaster after the Titanic. The cause of the accident may never be known without ROV's. Although technology used in ROV's have improved with advances in the fields of computer and communication, the user interface or the operator-controls of the ROV's have remained much the same. At present, the commonly used user interface of an ROV consists of joysticks for position control and a number of switches used for the operation of onboard equipment. Visual feedback is provided to the operator via the video monitor and indicators on the screen or on the control console. This method of control is demanding on the user as it requires undivided attention and good hand-and-eye coordination. The user is obviously unable to perform any other tasks while controlling the ROV. This form of operator-control is inconvenient in many circumstances and alternative methods of interface are desired. Use of speech technology will reduce the number of controls on the console which require hand-and-eye operations, thus allowing the addition of further controls without overloading the operator (Hollingum & Cassford, 1988).

This paper reports the development of a microprocessor-based speech recognition system designed to enhance the operator-controls of the TITAN range of ROV's manufactured by the Titan Cove Holdings Ltd. The TITAN products are low cost ROV's which can be operated down to a depth of 200 m. The microprocessor-based speech recognition system is based on a Motorola M68HC16 microcontroller unit (MCU) and the development was carried out on a Motorola M68HC16Z1EVB evaluation board. The speech processing and recognition algorithms based on the energy function and dynamic time warping techniques (Rabiner et al., 1978) are implemented in assembly codes for maximum efficiency. The algorithms are further enhanced with linear predictive coding, and cepstral coefficients are used for distance measurement (Rabiner & Juang, 1993). These were implemented and tested on a PC-486 microcomputer with a SoundBlaster Card from Creative Laboratory. The programs have been developed in C++ programming language and this approach has reduced the hardware and software development time. The following sections outline the basic architecture of the TITAN ROV's, a brief description of the speech processing and recognition algorithms,

implementations on the HC16 system and on the PC-486 computer, experimental results and conclusion.

ARCHITECTURE OF ROV

Development and applications of underwater ROV's are not new, most systems in present day market however are expensive, bulky and mostly designed for specific purposes. The lack of suitable low cost ROV's has deprived possible applications of ROV's in many situations. The range of TITAN products have been developed to fulfill the needs of such sectors of the market and these products show great potential. A basic TITAN ROV consists of a power pack, a control console and an underwater vehicle (UWV). The UWV can be manipulated in three degrees of freedom. An example of such system is shown in Figure 1. A tether or umbilical cable between the power pack and UWV provides electrical power to the UWV and it also serves as the communication channel for the control and video signals. The power pack converts the mains input electrical power to appropriate voltage for transmitting down to the UWV. The control console provides user interface for motor control together with the onboard equipment such as lamp, auxiliary power and manipulator. Microcontroller-based systems are installed in both the control console and the UWV. These systems are responsible for the tasks of monitoring, control and communication. Communications between the surface and UWV are via customised protocols in digital form. Optional onboard sensory equipment may also provide environmental data from the UWV to the sea surface for monitoring purposes.

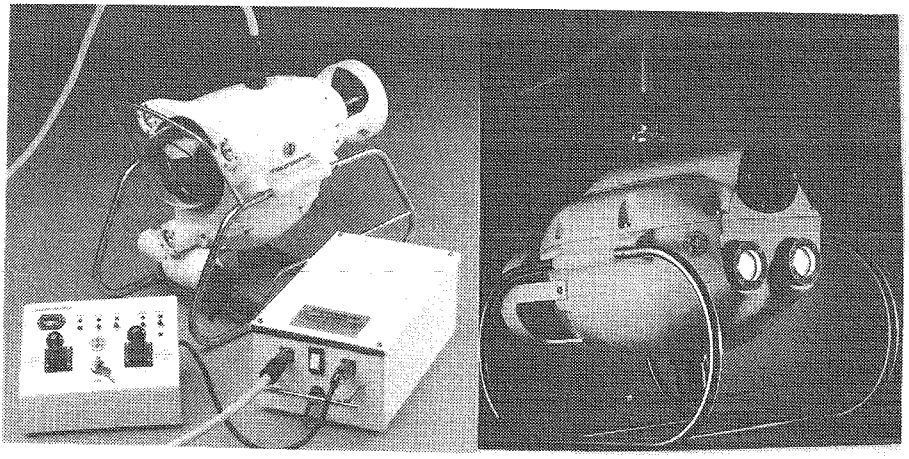

Figure 1: The TITAN ROV

SPEECH RECOGNITION

A generalised process for a speech recognition system is illustrated in Figure 2. The four steps are described as follows:

(i) Signal Acquisition: This process transforms the physical speech into a data format which the computer can deal with. The data is also stored for further processing. The in-built analogue to digital converter in the microcontroller is particularly suitable for this purpose.

(ii) Analysis: This stage reduces the amount of data from previous process and it extracts the required information which are necessary for recognition. In the present project, the energy function of the sample is calculated.

(iii) Similarity Calculation: This is the core of the recognition process. It involves the determination of the similarity between an unknown sample of speech and a reference template. The process measures the distance between two templates and the information is then used in the next process for decision making. The dynamic time warping algorithm is used to calculate the Euclidean distance in this case.

(iv) Decision Logic: The decision logic uses the distances measured from stage (iii). The distances between the input sample and various templates in the database are compared. The template which has the lowest calculated distance is considered to be a match.

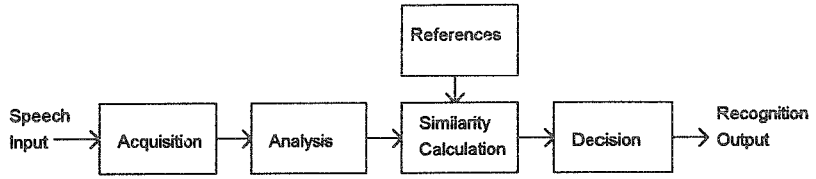

Figure 2: Generalised speech recognition process

MICROPROCESSOR-BASED SYSTEMS

The Motorola M68HC16Z1EVB evaluation board is used to design, debug, and evaluate the viability of M68HC16 based applications. This system is well suited to the project as it has the following features:

(i) extended digital signal processing language and multiplication addition unit;
(ii) availability of an analogue to digital converter for the sampling of speech input;
(iii) queued serial peripheral interface used to provide output indications;
(iv) extendible memory map for template storage; and
(v) a digital to analogue converter which can be used for playback facility to confirm proper recording of samples.

There is also a M68HCTK data acquisition board used for the development of an audio frequency analyser. It provides a programmable filter which has been programmed as an eighth order 10kHz lowpass filter. A summing amplifier on the device can also be used to combine audio input signals from left and right channels. This part of the circuit biases the audio signal at mid-point so that both positive and negative going waveforms can be sampled. This board greatly reduces the work on hardware development. Additional onboard LED's provide indication of the recognition results during the development stage.

IMPLEMENTATION OF SPEECH RECOGNITION ALGORITHMS

Speech detection

Prior to the analysis of the speech data, detection of the start of the speech data is essential. It is noted that the starting point of the speech data may be different to the starting point of the recorded data. Detection of the beginning of the speech input is based on the upper and lower threshold average magnitudes. The recorded input speech waveform is examined until the input has exceeded a pre-defined upper threshold value. The index at that point within the array is recorded as the initial beginning point of the speech waveform. Next, the algorithm steps backward through the speech waveform until it goes below a lower threshold. This is defined as the actual starting point of the speech waveform.

Signal acquisition

The input speech is sampled at 25kHz for 1 second. The above routine detects the presence of speech before commencing recording. Fifty samples are read in at a time and the energy of these samples are calculated. When the energy exceeds a predefined threshold, a period of exactly one second is recorded. This one-second recording is also checked against accidental triggering due to loud background noise or knocking of the microphone. A second level routine checks the energy between the 4600th and 4854th samples. This energy is compared against an upper threshold. When the sound exceeds this threshold value, the one second recording is accepted and the next stage begins.

Initial data compression

At a sampling rate of 25kHz, 25000 samples are obtained within a period of one second. This can be used in the playback stage for confirmation of template correctness. In order to reduce the amount of data so as to improve the processing speed, only every fourth sample value was used. This effectively reduces the sample size to 6250. This figure still gives a good representation of the sample as it exceeds twice the Niquist rate of 3000Hz. The value of 3000Hz is significant as it is assumed to be the upper frequency limit of speech. This reduced sample is now ready for the normalisation process.

Normalisation

The proportional normalisation technique (Das, 1982) is used to compensate for variations in energy due to varying distances between the user and the microphone. This technique is implemented by scanning each utterance in order to determine the maximum amplitude. This maximum value is then scaled to the largest possible number. All other samples are then scaled up proportionally with respect to this value. The following expression is used to determine the value of the normalised data.

$$V = U + (U-L)\,(M - Mx) / (Mx - L) \qquad (1)$$

where V is the normalised sample magnitude, U represents the unnormalised data value, L is the minimum amplitude, M is the maximum possible value and Mx is the maximum amplitude found after scanning through the entire waveform.

Analysis

The recorded signal is segmented into frames of 120 samples. Each frame is offset from the previous frame by 20 samples. Hence, each frame overlaps the previous frame by 100 samples. A short time average magnitude (STAM) which is the average of all the samples is calculated for each frame. The array of these STAM's forms the template and is used in the recognition process. This process effectively reduces the template size by a factor of twenty resulting in an array of 306 bytes in length. At this stage, the templates are now ready for distance measurement.

Distance measurement

This is the stage where the actual recognition process begins. The use of dynamic time warping allows correct matching of the templates even though they were recorded at different speeds. The optimum path between two templates forms a measure of the similarity between the sampled speech and a known template. Distances for all of the templates to be compared are obtained and they are used in the decision process.

Decision logic

The decision logic is a simple routine which compares the distance values obtained from the distance measurement process. The procedure compares the measured distances and determine the closest template. From this stage, the output of the recognition is determined and subsequent actions can be taken.

EXPERIMENTAL RESULTS

Due to the limitations of the memory capacity within the evaluation system, only up to four words are tested at a time. They are used to obtain the accuracy of different sets of vocabulary. Each word in the set are tested randomly and the average accuracy are obtained. A summary of the vocabulary used and the accuracy obtained are tabulated in Table 1.

From the experiment, it is observed that some words are more easily recognised as compared to others. The difficulty in differentiating some words is due to the similarity in the energy patterns between them. An example is the words Left and Right which gave only 35% accuracy between them. On the other hand, 100% accuracy can be obtained with the words Astern and Ahead. It is also observed the stored templates have great effects on the accuracy of recognition. If a badly recorded

template is taken to be the reference, the accuracy of subsequent recognition will be reduced dramatically. Hence it is necessary to have multiple templates or an average of more than one template should be used for each command. Alternatively, some form of adaptive process should be used to update the stored templates.

On	Auxiliary	Full	Forward	Up	Auxiliary	one
Off	Camera	3-quarter	Backward	Down	Manipulator	two
Ahead	Speed	Half	Left	Lock	Faster	three
Astern	Light	Quarter	Right	Unlock	Slower	four
85%	90%	83%	58%	60%	88%	70%

Table 1: Vocabulary tested and average accuracy

As regard to processing time, one second is required for the input and processing of each template at the teaching stage. During recognition, it requires three to four seconds to process an unknown word. If the vocabulary is expanded, the processing time will be longer. To improve the efficiency and to increase the number of stored vocabulary, the templates could be arranged in a tree-like structure so as to limit the number of templates to be compared at any one time. Due to limitations in the memory size within the evaluation board, approaches such as multiple templates and tree-like template arrangements have not been attempted on the HC16 system. The algorithms are further enhanced by incorporating a first order FIR filter to smooth the input waveform and a linear predictive coding algorithm. These algorithms are used for the analysis and distance measurement processes respectively. The enhanced algorithms and improved arrangements of vocabulary are implemented on a PC-486 and they are discussed in the following section.

IMPLEMENTATION OF ENHANCED ALGORITHM ON PC

A PC-486 with a SoundBlaster Card is used to test the enhanced algorithm. As compared to the HC16 evaluation board, the PC platform provides hardware and software resources which greatly facilitate the process. The added features include increased system memory, availability of secondary storage, increased clock speed, availability of high-level language compilers and the numeric co-processor for floating-point operations. The SoundBlaster Card is used because of the onboard sampling facility. The card used in the project is SoundBlaster version 2, which is only an 8 bit basic model. More advanced model such as the 16-bit SoundBlaster ASP card includes sophisticated onboard digital signal processing (DSP) functions. This card however was not used in the present project.

Compared to the algorithm implemented in the HC16 system, determination of the energy function is similar to that implemented in the microprocessor system except that initial data compression is not used. Also, the input signal is sampled at 15kHz instead of 25kHz and the short time average magnitudes are framed at 30 samples apart instead of 20. The size of each energy function template is therefore 500. In addition, linear predictive coding (LPC) algorithm is carried out and the cepstral coefficients are used to determine the optimum distances between templates. The process of the linear predictive coding algorithm is well documented and the steps taken in the present work are described as follows. The 15000 samples are first segmented into frames of 200 data values and each frame is offset from the next frame by 90 samples. This gives an array of 167 frames with 200 data in each frame. A FIR filter is then used to smooth the frequency spectrum of the input waveform for the linear prediction analysis. This is followed by a windowing function and the function effectively reduces the amplitude of the signals at the two ends of each frame. Autocorrelation analysis is then performed and each frame is reduced to 10 autocorrelation coefficients. The Durbin's algorithm is then used to covert the autocorrelation coefficients to the predictor coefficients. Finally these predictor coefficients are converted to 167 sets of cepstral coefficients. The dynamic time warping technique is used as before in the determination of similarity between templates of energy functions. However, the distance is not determined from the optimum path between two energy functions, rather, they are used as indices to obtain the respective cepstral coefficients of the two templates. This effectively takes into consideration of the frequency content of the speech waveforms. Since the dimension of the energy functions is 500 whereas the dimension for the cepstral coefficients is 167, the Euclidean cepstral distances of each third energy frame is added to obtain the total distance between two samples. The minimum distance now is used for the decision process.

The vocabulary is arranged in a tree-like fashion as illustrated in Figure 3. Results from experiment shows that near 100% recognition is obtained between words such as Propeller/Dive/Turn/Camera or Up/Down. However accuracy for words such as Left/Right is still comparatively low at 65%. Ongoing work is still carried out to determine an optimised set of vocabulary for the final system.

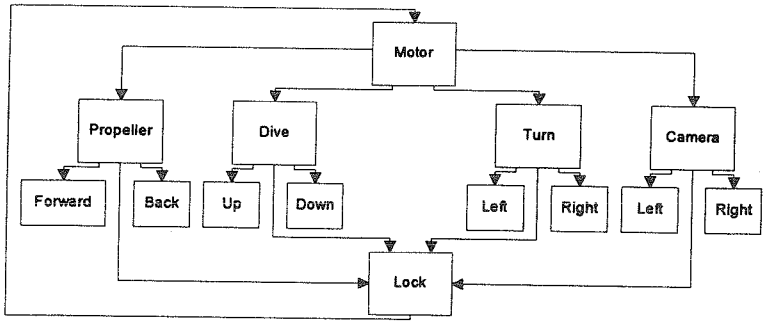

Figure 3: An example tree-like arrangement for vocabulary

CONCLUSION

Development of a HC16-based speech recognition system is reported. The experimental work has been carried out on a M68HC16Z1EVB evaluation board. Energy functions and minimum distance methods are adopted for the representation of speech templates and for the recognition process respectively. Due to limitation of memory capacity on the evaluation board, only a small vocabulary can be tested at any one time. The algorithm is enhanced with linear predictive coding and cepstral coefficients analysis. It is implemented on a PC-486 computer with a SoundBlaster card and the results are promising. A hierarchical structure arrangement of the vocabulary is used and this has improved the recognition process. Further investigations are carried out to optimise the set of vocabulary to be used for controlling of the ROV and a dedicated system is also needed to be developed. A 32-bit microcontroller with 12 to 16 bit analog-to-digital converter may be required for this particular application.

ACKNOWLEDGMENT

The authors wish to thank Mr. Ted Kneebone of the Titan Cove Holdings Ltd. for the support given to this project and for the permission to publish the results. The suggestions and expert advice from Mr. Kneebone are gratefully acknowledged.

REFERENCES

Arnoux, G. (1986) 'ROVs increase diver safety', *Proceedings of the ROV '86 Conference*, Society for Underwater technology and the Association of Offshore Diving Contractors, Aberdeen, June 1986, pp. 121-123.

Barrett, R.W. (1988) *Advanced Robotics under Water: the DTI Initiative*, Advances in Underwater Technology, Ocean Science and Offshore Engineering, Volume 14: Submersible Technology, pp. 247-255, (Graham & Trotman: London).

Das, D. (1982) 'Some Experiments in Discrete Utterance Recognition', *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-30, no. 5.

Hollingum, J. and Cassford, G. (1988), *Speech Technology at work*, IFS, Bedford.

Rabiner, L., Rosenberg, A. and Levison, S. (1978), Considerations in dynamic time warping algorithms for discrete word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-26, no. 6, pp. 575-582.

Rabiner, L. and Juang, B.H. (1993), *Fundamentals of speech recognition*, Prentice-Hall, New Jersey.