SPEAKER INDEPENDENT RECOGNITION OF SMALL VOCABULARY

Jason Chong and Roberto Togneri

Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia

ABSTRACT

This paper reports on the implementation of a real-time speaker independent isolated word speech recognition program on a PC Windows platform. The overall structure of the recognition engine is based on the Dynamic Time Warping (DTW) paradigm for computational efficiency. Furthermore, to decrease the recognition time and increase the recognition accuracy, the dictionary is limited to under 15 words. This severely restricts the vocabulary. To overcome this restriction, a new technique is introduced. Many dictionaries are linked in a hierarchical structure and each word in each dictionary will activate a new dictionary related to that word. This represents a basic form of language modelling which is suited for the menu driven interface found in many of today's applications. The results show that reasonable performance can be achieved by these methods.

INTRODUCTION

The fundamental objective of a speech recognition process is to identify a speech utterance which belongs to a given vocabulary with high accuracy. This problem becomes more complicated if the vocabulary is large, if multiple speakers are used or if the speech is continuous. The complexity of a speech recogniser can be significantly simplified if constraints to the system are imposed. These include decreasing the size of the vocabulary or limiting the recognition to isolated words.

Many user interfaces to computer systems are now menu driven and the number of options in each of these menus is usually limited. In this project, a speech recognition system has been designed and implemented which identifies an isolated utterance from a small vocabulary, such as those found in menus. The system utilises fast algorithms on a PC platform. The recogniser is also speaker independent and therefore a robust method of constructing the reference templates is used.

The combination of existing speech technologies has been carefully chosen to enhance not only the recognition rate but also the computational efficiency. The speech recognition system uses a pattern recognition technique. This consists of four main components: feature extraction, reference template formation, comparison of the test and references and a decision logic. Each of these components have been implemented on a PC platform.

This speech recognition system will provide many advantages on a standard PC platform. The recogniser will reduce the need for users to use a physical input device such as the keyboard and mouse to perform standard menu orientated tasks. It will also increase the user friendliness of the operating system environment as well as providing benefits for disabled and computer illiterate users.

SYSTEM STRUCTURE

Hardware

An IBM compatible Pentium 133Mhz system using the Microsoft Windows 95 operating system is used for both the data retrieval, processing and system development. This computer hardware setup will enable the speech recognition system to be used with popular Windows applications and the variety of soundcards which are supported by Windows. The development system also uses a Creative Labs Soundblaster 16 Plug and Play for recording the speech utterances. The entire system was developed under Visual C++ and hence some of their library routines were utilised for low level access to the sound device.

Software

The implemented speech recognition system is based on a pattern recognition technique. The main motivations behind using this approach include the short training time as compared to the Hidden Markov Model approach (Rabiner et al, 1993), its more straight forward implementation, and its flexibility.

The structure of the pattern recognition system is shown in Figure 1. There are four main procedures in the implemented pattern recognition approach. These will each be discussed below.

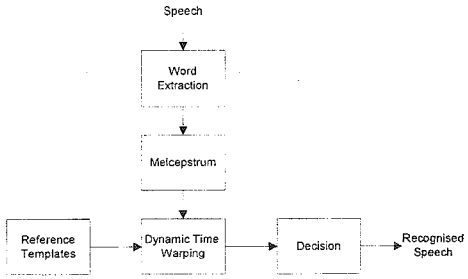


Figure 1: Block diagram of the system structure

For increased usability, the speech recognition system continuously samples data from the input microphone. This will cancel the need for the user to specifically record each command utterance to the system. The continuous monitoring of the audio input is performed with a circular buffer technique. By using a rudimentary energy threshold calculation, the buffers with significant sound data in them are passed on for further processing. However, as the speech recogniser will identify isolated utterances only, a technique must be used to determine the endpoints of the isolated words.

The method used in this speech recognition system to determine the endpoints uses the short time energy spectrum. This algorithm is based on that proposed by Rabiner and Sambur, but

without using the zero crossing rate (Rabiner & Sambur, 1975). Two different thresholds are defined. These are calculated as functions of the percentages of the maximum energy level in the speech signal. The other difference from the algorithm proposed by Rabiner and Sambur is that the final endpoint is determined in a forward time fashion. This is required since the system processes the speech in real time.

Once the isolated word has been segmented from the continuous signal stream, a discrete Fourier transform is computed for each 30ms frame width with a 15ms advance. The standard Hamming window is applied to each frame. From the DFT coefficients, 24 cepstral coefficients are calculated to produce a 12 dimensional melcepstrum feature vector to describe each frame.

The dynamic time warping algorithm was used to compare the test and the reference utterances. Since this algorithm is computationally expensive (Sakoe & Chiba, 1978), constraints have been imposed to reduce the complexity of this operation. One such constraint is to limit the area of the search path by restricting the average gradient of the path (Myers et al, 1980). The constraint implemented in this speech recognition system is the symmetrical condition defined as P=1. The allowable slope advancements are shown in Figure 2.

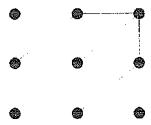


Figure 2: Allowable local path constraints

The speech recognition system developed also limits the calculation space of the paths to a narrow diagonal strip ±[N+D] frames from the diagonal line connecting the beginnings and ends of the test and reference words. The value D is the difference in the number of frames between the test and reference utterances whereas N is a variable but is in the order of 5 frames.

The pattern recognition approach to speech recognition requires no speech specific knowledge to be explicitly encoded into the system. This method is relatively ignorant of the choice of words, speaker, task or syntax. Therefore the reference templates are of great importance as these form the difference between a speaker dependent and a speaker independent system.

It has been shown (Rabiner et al, 1979) that a few carefully constructed templates can adequately represent a large speaker population for speaker independent word recognition. In the system developed, a k-means clustering algorithm is employed to generate 3 reference templates per word.

Linked Dictionaries

One small vocabulary (10-15 words) does not make a very useful or flexible system. The size of the vocabulary can be increased without decreasing the recognition rate by using many individual dictionaries which can be swapped to simulate a large vocabulary.

As the main task of this system is suited to menu driven systems, it is noted that menu systems consist of only a small set of options at one time. These options change as the user selects an item. Hence, the active dictionary of the implemented speech recognition system will change according to the previously spoken word.

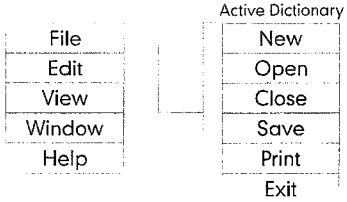
Each utterance in each dictionary will therefore need to be linked to following dictionary. This method is similar to a basic kind of language modelling (Martin, 1996). The method of linking the dictionaries together will initially be performed manually or via some graphical technique. The hierarchical structure of the dictionaries is ideal for the many menu driven interfaces used in computer applications.

By reducing the number of words in the active dictionary, the accuracy of the speech recogniser will be increased and the computation time for recognition of a single utterance will be decreased. However the many dictionaries will give flexibility by simulating a larger vocabulary.

The procedure by which the dictionaries are swapped is given in an example below.

In a typical menu system, the options are as follows:

If the user says the word "File", the active dictionary should change to reflect the new options which are available at this stage.



Each word in each dictionary will have an associated dictionary. Thus all the dictionaries will form a hierarchical structure that can be easily edited to suit a menu based application system.

EXPERIMENTAL RESULTS

For the testing, two sets of ten words from the September 1991 NIST TI 46-word Isolated Word Corpus CD were used. These were recorded at 16 bit 16 kHz quality in a low noise sound isolation booth. The first set of test words consisted of the 10 digits from zero to nine. The second set of data contains 10 isolated control utterances.

Eight different female speakers were used to generate the reference templates. Each word was uttered twice by each speaker. Three reference templates were then generated for the 16 tokens of each word. For the testing data, each of the 10 words were recorded three times for each of the 8 speakers.

Initial results showed an 88% recognition for both of the data sets. The most common errors were between similar sounding words such as "No" and "Go". However, in a real menu driven system where each word or command would be significantly (and deliberately) distinct from each other, these situations should not occur implying a higher recognition rate. The recognition time per word was in the order of less than 0.7 seconds per word.

The linked dictionary technique requires minimal time to swap from one dictionary to another as most of these would be loaded into memory before use.

CONCLUSIONS

A speaker independent speech recognition system was developed to recognise a small vocabulary of isolated words. All algorithms were implemented in software and the system runs in real time. The system can run on any Windows 95 platform with a compatible soundcard. The combination of efficient yet effective and simple existing algorithms were used for each part of the system structure. Twelve mel-frequency cepstral coefficients represented each frame of the utterance. The test and reference words were compared with the dynamic time warping algorithm.

Initial results show that the recognition accuracy is above 88% for each word and the recognition time is below 0.7 seconds. Therefore reasonable performance can be achieved by this system. The main feature of this system is the method in which it uses many linked dictionaries in a hierarchical structure to simulate a large vocabulary. This structure is ideal for the menu driven interfaces found in many of today's computer applications.

REFERENCES

Martin P., et. al (1996) *SpeechActs: A Spoken-Language Framework*, IEEE Computer, Vol. 29, No. 7, 33-40.

Myers, C., Rabiner, L. & Rosenberg, A. (1980) *Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 6, 623-635.

Rabiner, L. R. & Juang, B. H. (1993) *Fundamentals of Speech Recognition*, (Prentice-Hall: New Jersey).

Rabiner, L. R., Levinson, S. E., Rosenberg, A. E. & Wilpon, J. G. (1979) *Speaker Independent Recognition of Isolated Word using Clustering Techniques,* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, 336-349.

Rabiner, L.R. & Sambur, M.R. (1975) *An Algorithm for Determining the Endpoints of Isolated Utterances*, The Bell System Technical Journal, Vol. 54, No. 2, 297-315.

Sakoe, H. & Chiba S. (1978) *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, 43-49.