Jurnal Kejuruteraan 3 (1991) 41 - 46

Isolated Digit Recognition in Malaysian Language

Zainul Abidin Md. Sharrif, Masuri Othman &
Tan Shao Theong

ABSTRACT

This paper describes the use of a personal computer to store samples of uttered isolated digit signals and thereupon to recognize the appropriate digit uttered using time domain analysis and template matching. The fundamental of linear prediction analysis is briefly introduced and a general description on the hardware and software used in this project is then given. This project is undertaken as an initial step to implement an automatic translation system of daily used keywords terminology in Malaysian to English.

ABSTRAK

Kertas ini membincangkan penggunaan komputer peribadi untuk menyimpan sampel-sampel isyarat pertuturan dan mencam perkataan digit yang diturturkan dengan menggunakan kaedah analisa dalam domain masa serta perbandingan templat. Asas-asas kaedah pengkodan ramalan akan diperkenalkan secara ringkas diikuti dengan perkakasan dan pengaturcaraan. Projek ini dilakukan sebagai satu langkah permulaan kepada perlaksanaan satu sistem pencaman dan pengalihan bahasa secara automatik dalam bahasa malaysia kepada bahasa inggeris.

INTRODUCTION

The dream to be able to communicate with machines by speaking into microphone has long existed. Many techniques have been devised by speech technologists to come closer to this dream. Although communication through speech appears to be a simple process, making machine to understand them is not. Many techniques have been divised to achieve this goal and these techniques can generally be divided into time domain analysis and frequency domain analysis. There are currently a few successful digit recognition reported (Rabiner, et al. 1975; 1989) and, most of them used time-domain analysis such as combination of measurement on energy, zero-crossing rate and linear prediction. The method employed in our system is the linear prediction coding method. This technique is expected to become the predominant method for estimating basic speech parameters due to its speed of computation and its ability to provide extremely accurate estimates of basic speech parameters such as pitch and formants, spectra and vocal tract area function. (Rabiner et al. 1978; Thomas 1986)

FUNDAMENTAL PRINCIPLE OF LINEAR PREDICTION CODING

The basic idea behind LPC is that a sample of speech can be approximated by a linear combination of the past p speech sampels. This can be shown in a closed form as below.

$$\hat{s}_n(m) = -\sum_{k=1}^{p} a_k s_n(m-k) \qquad (1)$$

for $1 \leq k \leq p$

where s_n (m) : samples in a segment

k : lag

p order of LPC

The basic problem is to determine a set of predictor coefficient $\{a_k\}$ directly from speech signal. This is usually done by finding the set of predictor coefficients that will minimize the total squared error over a short segment of the time-varying speech signal. The total squared error is given by equation (2).

$$E_n = \sum_{n=n0}^{n1} (e_n^2(m))$$

$$= \sum_{n=n0}^{n1} [s_n(m) + \Sigma\, a_k s_n(m-k)]^2 \qquad (2)$$

This minimization is carried out by solving the equation below, which is obtained by setting $\partial E_n/\partial a_k = 0$ for $1 \leq k \leq p$.

$$\sum_{i=1}^{M} a_i c_{ik} = -c_{ok} \qquad (3)$$

where $1 \leq k \leq p$.

and

$$c_{ik} = \sum_{n=n0}^{n1} s_n(m-i)s_n(m-k)$$

In the autocorrelation method, the data sequence is treated as though it was zero outside the interval from $m = 0$ through $m = N - 1$. Then the coefficient set, $\{c_{ik}\}$, which form the elements of a symmetric positive definite Toeplitz matrix can be expressed in term of an autocorrelation sequence as below.

$$c_{ik} = r(|i-k|).$$

$$\text{where } r(k) = \sum_{i=0}^{N-k-1} s_n(m)s_n(m+k) \qquad (4)$$

The LPC coefficients sequence is then evaluated using Durbin's recursion method as discussed in (IEEE 1979).

SAMPLING AND PREPROCESSING

The isolated Malaysian Digit Recognition System is developed using an IBM PC/AT personal computer equipped with 512KB of RAM, a numeric coprocessor (80287) and an IBM Analog to Digital Card. An audio microphone is then connected to a home-built eight-pole maximally-flat low-pass (−3dB at 3.9kHz) presampling filter and a preamplifier. This preamplifier output is then connected to the input of the ADC card.

Sampling is triggered when the input signal exceeds a certain threshold value. This threshold value is automatically set by the system to be above the background noise prior to the audio input. Still, it was found that false triggering due to audio background noises occurs frequently and thus, a false triggering detection algorithm will be employed at a later stage of this investigation.

The sampling is conducted in a relatively quiet computer room with all the air conditioning switched off. The sampling rate used is at 8000 samples per second. The rate is chosen as it is enough to achieve the telephone quality speech. Once the sampling is done, the samples are then preprocessed by passing them through a pre-emphasis digital filter. By doing so, the higher frequency components are preemphasized and the dc component removed. The flow graph of this filter is shown in Figure 1. The overall hardware is shown in Figure 2.

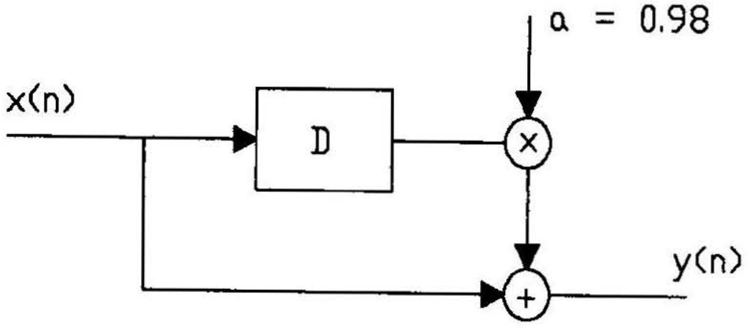

FIGURE 1. Pre-emphasis filter

TRAINING AND RECOGNITION

Before the systems is ready to recognize any digit, it has to go through a 'training' stage. In this stage, the user is required to utter the digit zero to nine in Malaysian three times. From these sampels, a good sample of each digit is then selected. These samples are then converted to LPC coefficients and stored as templates. The conversion is done by first windowing the samples with a shifting rectangular window and then, calculating the autocorrelation sequence and finally, solving for the ten LPC coefficient using Durbin's recursion. To ensure that at least two pitch cycles are enclosed by the window, a window width of 320 samples is chosen, since the pitch of natural speech ranges from 50Hz to 400Hz.

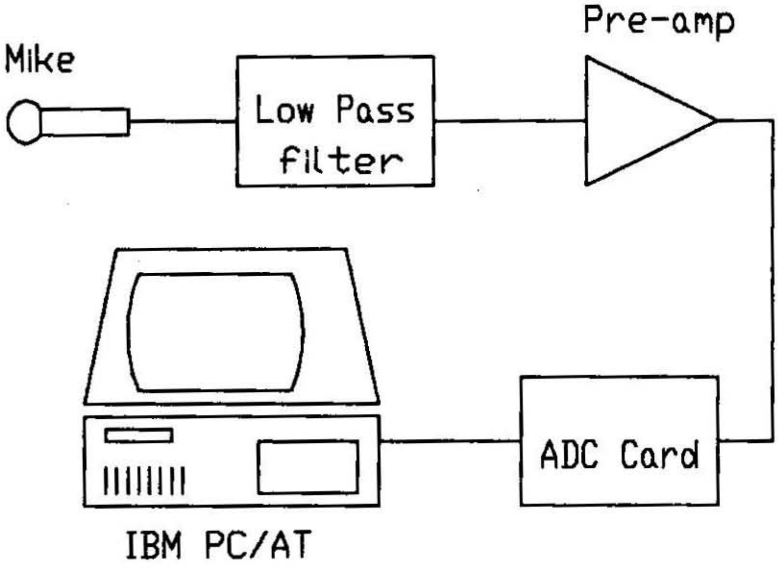

FIGURE 2. Hardware description

Recognition is done by using a simple template matching algorithm. Samples taken from the user at the recognition stage will also go through the same procedure as in the training stage. But, the ten LPC coefficients obtained will not be stored. Instead, it is used to determine a match with the stored templates (obtained during the training stage). The matching is done by utilizing the simple distance measure equation given below.

$$d_w = \sqrt{\Sigma_i \Sigma_k (a_{ik} - a_{rk})^2} \qquad (5)$$

where

d_w : linear distance.

a_{ik} : test LPC coefficients.

a_{rk} : templates LPC coefficients.

i : ith order LPC coefficients.

k : k th window

The digit recognized is the digit which has the smallest value of d_w.

TESTING

A simple test has been carried out where a user is asked to train the system before using it. The user has to exercise caution on the way he speaks during the training phase as well as in the recognition phase. The percentage of match is approximated to be above fifty percent if the system is used just after the training phase. The percentage of match will be lower if it used at different time by the same user as human speaks at a different rate at different time. A proper test is yet to be conducted to gauge the performance of the system.

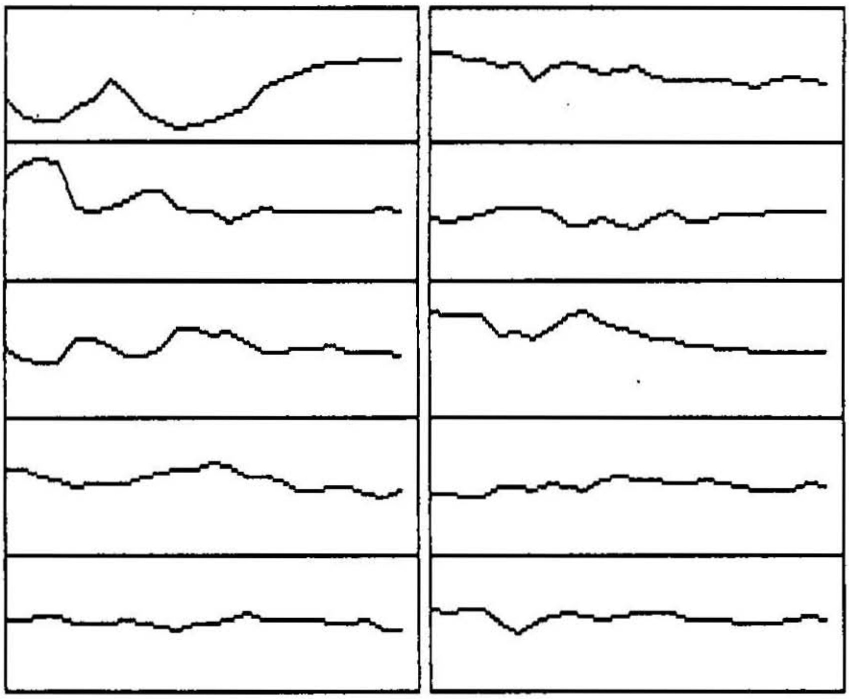

FIGURE 3. Graphs showing the magnitude of the ten LPC cofficients versus time for the word "satu"

DISCUSSION

Speech recognisers presently suffer from the following setbacks. Firstly, performance of these systems tend to decrease with time as people tends to speak at different rate at different time. Secondly, such systems tend to be speaker dependent even if the vocabulary size is small. This is because different people have different sized vocal tracts and therefore different templates is required for different user to bring the performance of the system to an acceptable level. Thirdly, most of the speech recognisers are required to operate on field where the background noise are unpredictable and mány times higher than a quiet computer room.

These problems are not new and many techiques have been devised and successfully implemented such as dynamic time warping which allow the templates to be expanded or compressed to obtained an optimum match. This will overcome the first problem to a certain extend. To over come the variability of different speaker, template transformation and statistical modelling techniques such as Hidden Markov Model has also been employed by other digit recognizer. (Rabiner and et al. 1989) Speech recognition in noise, however, is still at its early stage. The superior performance of people in recognising speech in noise has led to the suggestion that speech analysers which operates on the same principles as the human auditory system might work better than those conventional system.

The research in UKM on speech processing and recognition, is still at its early stage. Various methods such as Dynamic Time Warping, Hidden Markov Model, Vector Quantization and Coding and hopefully Neural Network are being looked at and large knowledge and experimental results are yet to be collected. Although the Malaysian Isolated Digit Recognition System shown here is simple, it serves as an initial step towards our objective of implementing an automatic system for translation of daily used keywords terminology in Malaysian to English and vice versa.

NOTATION

d_w	linear distance.
a_{ik}	LPC coefficients.
a_{rk}	templates LPC coefficients.
c_{ik}	autocoleration matrix
$r(k)$	sequence of autocoleration coefficients
$s_n(m)$	samples in the n th segment
k	lag
p	order of LPC
E_n	Error square at the n th segment

REFERENCES

Sambur & Rabiner L.R. 1975. Speaker-independent digit recognition system. *Bell System Technical Journal* 81–102.

Rabiner, L.R. 1989. High Performances Connected Digit Recognition Using Hidden Markov Models. *IEEE Transcation on Acoustic and Signal Processing* 37(8).

Rabiner L.R. & Schafer R. W. 1978 *Digital Processing of Speech Signals.* Prentice Hall: Englewood, New Jersey.

IEEE Committee of Speech Acoustics and Signal Processing. 1979. *Programs for Signal Processing.* New York: IEEE Press, John Wiley & Sons Inc.

Parson, Thomas. 1986 *Voice And Speech Processing.* McGraw Hill, Inc.

Witten, I.H. 1982 *Principles of Computer Speech* London Academic Press, Inc.

Fakulti Kejuruteraan
Universiti Kebangsaan Malaysia
43600 Bangi
Selangor D.E.