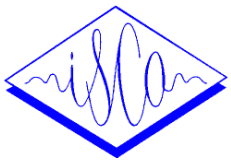

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

RECOGNITION OF STANDARD MALAYSIAN LANGUAGE PRONUNCIATION

Zainul Abidin Md. Sharrif
Masuri Othman
Mohammad Ibrahim AKB Maiden

Department of Electrical, Electronic and System Engineering
National University of Malaysia (UKM),
43600 UKM, Bangi,Malaysia

ABSTRACT

In line with the Malaysian Government policy of implementing standard Malay language pronunciation in schools and institutions of higher learning, research on the recognition of spoken Malay language is currently being pursued at the National University of Malaysia.The purpose of this research is to develop a computerised system which can recognize objectively standard Malay language. Uttered Malay words are sampled and broken into phonemes. The voiced and unvoiced phonemes are stored using different parameters. The zero-crossing rate and average magnitude of the signal are used to build the template for the unvoiced phonemes, while the normalized time sampled signal over a selected pitch period and the duration of the phonemes are being used as parameters for the voiced template.In the recognition part, Dynamic Time Warping (DTW) algorithm is used to compare between the voiced template and the input voiced samples. The recognition is done by comparing the sample within the pitch period for the voiced part. However in the case of unvoiced recognition, it is done through the computation of zero-crossing and average magnitude. A database of Malay words is used to compare the recognised words with all the words in the database. A decision as to whether the sound of the uttered word is in accordance with the standard Malay language is achieved through this comparison.Work is actively being undertaken to port the developed system from the IBM PC-AT to the TMS320-based system[1].

INTRODUCTION

Recently, The Ministry of Education, Malaysia, has announced the usage of standard Malaysian language pronunciation or what is known as the 'Bahasa Baku', for all purposes in teaching and learning in schools and institutions of higher learning. This announcement has attracted us at the department, to develop a speech recognition system that would be able to recognize the accuracy of uttered word pronunciation compared to the standard word which is pronounced in Bahasa Baku. This system is intended to be used as a teaching aid to fulfill the needs of those learning the language. In addition, there is no efficient method to determine the accuracy of the uttered Malay words pronunciation.

Bahasa Baku is a standard where each letter represents a phoneme, and is pronounced the same way regardless of its position in the word[2]. This concept is implemented directly as an algorithm in the research to develop the recognition system which is speaker independent.

Although many speech recognition systems are available, however most of these systems are limited to the recognition part only and cannot measure the accuracy of pronunciation compared to the standard word. Recognition of standard pronunciation is more difficult and challenging due to the more stringent requirement.

The concept of Bahasa Baku pronunciation has resulted in several possible recognition methods with new ideas and techniques. Uttered Malay words are sampled and broken into the basic unit of sound called phonemes. The sampled signal is first divided into voiced and unvoiced segments, and each segment processed separately. Voiced segments are subdivided into smaller unit of one pitch period each, while zero-crossing rate and average magnitude are computed over the unvoiced segments[3].

Dynamic Time Warping[4] is used to align the time scale, non-linearly, between the template and sampled signal over the pitch period, in order to obtain the minimum difference between the two. Recognised phonemes of the uttered word are then compared with the standard word. The difference between the two is calculated to determine the accuracy of pronunciation.

CLASSIFICATION OF STANDARD MALAYSIAN LANGUAGE SOUNDS.

Uttered Malay words can be classified into three major categories of sounds : 1) vowels 2) diphthongs 3) consonants. This classification is shown in Fig. 1.

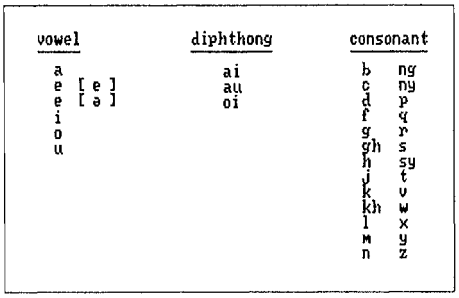

Fig. 1: Phonemes in Malay words[2].

Combination of vowels and consonants or diphthongs and consonants produce a syllable.

Malay syllables				Malay words	English equivalent
e	+nak		=	enak	(delicious)
sa	+ya		=	saya	(me or I)
ka	+wa	+san	=	kawasan	(area)
ba	+tu		=	batu	(stone)
ba	+ha	+ya	=	bahaya	(danger)

Finally, the possible combination of these syllables produces a Malay word.

RECOGNITION OF STANDARD MALAYSIAN LANGUAGE PRONUNCIATION.

Separation of voiced and unvoiced parts of the word.

The average magnitude level is used to determine the beginning and ending points and also to identify the voiced and unvoiced parts of the word.

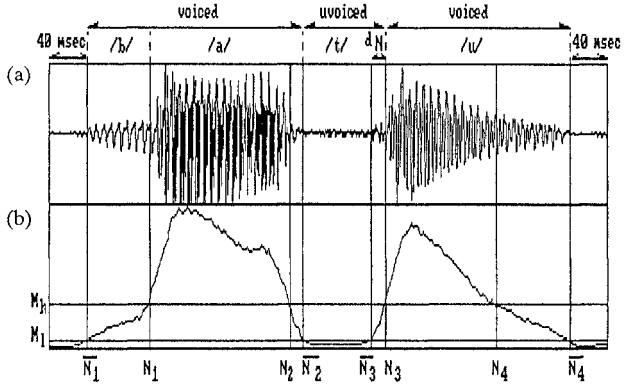

Fig. 2: Average magnitude level representation of the word 'batu'. (a) input signal (b) average magnitude level.

Average magnitude of the signal is calculated over the window size of 30 msec. An overlap ratio of 14/15 is used. Average magnitude of k^{th} frame is given by:

$$M_k = \frac{1}{N} \sum_{n=k-(N/2-1)}^{k+(N/2)} s(n) \qquad (1)$$

where N is size of window, s(n) is the input signal. The average magnitude level representation is shown in figure 2.

The maximum average magnitude of the word is determined. Base on experience, the high threshold M_h is selected as 30% of the maximum average magnitude while the low threshold M_l is taken as 5% of the maximum average magnitude.

The average magnitude profile is searched so as to determine the N_1, N_1, N_3 and N_3 points as shown in figure 2. From these points, the interval d_N at which the average magnitude is between M_h and M_l is then determined.

If d_N is less than 40 msec or the zero-crossing rate in this interval exceeds a threshold, 3000 crossings per second, then this interval is considered as the end part of the unvoiced segment. The whole unvoiced segment can then be determined as the interval between N_2 and N_3 as shown in figure 2.

Recognition of unvoiced segments.

For unvoiced segments, recognition is done through the computation of zero-crossing rate and average magnitude. Unvoiced part of Malay words can be classified into two major categories:

i) plosive unvoiced ; /k/,/t/,/p/,/q/ and /c/ where a plosive occurs at the end of the segments.

ii) non-plosive unvoiced ; /s/,/f/,/h/ and /sy/ where the energy level and zero-crossing rate of these segments more uniform compared to the plosive unvoiced.

Using the sampling technique, the first 100 msec of the interval contains no speech (background noise). The average magnitude (M_o) and zero-crossing rate (Z_o) are computed for this interval.

Recognition of the unvoiced segments is done via four parameters (M_1, M_2, Z_1 and Z_2) as follows:

$$M_1 = \frac{M_L}{M_F} \qquad (2)$$

$$M_2 = \frac{M_L}{M_o} \qquad (3)$$

$$Z_1 = \frac{Z_L}{Z_F} \qquad (4)$$

$$Z_2 = \frac{Z_L}{Z_o} \qquad (5)$$

where,

M_F = average magnitude over the first 40 msec interval,of the segment,

M_L = average magnitude over the last 40 msec interval,of the segment,

Z_F = zero-crossing rate over the first 40 msec interval of the segment,

Z_L = zero-crossing rate over the last 40 msec interval of the segment.

M_1 and Z_1 are used to differentiate between the plosive and non-plosive unvoiced while M_2 and Z_2 are used to recognize the unvoiced phonemes. As an example, plosive /k/ has fairly large M_1 and small Z_1 compared to the fricative /s/ with both M_1 and Z_1 approximately equal to one.

The recognition of plosive unvoiced such as /k/ and /t/ possesses a rather difficult problem as they exhibit similar properties. However this problem can be overcome at the end of the recognition phase by referring to the Malay words dictionary.

Recognition of voiced segments.

For voiced segments, pitch period detection is done over the entire segments. A simple procedure is used at which the peaks and the valleys of the signal are located[3]. An impulse equal to the maximum swing between the peak and its valley neighbours is being identified as shown in figure 3.

A 10 msec window with an overlap ratio of 1/2 is being applied to the voiced segments. The maximum magnitude (M_{max}) of the signal within this window is identified. The impulses that exceed the threshold A_r (70% of M_{max}), are obtained. By comparing successively the impulses that are greater than their nearest neighbours, we will be able to determine the pitch period of the voiced segment.

During the training mode, time normalized sampled signal over the entire pitch period of different phonemes are stored as templates. The number of samples and peaks within the pitch period is also being kept in the template. Recognition is done by comparing the waveform in one pitch period with templates. Comparison is done through the computation of mean squared error between the sampled signal and template.

$$e_1 = \sum_{n=0}^{N} [s(n)-s_t(n)]^2 \qquad (6)$$

where s(n) is the sampled signal over the pitch period at n^{th} sample, $s_t(n)$ is the template signal at n^{th} sample and N is number of normalized samples.

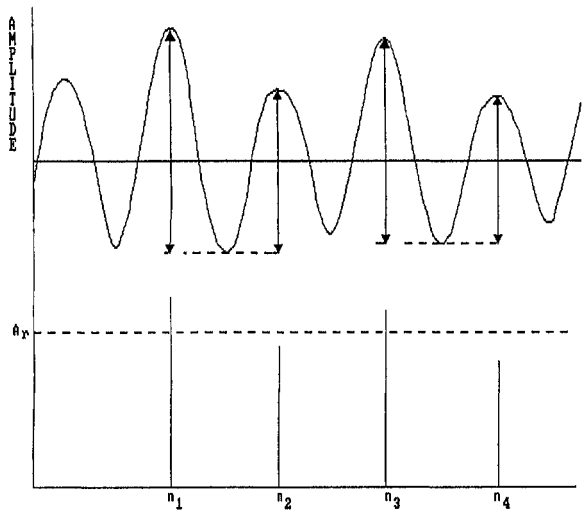

Fig. 3: Impulse trains generated from peaks and valleys.

However, the pitch period is different for the same word being uttered by the same person at different time or by a different person.

To overcome this problem, Dynamic Time Warping (DTW) is used to align the time scale of the sampled signal with the template. DTW simply changes the time scale of the input signal, non-linearly by matching the peaks and valleys, in order to obtain the minimum difference between the two as shown in figure 4.

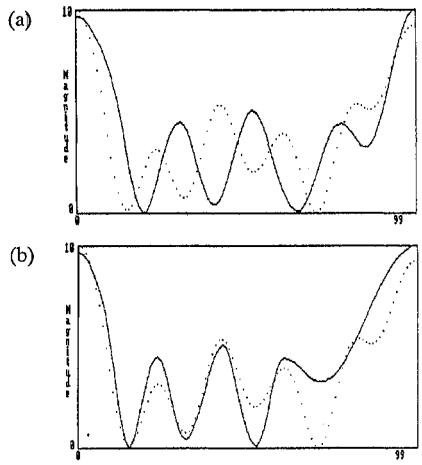

Fig. 4: Signal for /a/ and template, (a) before time warping, (b) after time warping.

Mean squared error computed between the aligned sampled signal and the template is used for comparison.

$$e_2 = \sum_{n=0}^{\overline{N}} [\overline{s(n)} - s_t(n)]^2 \qquad (7)$$

where $\overline{s(n)}$ is the aligned signal, $\overline{N}$ is the number of samples within pitch period after alignment.

e_1 and e_2 (from equation (6) and (7)) are compared to get the minimum error between the sampled signal and template. Similarly, the process is repeated with other phonemes in the table of templates. The template which provides the minimum difference with the input signal is considered as the recognised phoneme.

Finally, in the recognition part, all the recognised phonemes, from unvoiced and voiced segments, are then recombined in the order to form a word. This predicted word is then compared with the available vocabulary in the dictionary. The word which provide the minimum difference is considered the spoken word.

Recognition of the pronunciation.

To determine accurately how the spoken word is being pronounced in accordance with Bahasa Baku, the difference between the recognised word and the Malay dictionary is calculated.

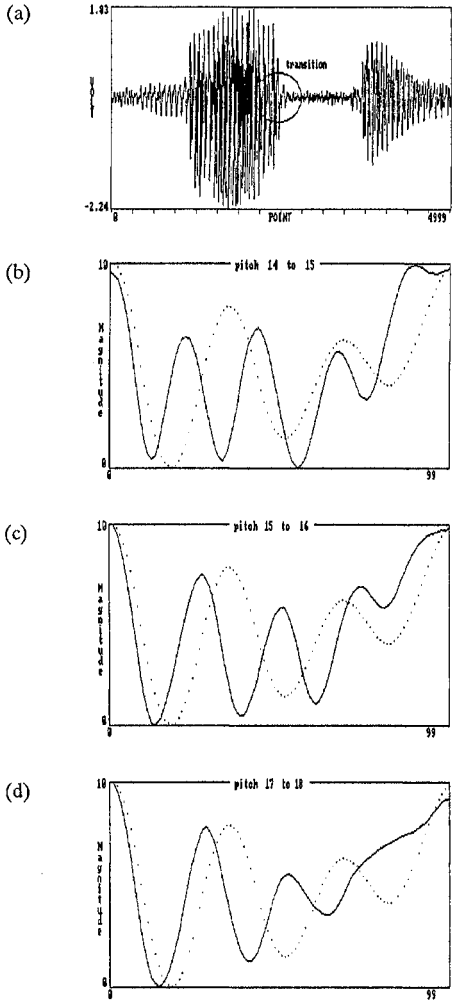

Fig. 5: Transition in the word 'batu' between the /a/ and the /t/, the doted curve represents the template for phoneme /e/, fig. (a) the input word, (b) phoneme /a/ pitch period 14-15, (c) pitch period 15-16, (d) pitch period 17-18 where the signal is close to the shape of template /e/.

As an example, the word 'saya' (me or I) is usually pronounced as 'saye' which will result in the existence of different phonemes at the end of the word compared to the standard word. Three of the phonemes are pronounced correctly. This will result in 75% accuracy in the pronunciation for the standard Malay word 'saya'. The smaller the difference of the phonemes between the two words, the more standard is the pronunciation.

However there are problems. For instance the word 'batu' (stone) should contain the phonemes /b/,/a/,/t/ and /u/ only. But on examining the signal waveform closely, it reveals the existence of an /e/ phoneme between the /a/ and the /t/ which is the result of the transition between the two phonemes as shown in figure 5.

To overcome this problem, all of the phonemes that exist within the transition and the actual phonemes are stored together as the template word. As an example, the template for word 'batu' is stored as 'baetu'.

CONCLUSION

Though the research on recognition of Bahasa Baku has just started (early of 1990), we have achieved some degree of success. A number of methods in speech recognition has been implemented with new ideas and techniques to achieve the objective of this project to produce a recognition system which is speaker independent. Work is currently being undertaken to overcome the problems associated with speech recognition due to the large variation in human speech. The algorithms used in this recognition system will be implemented on a digital signal processing microcomputer system (Texas Instrument TMS320C30). Step towards realization of these algorithms in real time is already in progress so as to develop a teaching aid system for the recognition of standard Malaysian language pronunciation.

REFERENCES

[1] Zainul Abidin Md. Sharrif, Masuri Othman, Wan Jalaludin Wan Hussein, Kader Ibrahim, Mohammad Ibrahim AKB Maiden, "Man Machine Interface : Speech", International Conference On Information Technology, ICIT '90. 17-20 Sept 1990, Kuala Lumpur, Malaysia.

[2] Educational Technology Section, "Panduan Sebutan Baku Bahasa Melayu", Dewan Bahasa dan Pustaka, Ministry of Education, Malaysia, 1989.

[3] L.R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signals", Prentice - Hall Inc, 1978.

[4] M.A. Rashwan. PhD, Prof. M.M.Fahmy, "New Technique for Speaker-independent Isolated-word recognition",IEE PROCEEDING, Vol. 135, Pt.F.No.3, JUNE 1988.