AN AUTOMATIC SPEAKER RECOGNITION SYSTEM

BY

YU CHUN KEI

(余振琪)

A MASTER THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT

FOR THE DEGREE OF

MASTER OF PHILOSOPHY

IN

THE DEPARTMENT OF ELECTRONIC ENGINEERING

THE CHINESE UNIVERSITY OF HONG KONG

HONG KONG

SEPTEMBER, 1989.

ACKNOWLEDGEMENTS

From the beginning of this research project, my supervisor, Dr. P.C. Ching, has been giving me lots of valuable guidance and instructions till to the finish of this thesis. Experienced suggestions and corrections with much patience have been given me from him through out the research period. Thanks are also due to Mr. M.H. Ko who has been very helpful in sharing his experience and knowledge in the technical area in the research and preparation of this thesis. In addition, I would like to give thanks to Miss W.M. Lai in the provision of her past research materials. Finally, appreciation should be given to those 3rd year students who have participated in the recording.

ABSTRACT

A fully automatic speaker recognition system, which is composed of a speaker verification (SV) module, a speaker identification (SI) module and a speaker independent isolated word recogniser (IWR), has been designed for mono-syllabic language, specifically for Cantonese. Energy-time profiles (ETP), the segmental energies extracted from the outputs of 5 consecutive bandpass filters with cutoff frequencies 150-500, 500-850, 850-1.2k, 1.2k-1.8k and 1.8k-3.2k Hz respectively, have been used as feature parameters to carry the acoustic characteristics for both speaker identity and speech contents. Due to the simple phonetic structures of Cantonese, instead of using dynamic time warping (DTW), linear time warping (LTW) is applied for aligning of the testing and reference tokens during template matching so that hardware implementation for low-cost high speed real time processing is possible.

For speaker verification, a combination of M distinct Cantonese digits is used as input utterance which at the same time contains information of the identity of the user. The claimed identity is being extracted by a speech recognition algorithm such that corresponding references can be retrieved for comparison. During verification, instead of treating the complete sequence of digits in its entirety, the input utterance is considered as different units of discrete word and comparisons are made on a digit-by-digit sequential order. Final decision depends upon the overall recognition results obtained from each digit. Using LTW for time alignment of each individual digit, a verification accuracy of 99.39% is obtained with M=5. This is comparable to that obtained by using DTW with M=3, but with a 10 fold increase in computation speed. Higher accuracy can still be possible with a larger value of M.

For identification of a speaker, the user is requested to utter a randomly selected combination of 5 digits and recognition is again carried out on a digit-by-digit approach. After comparison with the appropriate reference templates for each registered candidates, the speaker will be identified as either one of the qualified

candidates or a non-registered user by using a set of decision criteria based on a majority rule. Identification accuracy of 96.21% is obtained on using LTW. When compared with the results obtained by using DTW, only an increase of 2.61% in rejection rate of legal user is found but with the advantage of a large amount of time saving in the identification process.

Finally, a probabilistic approach is employed in the IWR algorithm for extracting a user's personal identity code from his input token in the SV system. Instead of using traditional template matching technique, decision is based on a statistical criterion. Template generation is by means of clustering the feature parameters, viz. the ETP, of a large training set on a temporal basis, and a probability matrix is computed which gives a similarity measure between the vocabulary and that of the cluster centre. Although the training process is relatively lengthy, it can be done off-line and hence real time performance will not be affected. Average recognition score up to 97.88% is attained for trained speakers which is fairly satisfactory for the prescribed application in view of the system simplicity.

Contents

Chapter 1

Introduction

Computer, being originally designed to perform simple but repetitive arithmetics, has emerged to be an essential tool for scientific development nowadays. Many of the human tasks, especially those involve extensive and complicated calculations, are performed by computers under human instructions in a fast and ordered manner. In order to cope with the demands existed in the development of different areas of science, the communication medium between human beings and computers has evolved from a machine level (using digital code) to a more accessable way (using human language in words). " Can computer communicate with man in a more efficient way?" has been an unceasing question to scientists over the past couple of decades. Speech, one the most natural and frequent medium that man communicate among themselves has been considered as one of the possible and ideal medium for man-computer communication. This is particularly true for many disabled people because speech often remains one of the most important biometric attributes they can easily access. Hence, teaching the computers to listen and understand as well as to speak are the prime goals of many researchers.

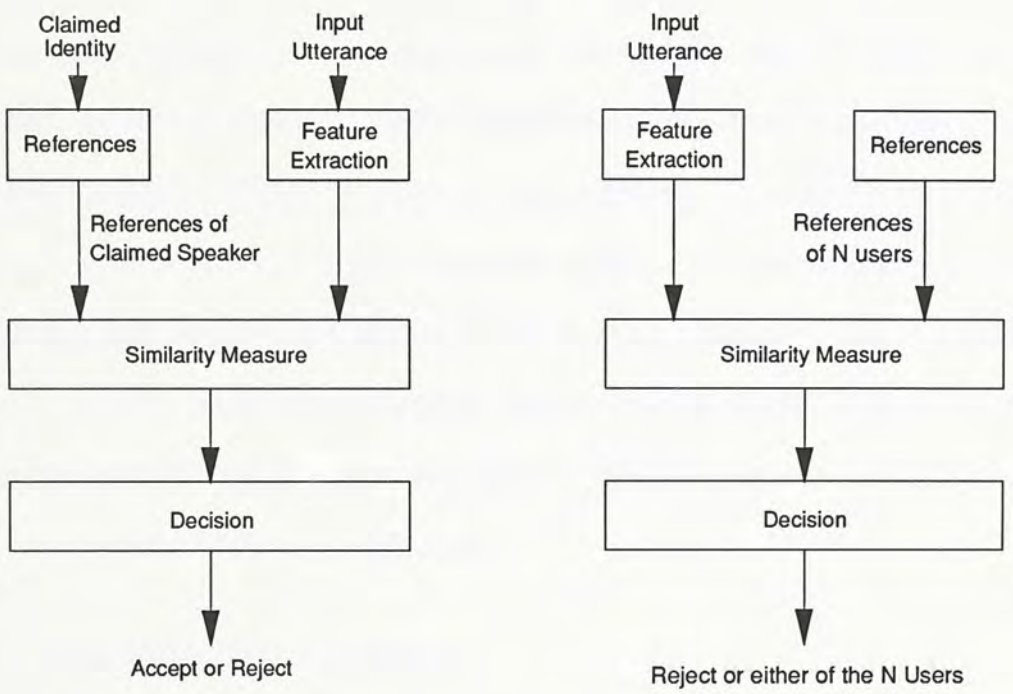
During human conversations, not only can a listener extract the language contents from the speech spoken by a speaker, it is also highly probable that he can obtain extra information from the speech such as the identity of the speaker (who is speaking), the emotion of the speaker (angry or calm, etc.) and even the sex of the speaker (a man or a woman). For more than twenty years ago, scientists have started to investigate whether computers can recognise a person by the way how he speaks as well as to understand what is he speaking. Automatic speaker recognition has then become a major research interest and enormous amount of work has already been done in this area.

In fact, to identify a person, many methods have been found useful and robust. These include the finger print recognition, retina pattern recognition and many methods along with physical attributes on a person's body. These are said to be static which cannot be easily changed even intentionally. On the other hand, signature recognition which is said to be dynamic as the performance depend upon the signer's act, has been used broadly all over the world. However, none of the above recognition methods can be performed easily under merely computer supervision, i.e. full automation. To represent one's identity in a convenient way, some artifacts such as magnetic card and pass code have been used which however, being extrinsic to the user, may be forgotten, lost or more seriously, stolen by someone else which will cause inconvenience and lost of property to the user. Human speech, the future medium for man-machine communication, has been suggested to be a potential attribute from which speaker's identity can be determined automatically by computer after processing. There are many real world applications for automatic speaker recognition such as the physical access control to some restricted area, direct access or remote access control through telephone channel to some personal and confidential information or data, forensic science for investigation of evidence to be used in the court and many others.

1.1 Classification of Speaker Recognition

With different applications in mind, speaker recognition can basically be classified into two main categories: Speaker Verification and Speaker Identification. A speaker verification (SV) system, having an utterance and an identity claim as input, discerns whether the input utterance belongs to the claimed speaker or not. Rejecting or accepting the claim are the two possible decisions. On the other hand, a speaker identification (SI) system, having only an utterance as the input item without any other information about the speaker's identity, has to determine to whom this input utterance belongs among the N legal and pre-registered users or to confirm

the speaker is not a qualified user. Therefore there are altogether N+1 possible outcomes for such a system. It has been found that the performance for speaker verification, which being theoretically independent of the size of the registered population, is always better than that of speaker identification which varies inversely to the number of legal users. Fig.1-1 shows the schematic structure of these two kinds of speaker recognition system.



(a) Speaker Verification (b) Speaker Identification

Figure 1-1 Speaker Verification and Speaker Identification

Despite of the differences between SV and SI, the extraction of feature parameters and the exploitation of the speaker-dependent acoustic events from the input utterances are the most crucial parts to both systems which directly affect the recognition performance. A successful parameter extraction process is determined by the appropriate choice of robust acoustic features from the spoken words which can

reflect distinctively the speaker's characteristies. Basically, there are two approaches in which extraction of acoustic features to characterize a specific speaker can be based upon, namely, "text-independent" and "text-dependent".

For text-independent speaker recognition system, distinctive speaker characteristics are supposed to be contained frequently and naturally in ordinally human speech irrespective to the embedded contents.Thus, no fixed text is necessary for recognition as long as the input sentence is properly uttered. The advantage of text-independent speaker recognition lies on the fact that speaker characteristics can simply be extracted from ordinary conversation so that cooperation from and pre-acknowledgement to the speaker are not necessary. Contrary to text-independent speaker recognition system, a text-dependent speaker recognition system requires the unknown speaker to cooperatively utter a fixed word, a full sentence or a combination of words, out of a confined vocabulary. In this case, acoustic features are extracted as a function of time and the speaker's characteristics are then reflected by how these time-varying features change with time.

1.2 Speaker Recognition Techniques

To recognise automatically a speaker by computer, many techniques have been developed and tested. These techniques involve various ways for (a) extracting useful speaker dependent feature parameters, (b) increasing speaker information and finally, (c) exploiting the parameter sets for recognition. Although under different emphasis, they are usually designed in accordance to the utterance nature (fixed text or not) and phonetic structures of the language employed (English, Japanese or Chinese), and with considerations of the efficiency and effectiveness for the applications in mind.

To represent a speaker by the way he speaks, many acoustic features evaluated through signal processing theories have been found to be very useful. Out of them, the linear predictive coefficient (LPC), the pitch, the gain and the short time spectrum have been used widely for both speech and speaker recognition. The LPCs, being the coefficients of an adapted model to an all-poles filter which simulates the human articulary system, has been found to be a very effective parameter set to represent speaker's characteristics. It can used either directly [1], or after undergoing some transformations [2]. Different methods for its evaluation has been proposed and useful properties in deriving more information from the speech signals have been discovered [3]. However, the computational effort in obtaining the LPC coefficients is very large and, therefore, despite of the popularity of its use in speaker recognition, economic real time application has been found very difficult, if not impossible. On the other hand, pitch (the foundamental frequency variation of the speech signal) and the gain (the intensity variation of the speech signal) are some of the features which can be extracted easily through simple methods and hardware. However, representing a speaker by the pitch or the gain of his utterance only is usually inadequate. In fact, they are often the side-products of the extraction process of other features which makes them, in many cases, only part of the parameter set in speaker recognition [4,5]. Besides these features, the short-time spectrum, showing the 3-dimensional power spectrum of the speech signal along time, is another very useful parameter to distinguish a speaker. It has been refered to be the "voice print" in analogous to the "finger print" in identifying a person and had been used semi-automatically in speaker recognition in the initial development of automatic speaker recognition techniques. Filter-bank approach to approximate the power spectrum of the speech signal is yet another method to obtain acoustic features in a simple way. However, the number of channels usually employed in the frequency bank is around 20 [6] which complicates the system configuration and requires a relatively large amount of compu-

tation or hardware. To put into low-cost real time applications, reduction in the number of channels is neccessary but which, on the other hand, will affect the recognition performance.

The extraction of useful feature from the speech signal is very crucial for automatic speaker recognition. However, the way for exploitation of these useful feature parameters is also very critical for a successful system and depends highly on whether the system is text-dependent or text-independent. One of the common ways in handling the parameters in text-independent speaker recognition system is the long-term average statistical approach [7,8]. Acoustic events extracted in certain time intervals are averaged throughout the whole input utterance to obtain some statistics which is said to be speaker dependent. However, this method requirs quite lengthy utterances to ensure stable and reliable statistics for satisfactory performance. Another frequently used method is the detection of certain phonetic events which occur frequently at different location in the utterances [9,10] and comparison will be made upon on the feature characteristics within the specified region. The location of phonetic cues, which is the determining factor for good recognition performance for this method, is however not easy to do. In addition, as a matter of facts, long duration of input utterance is required to ensure adequate existence of the events for good performance. Owing to these reasons, text independent speaker recognition has not yet reached to a prominent result that leads to practical usefulness. Although in the past few years, much effort has been contributed in the text-free speaker recognition, the performance is still lagged behind those employing a fixed text pattern as testing utterance and in fact, many practical applications are basically a text-dependent speaker recognition system [11,12].

One of the approaches in the comparison of speaker identity for a text-dependent speaker recognition system is the statistical methods [13,14]. The probability densities of the dynamic features is estimated from the training data set of a fixed

input token to represent the speaker and recognition is performed on statistical decision theory. Although only a fair amount of computation is needed to acquire the statistics, it requires quite a large number of utterances for training and this requirement is usually very difficult to satisfy in real-time operation. The template matching approach, which has been a traditional method employed in many of the speech recognition systems, is another method popularly used for text-dependent speaker recognition. In this approach, the speaker is represented by the time series of the features extracted from the training utterances of the speaker. Averaging or vector quantization [15] is usually employed to form the reference templates or codebook. The comparison will then be done between the time series of acoustic features of the reference templates and that of the input utterance with identical context. The possibility that the input words and a certain set of reference template belongs to a specific speaker will be evaluated according to the degree of resemblance between them. However, in real situation, the duration of the utterances differ from speaker to speaker and from time to time. In order to achieve a meaningful measure on the utterances' similarity, the parameter sets must be matched properly on the temporal basis. Unfortunately, the variation in the speaking rate has caused non-linear fluctuations in the speech pattern along the time axis which makes linear time warping (LTW) for time alignment becoming inadequate for satisfactory comparison. This phenomenon is more prominent in utterances of long duration and complicate phonetic structures. To cope with this problem, dynamic time warping (DTW) is introduced. With this algorithm, distinctive time matching function to achieve maximum resemblance for any two utterances is evaluated through dynamic programming under some preset constraints. Every possible functions within the possible region would be evaluated and finally the utimate solution is determined at the one giving the best matching of the two. Though the achieved alignment is satisfactory, the amount of computation involved is so large that real time application of the algorithm in speaker recognition has been kept unreachable for low cost hardware.

However, in many of the proposed speaker recognition systems, the non-linear fluctuation is serious as the utterances is formed of languages rich in phonetic variation such as English and Japanese, DTW needs to be used to achieve good recognition results [12,16].

Hong Kong, a highly industrialized and commercialized area and also a financial centre in the South East Asia, demands very much on an effective and efficient communication of data and information. Automation through computer has been therefore used enomously in nearly every areas to achieve a successful business. Security in the access of data and information, or in the entry to a restricted area then becomes a necessity. A code is popularly used to represent one's identity which is usually in the form of a password accompanied with a magnetic card. However, all these are extrinsic to the user and subjected to forgetting or being stolen by other persons. The needs for an efficient automatic system in the recognition of the user's identity are therefore existed. However, for a wide range of applications to be possible, real time but low cost becomes a "must".

A novel automatic speaker verification system, aiming at low cost real time implementation, is proposed in this project. The system uses a testing utterance composing a sequence of Cantonese words as Cantonese is the mother tongue of most of the Hong Kong residents. The template matching approach with a minimum distance measure is employed. To allow real time implementation with economic hardware possible, DTW will not be suitable for the time alignment of utterances. However, for a complete utterance consisting a sequence of Cantonese words, the variation in the duration and the non-linear fluctuation of speech pattern will be very serious from utterance to utterance. Since the intensity variation for Cantonese words is simple so that the words in a sentence can be separated without difficulty as long as the number of discrete words is known. Instead of comparing speaker's similarity with the input utterance as an entirety, each of the individual words in the input

sequence are separated and comparison is made on a discrete word basis. Since Cantonese, a mono-syllabic language, has very simple phonetic structure for each single word and therefore LTW is applied without serious degradation in the time alignment between discrete words. The short-time spectrum from a 5-channels filter bank, evaluated on each of the discrete words, is extracted to represent the speaker's identity. The name Energy-Time Profile (ETP) is given for this set of parameter. Verification decision will be based on the accumulated result of the similarity comparison of the ETP on each of the single word in the input sequence.

Moreover, the sequence of words in the input utterance is used to represent the speaker's claim of his identity simultaneously and is recognised by a built-in isolated word speech recogniser. Again, ETP is used as the feature parameter in the recognition of the words, but with a small modifications. Instead of using template matching with a minimum distance measure for the decision, a probabilistic criterion is proposed for the recognition of the words.

In order to remedy in the proposed verification system for the case that the speaker has forgot his identity code which is needed for verification, a speaker identification system is included. An input sequence composing a random combination of discrete digits is requested by the system for identification. The identity will be determined as either one of the legal users or an illegal user using an approach similar to that used in the verification system. However, a different decision strategy is employed to ensure identification accuracy. Fig.1-2 shows a block diagram of the automatic speaker recognition system. The system is evaluated on a pool of speakers with the 10 Cantonese digits as the vocabulary.

This report describes in details the above automatic speaker recognition system and experimental results will be given. In Chapter2, the way for extraction of the ETP will be descrided. The measure of speaker similarity will also be defined

together with the description of the 2 alignment techniques, DTW and LTW, used in the experiment. In Chapter 3 and 4, the details on how to verify and identify a speaker will be explained respectively and their respective evaluation results will be given. The description of the isolated word speech recognition algorithm, inluding the modification of parameter set, the training methodology and the probabilistic decision criterion, will be included in Chapter 5. Finally, a conclusion of the project and the suggested future work will be given in Chapter 6.

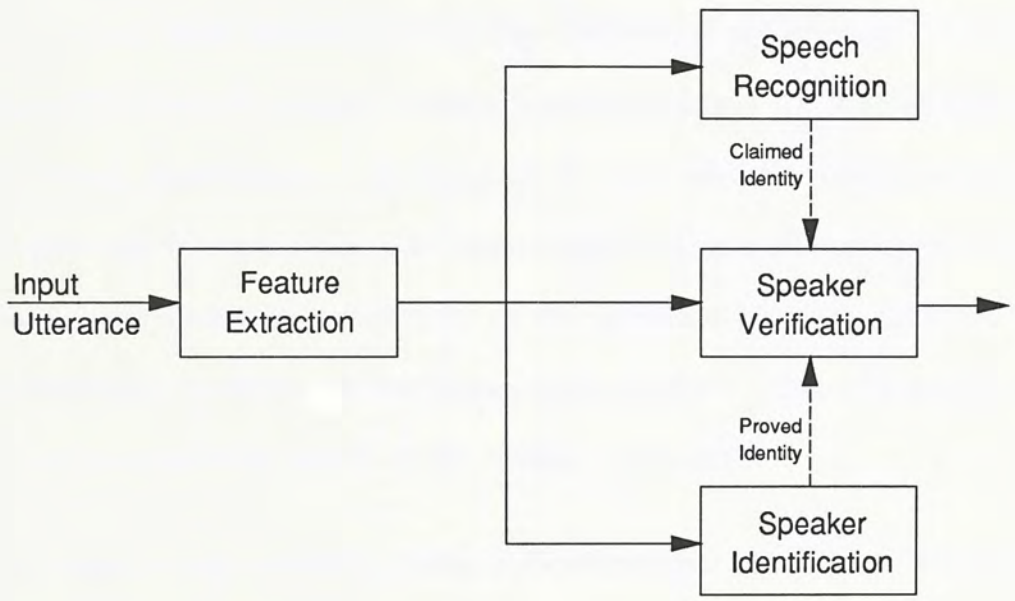


Figure 1-2 Automatic Speaker Recognition System

Chapter 2

Speaker Recognition Using Energy-Time Profiles

One of the difficulties in text-dependent speaker recognition system using temporal acoustic features is the availability of an efficient but accurate method for aligning two utterances having different durations such that meaningful comparisons can be made. Dynamic Time Warping (DTW) has been employed by many researchers to handle the alignment. However, the computation involved is laborious and it makes real time implementation very difficult, if not impossible. Linear Time Warping (LTW), which requires a much less computation for time alignment, is on the other hand, insufficient in many cases to cope with the nonlinear relationship between time and the variations of acoustic features. This nonlinearity increases when the duration and structural complexity of the utterance increase. LTW is therefore seldom employed in those speaker recognition systems using discrete or a whole sentence of poly-syllabic words as the testing utterances.

Cantonese, a very common dialect in Southern part of China as well as in many communities overseas, is composed of words only in single syllable only, usually known as monophone. Each of these mono-syllabic words has the following simple phonetic structure

$$\text{syllable} = \underbrace{(C)}_{\text{initial}} + \underbrace{V + (C)}_{\text{final}} \tag{2.1}$$

where V is a vowel and (C) is an optional consonant. Table 2-1 gives a list of the possible finals and initial consonants of Cantonese [17,18]. There are altogether 19 initials and 53 finals in Cantonese. Combined with the 9 tones of speaking, they form the whole vocabulary of phonetically possible Cantonese words - though not all of the combinations are meaningful and defined. Because of this simple structure,

the above mentioned nonlinear phenomenon that occurs in Cantonese discrete words is much smaller than those in poly-syllabic words. [19] also showed that for small variations of speech periods, DTW's performance is not overwhelming when compared with LTW. Hence, LTW is expected to perform reasonably comparable with DTW for proper time alignment for Cantonese words. A speaker recognition system is, therefore, proposed in which speaker characteristics are extracted from isolated Cantonese words. Both dynamic and linear time warpings are employed and comparisons between them have been studied.

Finals (53)											
Vowels	a			ɛ	i		ɔ	œ	u	y	
Diphthongs	ai	ɐi	ei				ɔi		ui		
	au	ɐu			iu	ou					
								œy			
Nasals only or after Vowels	am	ɐm			im						m̩
	an	ɐn			in		ɔn	œn	un	yn	
	aŋ	ɐŋ		ɛŋ	iŋ		ɔŋ	œŋ	uŋ		ŋ̩
Plosive after Vowel	ap	ɐp			ip						
	at	ɐt			it		ɔt	œt	ut	yt	
	ak	ɐk		ɛk	ik		ɔk	œk	uk		
Initials (19)											
b	d	dz	f	g	gw	h	j	k	kw	l	m
n	n	p	s	t	ts	w					

Table 2-1 The Cantonese Finals and Initials

2.1 System Configuration

Energy-time Profiles (ETP) of a word at different frequency bands are used as parameters to characterize a specific speaker in the proposed speaker recognition system. A similar approach has been used in speech recognition of Cantonese words

[20], and experimental results have indicated a distinctive dependence of this parameter on different speakers. Fig.2-1 is a block diagram of the speaker recognition system.

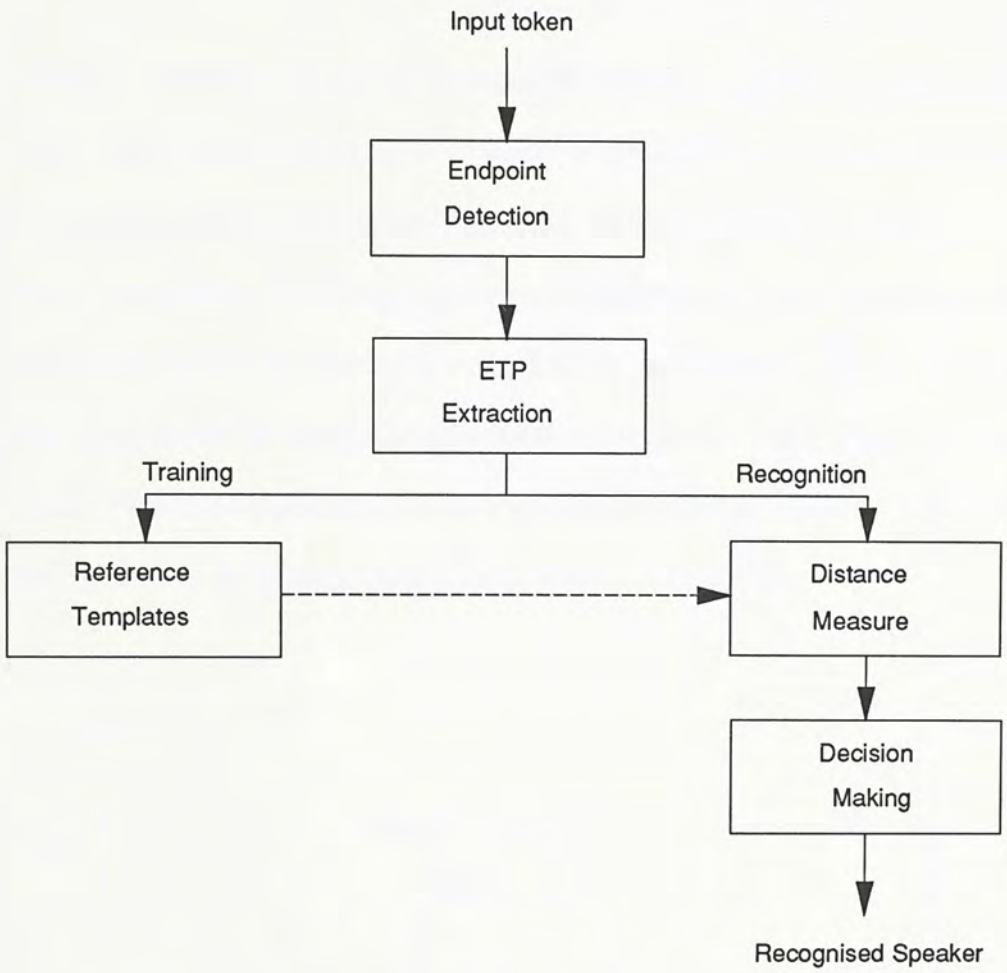


Figure 2-1 Speaker Recognition System Using ETP.

After determining the beginning and end of the utterance, the input token containing a discrete word is then passed through a series of bandpass filters and segmental energies are computed from the filters' output. In the training process, the evaluated parameter sets are stored as references. A distance threshold is also calculated from the references and will be used for decision making. In the recognition process, the

input token is subjected to a similarity measure with the reference tokens. The result will then be determined by a set of decision rules. Details of each of the functional blocks shown in the figure will be described in the following sections.

2.1.1 Endpoint Detection

In order to include all important acoustic events and also to ensure meaningful comparison of the similarity of utterances of a specific word from various speakers, the beginning and end of an utterance should be positioned as accurate as possible. In addition, elimination of useless information actually implies a saving of memory and computation which is essential in real time application. Three common techniques have been used for endpoint detection of isolated words. They are the explicit, implicit and hybrid techniques [21]. In Fig.2-2, the typical values of the energy and zero-crossing rate of a discrete Cantonese word is shown.

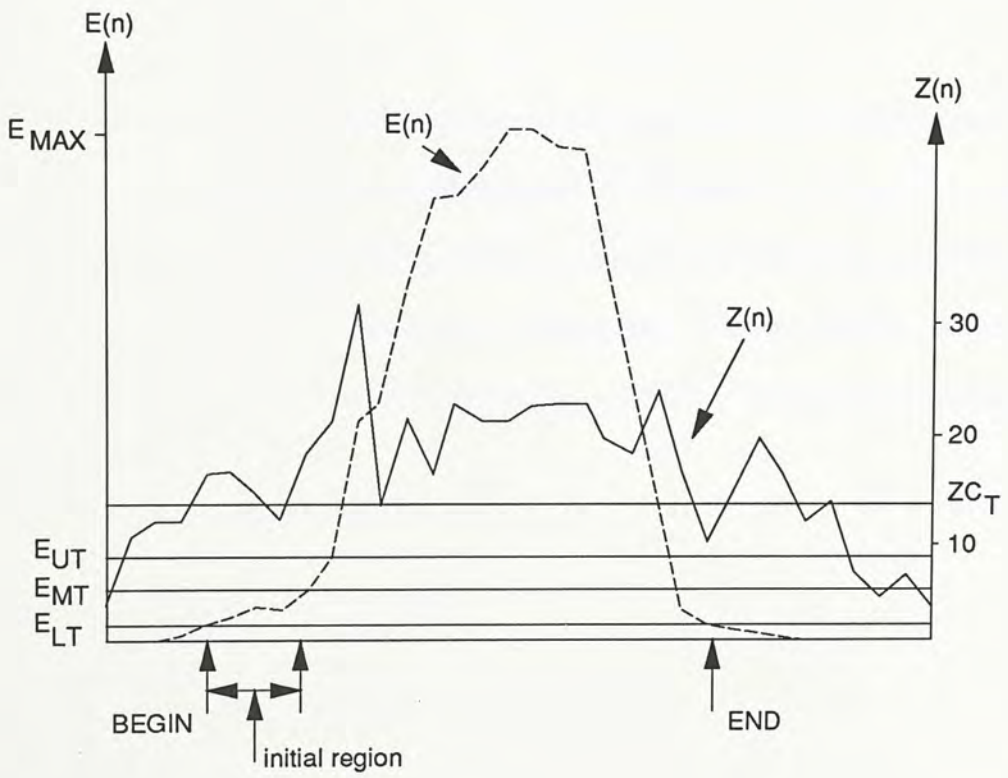


Figure 2-2 Contour of segmental energy and zero-crossing rate of a Cantonese digit "1"

The mono-syllabic characteristic of Cantonese is clearly shown by its simple shape of variation of energy with time. The explicit technique which is based merely on energy detection is therefore adequate for the positioning of endpoints of discrete Cantonese words. However, to refine the endpoints of words for those with fricative beginning and end, zero-crossing rate is used together with energy to determine the endpoints [20]. This efficient method has been used for speech recognition of discrete Cantonese words and will also be adopted in the proposed speaker recognition system.

The input speech, after bandpass filtered at 100-3.3kHz and sampled at 8 kHz with 12 bit resolution, is divided into segments of 10ms duration. The 'segmental energy' E(n) is evaluated by the 80 samples inside the nth segment in the following way:

$$E(n) = \sum_{i=0}^{80} | S_n(i) | \tag{2.2}$$

where $S_n(i)$ is the ith sample amplitude in the nth segment. The maximum 'segmental energy' is extracted among all segments and is denoted by E_{MAX}. The 'segmental energy' of 10 segments within the speechless portion (the initial region of the speech) are averaged to give the energy E_{SIL} during the silence period. Three energy thresholds are then defined as a function of E_{MAX} and E_{SIL} by the following equations:

$$E_{UT} = 0.2 \times (E_{MAX} - E_{SIL}) + E_{SIL} \tag{2.3}$$

$$E_{MT} = 0.1 \times (E_{MAX} - E_{SIL}) + E_{SIL} \tag{2.4}$$

$$E_{LT} = 0.01 \times (E_{MAX} - E_{SIL}) + E_{SIL} \tag{2.5}$$

The zero-crossing rate Z(n) in the nth segment is computed at the same time with the energy as follow,

$$Z(n) = \sum_{i=1}^{80} | Sgn[S_n(i)] - Sgn[S_n(i-m)] | \qquad (2.6)$$

$$\text{where } Sgn[S_n(i)] = \begin{cases} 1 & \text{if } S_n(i) > 0 \\ 0 & \text{if } S_n(i) = 0 \\ -1 & \text{if } S_n(i) < 0 \end{cases}$$

$S_n(i-m)$, with $m \leq 1$, refers to the nearest non-zero sample that is m samples before the ith sample. A zero-crossing threshold, ZC_T is chosen as the minimum between a fixed value of 25 and a threshold which is given by the sum of the mean (μ_{ZC}) plus twice the standard deviation (σ_{ZC}) of the zero-crossing rate during the 10 silence segments stated previously.

To locate the beginning of the utterance, the immediate segment having energy just below the threshold E_{MT} is first found by searching backward from the maximum energy segment. The beginning is then assumed as the j segment in front when one or more of the four conditions is satisfied:

1. $E(j-1) < E_{LT}$ and $E(j-2) < E_{LT}$
2. $E(j-1) < E_{LT}$ and $Z(j-1) < ZC_T$
3. $(E(j-1) - E(j)) > E_{LT}$ and $Z(j-1) < ZC_T$
4. $(VOWB - j + 1) > 10$

VOWB is the first block whose 'segmental energy' is just above E_{UT} during the backward searching process. By searching forward from the maximum energy segment, the end of the utterance is simply fixed at the last segment having 'segmental energy' just above E_{LT}. By using this simple and efficient algorithm, the endpoints of utterances containing a discrete Cantonese word can be correctly positioned. Followed will be the extraction of speaker's characteristics from these confined words.

2.1.2 Feature Extraction

Many of the parameters have been found to possess various characteristics which are speaker dependent. The pitch contour, linear prediction coefficients, spectral information and intensity can all be employed to distinguish a speaker's identity. Among them, linear prediction cofficients and their derived parameters seem to be the most commonly used. Specifically, they have been used in many speaker recognition systems and have shown to give satisfactory performance [27]. However, the evaluation of the parameter sets are too complicated which is almost impossible for real time application without expensive hardware. Pitch contour and intensity, which can be extracted in simple and efficient ways [22] are ideal to achieve high speed recognition. However, these parameter sets are usually used in conjunction with other uncorrelated parameters such as the LPC [23] and formant frequencies [24] to give satisfactory results and, therefore, system simplicity can hardly be maintained. Having considered the above tradeoff, a spectral energy approach which makes use of the short time energy of a filter bank outputs as feature parameter is defined and investigated.

The input speech sample is first sent to a bank of five bandpass filters with passband at (i) 150-500 Hz, (ii) 500-850 Hz, (iii) 850-1.2k Hz, (iv) 1.2k-1.8k Hz and (v) 1.8k-3.2k Hz. Elliptical filters having specification shown in Table 2-2 are used. The ranges of filter are fixed with the purpose to extract sufficient information of speakers from different spectral regions. The choice has been made on a compromise between speaker characteristics and computation time for recognition. Increasing the number of filters by narrowing each filter's range will increase speaker's information on one hand and computation of recognition on the other.

Channel no.	Lower Stopband Frequency (Hz)	Lower Passband Frequency (Hz)	Upper Passband Frequency (Hz)	Upper Stopband Frequency (Hz)	Stopband Ripple (db)
1	100	150	500	550	-33
2	450	500	850	900	-35
3	800	850	1250	1250	-36
4	1150	1200	1800	1850	-32
5	1750	1800	3200	3250	-32

Table 2-2 Design specification of the bandpass filters.

Short-time energy under a 20ms rectangular window are then computed every 10ms on the five output signals from these filters together with the wide-band signal. Let $E_q(k)$ be the energy of the kth segment from the qth band filter and can be computed from

$$E_q(k) = \sum_{i=-\infty}^{\infty} [S(k,i)\, W(k,i)]^2 \tag{2.7}$$

where S(k,i) is the ith sample in the kth segment and W(k,i) is a 20ms rectangular window given by

$$W(k,i) = \begin{cases} 1 & \text{if } 1 \le i \le 160 \text{ in the kth segment} \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

It has been stated that Hamming window is more appropriate for the extraction of short time information than rectangular window [25]. However, for the purpose of simple impementation and efficient evaluation, rectangular window is used in all our tests. The total number of segments varies from 1 to N where N depends on the length of the utterance. Let E_{0M} be the maximum short-time energy evaluated from the wide-band signal (i.e., q=0), the self-normalized short-time energy $\hat{E}_q(k)$ is as follow,

$$\hat{E}_q(k) = \frac{E_q(k)}{E_{0M}} \tag{2.9}$$

The speaker's identity can then be represented by a matrix composed of N time vectors. Each of the time vectors contains five normalized short-time energies in the 5 different frequency bands. Each time vector is actually an approximation of the short-time spectral energy of the utterance while its change indicates the temporal variation. The speaker similarity will then be defined equivalently as the degree of resemblance of their corresponding ETP matrices in terms of a distance measure in the proposed system.

2.2 Distance Measure

Similar to most of text-dependent speaker recognition systems, template matching approach is adopted in our system. It is therefore necessary to define a distance measure between an input token and reference tokens which are now represented in the form of time vectors of spectral energy. The distance between the input token and any reference pattern is the total vector distance between their time vectors. Smaller the distance, the greater the possibility that they are of the same speaker. In our case, using energy-time profiles as parameter, the distance d(X,Y) between utterance X and Y is defined as the sum of a distance measure d_m of the short-time normalized energy in utterance X to that in utterance Y. Hence,

$$d(X,Y) = \sum_{j=1}^{N} \sum_{i=1}^{5} d_m(E_i(j)_X - E_i(T(j))_Y) \tag{2.10}$$

where T is a mapping between the time index of utterance X and Y for proper time alignment. Many formulae have been proposed for the distance measure d_m according to different applications. Euclidean distance, defined by the formula

$$d_m(E_i(j)_X, E_i(T(j))_Y) = (E_i(j)_X - E_i(T(j))_Y)^2 \tag{2.11}$$

is one of the most common distance measure. However, in our case, the temporal changes of the energy-time profiles are so great that a few order of difference exists in the ratio between the high energy portion (around the middle of the word) and low energy portion (around the ends of the word). With the same percentage of difference, Euclidean distance, giving the sum of the squared differences, will be completely dominated by that in the high energy portion. One possible solution is to introduce normalization in the formula which can be expressed as

$$d_m(E_i(j)_X, E_i(T(j))_Y) = \frac{(E_i(j)_X - E_i(T(j))_Y)^2}{E_i(j)_X + E_i(T(j))_Y} \tag{2.12}$$

The modified formula allows a more even contribution to the overall distance from different regions of an utterance. A further simplification on the distance measure can be obtained by replacing the squared distance with an absolute distance, i.e.,

$$d_m(E_i(j)_X, E_i(T(j))_Y) = \frac{|E_i(j)_X - E_i(T(j))_Y|}{E_i(j)_X + E_i(T(j))_Y} \tag{2.13}$$

Equation (2.13) is adopted and is fairly effective in speaker recognition in which similarity is measured between contextually identical utterances among a group of speakers. With this fixed distance formula, we have to determine which of the time vector of utterance Y should be compared to that of utterance X. In simple words, the function T in equation (2.13) is to be determined to ensure a sensible comparison. Two mapping approaches are used in our system - the dynamic and linear time warping.

2.2.1 Dynamic Time Warping

Any two utterances having identical context, either from the same speaker or different speakers, are seldom to have the same duration. Unfortunately, the acoustic features for the utterances have been found to be non-linearly changing with time.

Therefore, a linear compression or expansion of temporal acoustic parameters will cause an improper time alignment between the two utterances when they are being compared. Moreover, this non-linearity varies in an undeterministic way between utterances. Dynamic programming is therefore adopted to determine an optimal matching between the acoustic features of any two utterances in the time domain.

The utterances to be aligned are first assumed to be similar. The alignment that gives greatest similarity between the utterances is to be found under a set of constraints. Different sets of constraint have been proposed [26]. The one we have used in our experiment is a foundamental one which has been used in many applications. Considering input utterance X and reference utterance Y with time indices n and m respectively, where n=1,2,...N, and m=1,2,...,M, a mapping T between the time indices n and m is to be determined so that a matching with greatest possible coincidence between them can be obtained. The following conditions guide this mapping:

(1) $1 = T(1)$

(2) $M = T(N)$

(3) $T(n+1) - T(n) = \begin{cases} 0,1,2 & \text{if } T(n) \neq T(n-1) \\ 1,2 & \text{if } T(n) = T(n-1) \end{cases}$

Conditions (1) and (2) are in fact the boundary conditions with the assumption that the beginnings and ends of the two utterances are determined accurately and should be aligned first. The remaining continuity constraint limits the function T to be monotonically increasing with a maximum slope of 2 and a minimum slope of either 0 (if the slope of the preceding frame is non-zero) or 1 (if the slope of the preceding frame is zero). Fig.2-3 indicates the possible region for the mapping T, bounded by the four lines which are evaluated according to the constraints. A typical function is also shown in Fig.2-3.

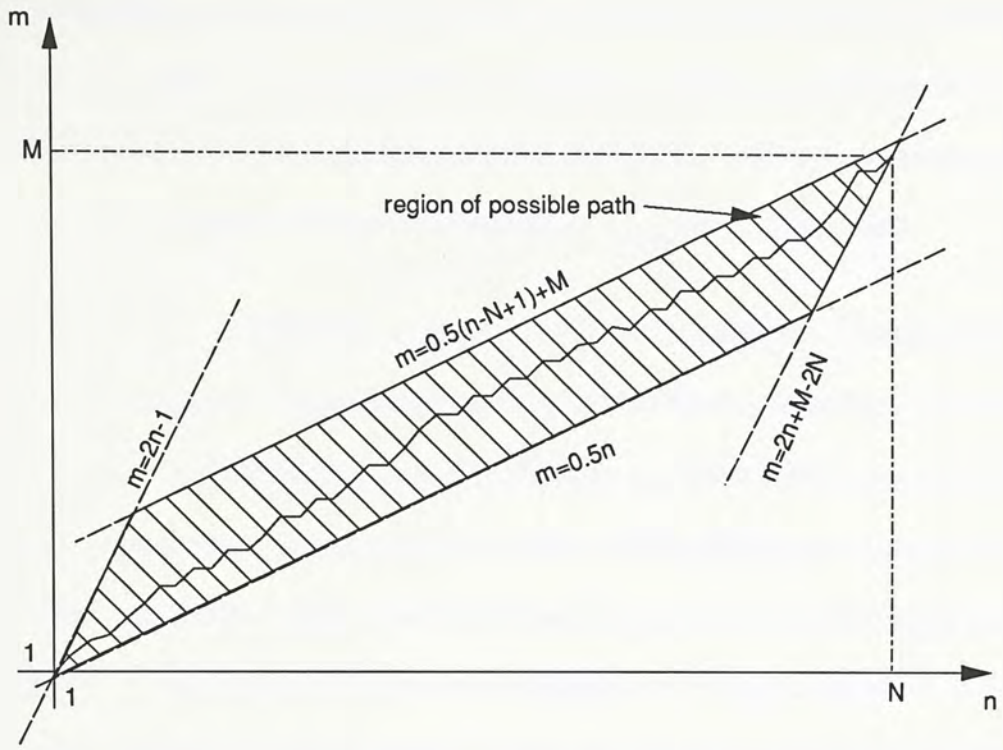


Figure 2-3 Possible region and typical mapping in DTW

Using the previous symbols, we define a distance measure d'(n,m) between the time vector of utterance X at time index n and that of utterance Y at time index m, i.e.,

$$d'(n,m) = \sum_{i=1}^{5} \frac{|E_i(n)_X - E_i(m)_Y|}{E_i(n)_X + E_i(m)_Y} \tag{2.14}$$

An accumulated distance $d_{acc}(n,m)$, measured as the minimum possible distance started from the coordinate (1,1) to the coordinate (n,m), is defined using the recursive formula

$$d_{acc}(n,m) = d'(n,m) + \text{min. value between} \begin{pmatrix} d_{acc}(n-1,m), \\ d_{acc}(n-1,m-1), \\ d_{acc}(n-1,m-2) \end{pmatrix} \tag{2.15}$$

A value of infinity is set to $d_{acc}(n-1,m)$ if the slope in the frame of time index n-1 is zero. All $d_{acc}(n,m)$ is not defined in the formula if the coordinate (n,m) is out of the bounded region. In fact, all the possible path for T will be evaluated and a final one giving the minimum distance denoted by $d_{acc}(N,M)$, is selected.

Owing to the continuity constraint, the ratio between utterances' length is limited to 2. However, because of its simple syllabric structure, the duration of Cantonese word is usually short. It is therefore common for utterances having duration ratio greater than 2. In order to allow the application of previous dynamic programming for the time alignment on Cantonese discrete words, linear compression or elongation of the time axis is performed ahead of the application of the algorithm whenever this situation occurs. This is done by transforming the short-time spectral energy $E_i(n)$ of time index n, to $E'_i(n')$ of time index n' using the following formula:

for compression

$$E'_i(n') = E_i(n) \qquad \text{with } n=2n'-1 \qquad (2.16)$$

for elongation

$$E'_i(n') = E_i(n) \qquad \text{with } n=(n'-1)/2 \text{ and } n' \text{ odd} \qquad (2.17)$$

and

$$E'_i(n') = \frac{E'_i(n'-1)+E'_i(n'+1)}{2} \qquad \text{with } n' \text{ even} \qquad (2.18)$$

where i=1,2,...,5, n=1,2,...,N and m=1,2,...,M. The transformed set of short-time energy $E'_i(n')$ having time index n' is then used in the dynamic programming to obtain maximum coincidence with the other set of the matching utterances.

2.2.2 Linear Time Warping

It is not difficult to understand from the recursive procedure according to equation (2.15) that the amount of computation for DTW is enormous. There is no definite function to evaluate the amount of computation. However, the time for computation will definitely increase with M, N and their ratio as it approaches 1. DTW is therefore not suitable to be used in real time application for speaker recognition. As stated previously, the simple structure and the short duration of discrete Cantonese word provide a chance for LTW without serious degradation of the speaker's information contained in the utterance.

LTW is implemented in the following way. For each pair of utterances, despite of their length, their time axes are warped on to the same time axis with a fixed number of time vectors, say P. As to the previous utterances X and Y which have N and M time vectors respectively, they are now transformed to have P time vectors containing short-time spectral energies $E_i(p)_X$ and $E_i(p)_Y$ which are given by

$$E_i(p)_X = E_i(n)_X + K \cdot [E_i(n+1)_X - E_i(n)_X] \tag{2.19}$$

$$\text{with} \quad n = \text{integral value of} \left[\frac{(p-1)(N-1)}{(P-1)} + 1\right] \tag{2.20}$$

$$\text{and} \quad K = \frac{(p-1)(N-1)}{(P-1)} + 1 - n, \qquad \text{a real constant} \tag{2.21}$$

and

$$E_i(p)_Y = E_i(m)_Y + k \cdot [E_i(m+1)_Y - E_i(m)_Y] \tag{2.22}$$

$$\text{with} \quad m = \text{integral value of} \left[\frac{(p-1)(M-1)}{(P-1)} + 1\right] \tag{2.23}$$

$$\text{and} \quad k = \frac{(p\text{-}1)(M\text{-}1)}{(P\text{-}1)} + 1 - m, \qquad \text{a real constant} \tag{2.24}$$

The time mappings defined by equation (2.20) and (2.23) are shown in Fig.2-4.

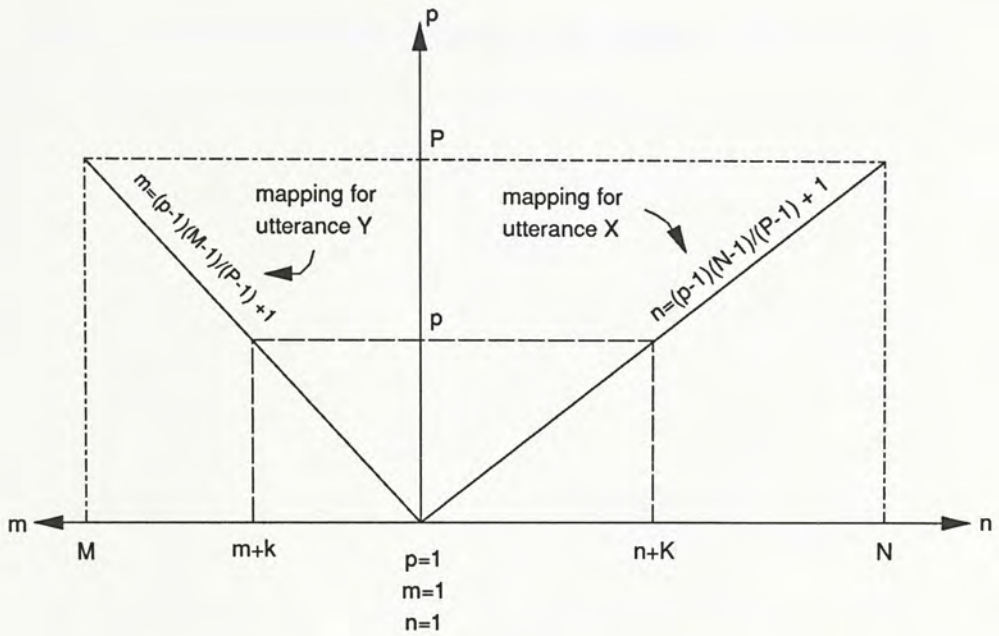


Figure 2-4 Linear mappings of two different length utterences to a fixed duration.

Consequently, the distance d(X,Y) between utterances X and Y is calculated according to the equation

$$d(X,Y) = \sum_{p=1}^{P} \sum_{i=1}^{5} \frac{| E_i(p)_X - E_i(p)_Y |}{E_i(p)_X + E_i(p)_Y} \tag{2.25}$$

which involves a definite amount of computation for each distance measure.

With the ETP representing the phonetic characteristics of a speakers, a template matching approach on a minimum distance measure according to equation (2.13) is adopted in the recognition of speakers. In the time alignment of input tokens and the reference templates, LTW obviously requires much less computation time and

is, therefore, employed for the function T in the distance measure defined in equation (2.13), with P being selected to be 16, in order to allow real time processing. However, the experimental results using DTW will be presented together so that the comparison on their corresponding performance can be studied. In chapter 3, we will studied how the ETP and distance measure are used in the verifying a speaker. Identification of speaker will be studied in the chapter following it.

CHAPTER 3

Speaker Verification System

To verify a speaker, two inputs are required, namely, the identity claimed by the speaker and the utterance that carries his characteristics for verification. However, in many systems, identity claim is sometimes extracted by some other means such as magnetic card and keypad while the input speech consists merely a specific sentence or a combination of words, not necessarily meaningful. To facilitate a thorough speech automation, a novel speaker verification (SV) system is therefore proposed in which the speaker's identity is extracted simultaneously from the input token. The contents of the input token can be a sequence of discrete Cantonese words which represents the speaker's identity in the form of either his name or his code number. Unlike many systems that carry out template matching on the whole speech sentence, the proposed system performs matching on a discrete word basis and each individual word contained in an input utterance, in this case, a monophone, is considered as one entity. Each input token is either partitioned into units of discrete word by an end-point detection algorithm that makes use of the simple energy variation of Cantonese, or simply by uttering the sentence in a word-by-word sequential manner. Energy-time profiles are extracted from each of these discrete entities and are used as speech parameters to represent individual speaker's characteristics.

To derive the speaker identity from the contents of the input token, speaker independent isolated word recognition (IWR) is required. There are many IWR methods which use different parameter sets for recognition, however, to maintain system simplicity, ETPs are again employed to represent the speech features. Instead of using Euclidean distance for template matching, a probabilistic approach [28] is investigated which has been found effective for discrete utterance recognition with

a small vocabulary. The details of the algorithm will be given in chapter 5. In this chapter, we shall concentrate on discussing the method for verifying the identity claim (which is supposed to be correctly recognised here) based on the user's input utterance of mono-syllabic words. A block diagram of the system configuration is depicted in Figure 3-1. In the following sections, the topology of defining distance measure for matching and the rules of decision will be given. Finally, statistic calculation, tests for system evaluation as well as experimental results will be described.

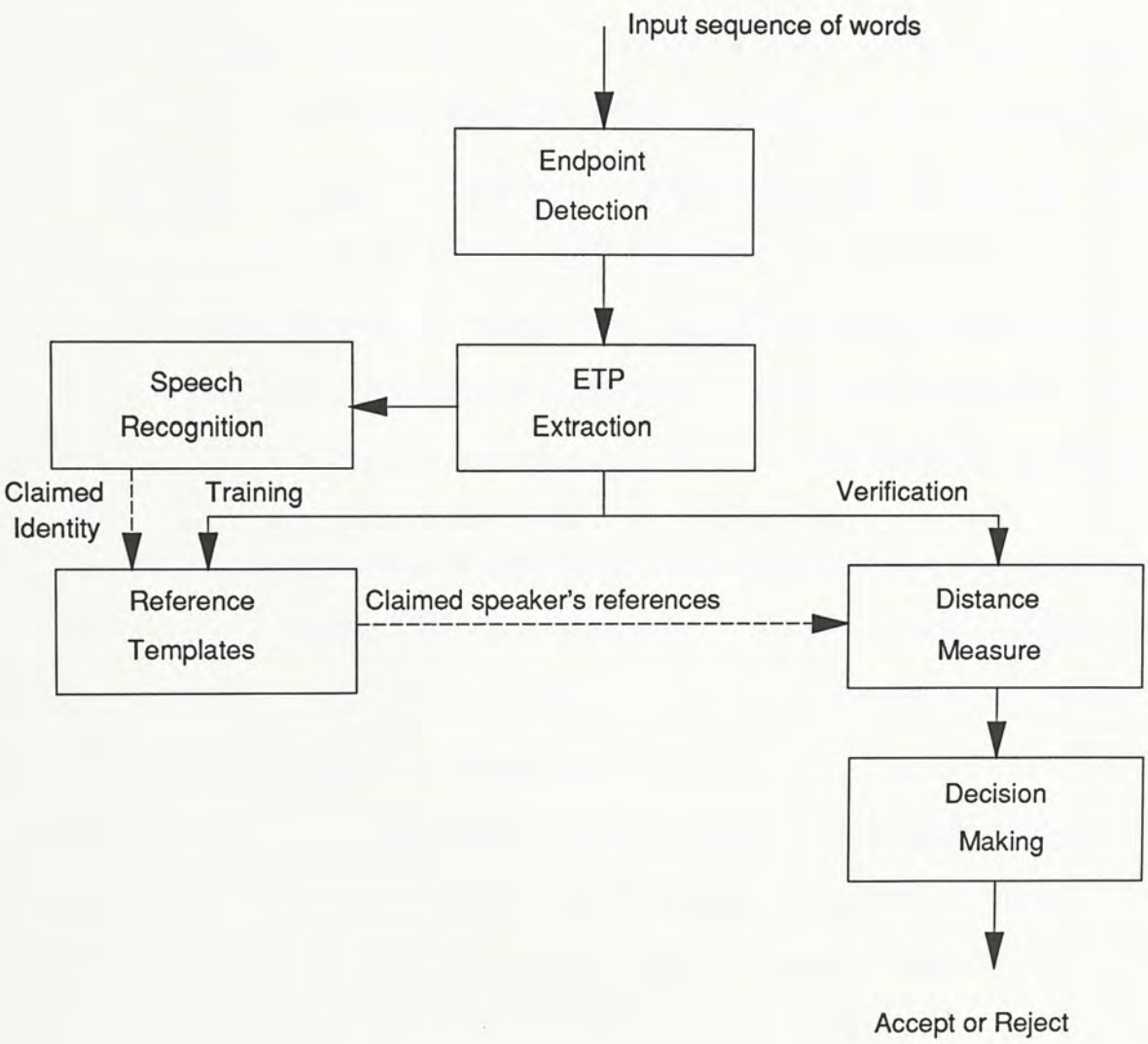


Figure 3-1 System Configuration of the proposed speaker verification system.

3.1 Template Matching

To verify a user (anyone who comes forward to request for a verification) from his utterance, the proposed system employs a simple template matching approach. The input words composing the utterance will be compared with the references of the claimed speaker and the determination to accept or reject will be made on the overall results obtained for each of these input entities. Each of the candidates, i.e. the registered users, are requested to utter each word in the vocabulary for a few times as training references and the energy-time profiles will be extracted to form the ETP matrices. No clustering or averaging is performed on these ETP matrices to from the reference templates and instead, all of them will be used directly as references during verification. Consider a system consisting of S candidates with each of them utter N times for each word in a vocabulary size W, the total number of reference templates in the form of ETP matrices in the system memory will be NxSxW. Of course, the memory requirement might be excessingly large if there are a lot of users and also if the vocabulary size is big, so one must be careful to consider the tradeoff between a complicated training process or a relatively larger memory.

During verification, each of the words composing the input utterance will be matched to the corresponding reference patterns of the claimed speaker in a word-by-word basis. Suppose one of the words in the input utterance is designated by "m", it will be subjected to a distance measure with the N templates of the claimed speaker of the same word "m" according to equation (2.15) or (2.25). Among these N distances, the minimum is selected and is defined as the smallest distance D(m) of this testing word "m" to the claimed speaker's references. That is,

$$D(m) = \min[d_i(m), \quad i=1,2,...,N] \tag{3.1}$$

Where min[ ] is a minimum selection operation amongst the arguments within the bracket. Suppose an input sequence is consisting of M words designated by "1", "2", ..., "M", the above procedure will then be applied to all of these M words. M smallest distances will, therefore, be obtained correspondingly and at last, a final average distance, D_{FA}, which gives the closest assemblance of the input utterance to the claimed speaker's references, is computed by averaging the M smallest distances, i.e.,

$$D_{FA} = \frac{1}{M} \sum_{i=1}^{M} D(i) \tag{3.2}$$

The evaluation of the M smallest distances D(m) and the final average distance D_{FA} is shown in Figure 3-2. Finally, in the decision making process, D_{FA} will be compared with a preset threshold to determine whether to accept or reject the claim made by the user.

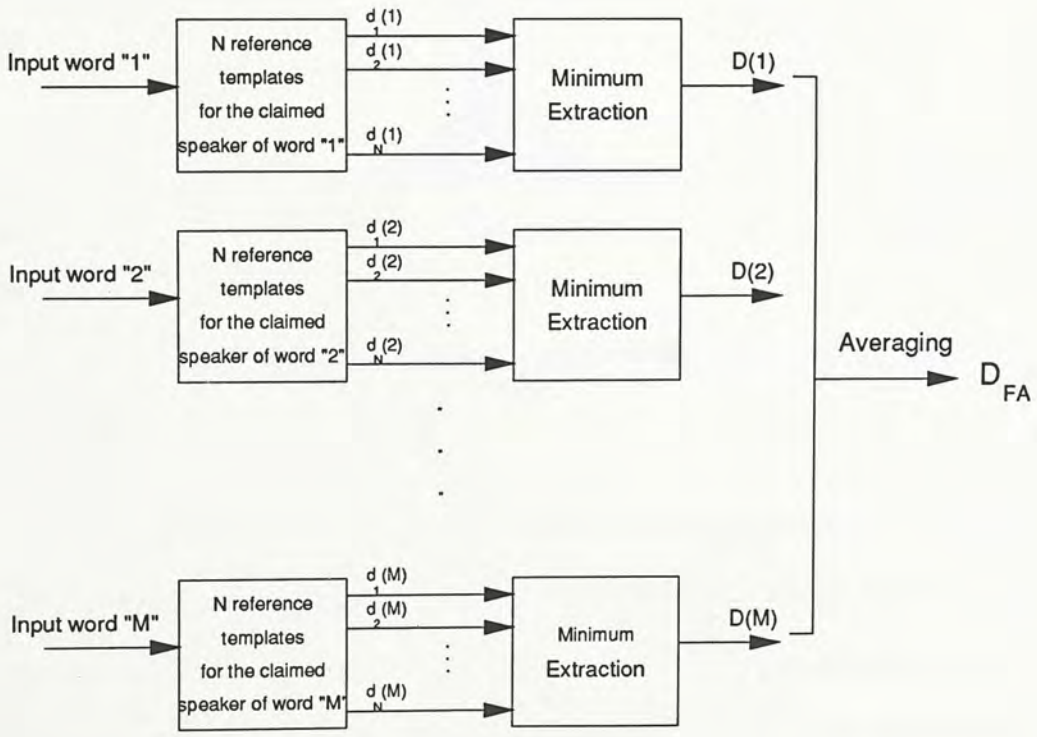


Figure 3-2 The evaluation of the smallest distances and the final average distance.

3.2 Decision Making

Although utterances of the same word made by the same speaker are quite similar, the measured distance, or the so called intra-speaker distance, between any two of them are expected to follow a Gaussian distribution. On the other hand, the distances, or the inter-speaker distances, between utterances of the same word made by different speakers are also expected to have the same type of distribution, but with a greater mean and deviation. Figure 3-3 shows an example of typical intra-speaker and inter-speaker distance distribution.

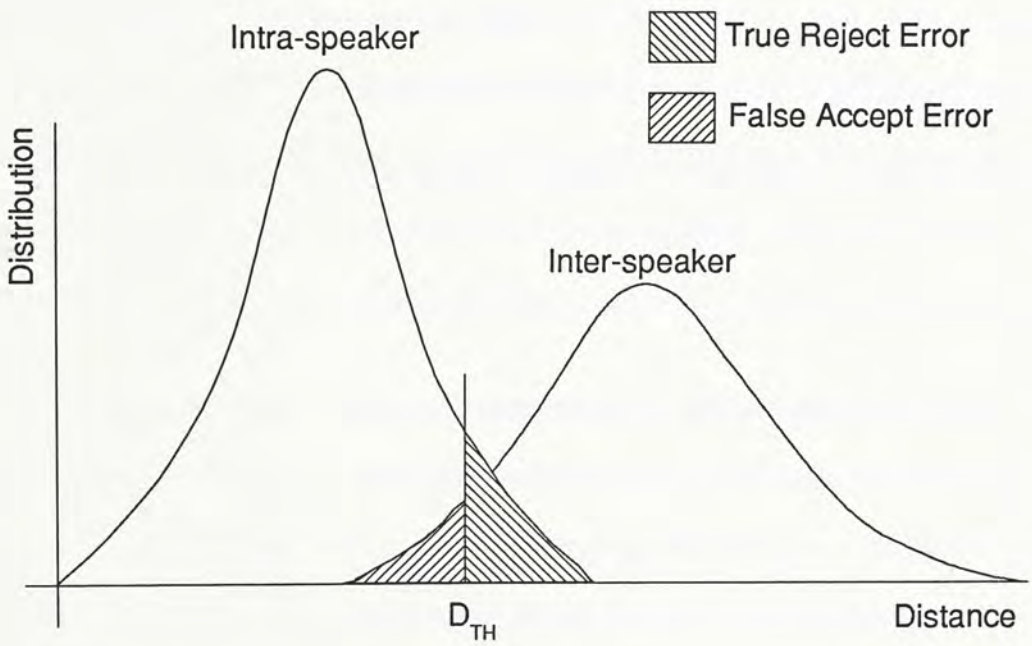


Figure 3-3 Typical intra-speaker and inter-speaker distance distribution.

A distance threshold D_{TH} is selected such that the claims will be rejected if the measured distance between testing and reference tokens is greater than this threshold, otherwise the claim will be accepted. The expected errors in rejecting a true claim together with that in accepting a false claim are indicated by the shaded regions as shown respectively. This D_{TH} is usually fixed at a compromise between the false

accepting error and the true rejecting error. The larger the D_{TH}, the greater will be the false accepting error but the smaller the true rejecting error. It is therefore not difficult to understand that the error rate for a SV system is directly related to the area of the overlapping region and which is in turn depends on the effectiveness of the parameter set in distinguishing each speaker's identity as well as how the distance measure is defined. The smaller the overlapping region, the higher the verification accuracy that can be possible.

For the proposed SV system, ETP matrices are used to represent speaker's characteristics while the distance measure is made the final average distance D_{FA}. A similar intra-speaker and inter-speaker distance distribution under this condition is expected. However, before the investigatine the intra- and inter-speaker final average distance distribution, the distributions on the smallest distance D(m), which compose D_{FA}, on each of the single word have first been studied. This will gives us some insight on various aspects of system design to achieve satisfactory performance.

With no exception, the intra- and inter-speaker smallest distance using ETP on each of the single words had also followed the same Gaussian distribution. Figure 3-4(a) to 3-4(j) and Figure 3-5(a) to 3-5(j) in page 40-43 show the distributions of the intra- and inter-speaker smallest distance on the ten Cantonese digits using DTW and LTW respectively. These curves will be described in detail in the next section and the results will be used to estimate analytically the distribution of D_{FA} with different number of words in the input sequence.

To simplify the derivation by making use of the distributions shown in Figure 3-4 to 3-5, we first assume, without loss of generality, that the distribution of intra-speaker and inter-speaker smallest distance D(m) be the same on each of the single words. Let the mean and standard deviation of the intra-speaker distance distribution be $\mu^{(1)}_{intra}$ and $\sigma^{(1)}_{intra}$ while that of the inter-speaker distance distribution

be $\mu^{(1)}_{inter}$ and $\sigma^{(1)}_{inter}$ respectively. The superscript (1) indicates that the distance is measured on a single word. Consider the intra-speaker distance distribution with distances being measured as D_{FA} over M words, the mean $\mu^{(M)}_{intra}$ and the standard deviation $\sigma^{(M)}_{intra}$ are then given by

$$\mu^{(M)}_{intra} = \mu^{(1)}_{intra} \tag{3.3}$$

and

$$\sigma^{(M)}_{intra} = \frac{\sigma^{(1)}_{intra}}{\sqrt{M}} \tag{3.4}$$

Similarly, for inter-speaker distance distribution,

$$\mu^{(M)}_{inter} = \mu^{(1)}_{inter} \tag{3.5}$$

and

$$\sigma^{(M)}_{inter} = \frac{\sigma^{(1)}_{inter}}{\sqrt{M}} \tag{3.6}$$

Obviously, the above formulae indicate that both the distribution of intra-speaker and inter-speaker distance on D_{FA} of M words will stay at the same mean but with a smaller deviation, i.e., a narrower shape, as M increases. This implies that the area of the overlapping region will decrease with M and, therefore, higher verification accuracy can be achieved. Or simply, in defining the distance by D_{FA} over M words, more information is obtained in distinguishing between speakers than measured merely on a single word. Although the derivation is purely theoretical, the investigations done which will be described in a later section did confirm the above hypothesis on D_{FA} over M words, where M runs from 1 to 5.

However, increasing the number of words in the input sequence will not only allow a higher verification accuracy, but the computational time for verification will also be increased. The selection of M is therefore based on the application and the accuracy required. System evaluation using different M has been done on a data base and will be described in the following section.

3.3 System Evaluation and Results

The above SV system is evaluated on a data base consistion of 6 male speakers and 5 female speakers having the vocabulary composed of the ten Cantonese digits from "1" to "10". Twelve times of each of the digits were uttered separately by each speaker in a quiet chamber and were recorded through a microphone. The tokens were band-passed at 100-3.3kHz to simulate the telephone speech quality and digitized to 12 bit resolution at 8kHz sampling rate. After endpoints detection, energy-time profiles were extracted from each token containing a single digit only. 5 utterances were selected as training references from each speaker for each digit, i.e., N=5, while the remaining 7 utterances are used as testing data.

Statistics on the intra-speaker and inter-speaker distance distribution were carried out for each of the ten digits on the eleven speakers. Four different sets of training and testing utterances were used for the statistics. Labelling the 12 utterances from each speaker for each word by U1, U2,..., U11 and U12, the four sets were selected arbitrary and were recorded in Table 3-1. The purpose of chosing 4 different sets of training and testing data is to obtain a more thorough and unbiased distribution after averaging the results from each set.

Trial no.	Reference set utterances
1	U1, U2, U10, U11, U12
2	U4, U5, U6, U7, U8
3	U1, U4, U5, U6, U9
4	U2, U3, U6, U9 ,U12

Table 3-1 Reference sets for distance statistics.

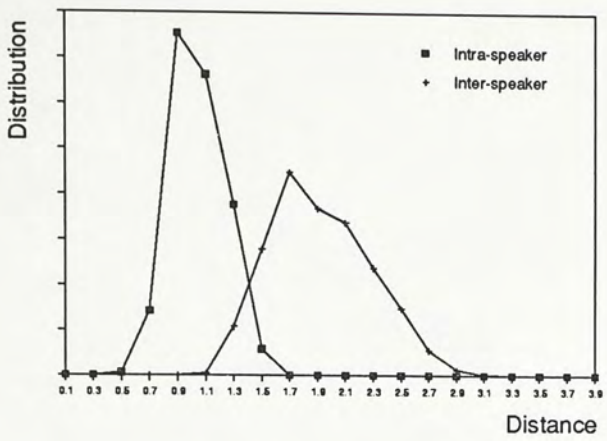
Each of the utterances from the testing set was used as an input token and was compared with the references of the same digit for all the speakers. The smallest distances for a certain input digit to each of the 11 speakers were extracted. The one measured from the same speaker was used as an entry to the intra-speaker distance distribution while the remaining 10 were used as entries to the inter-speaker distance distribution. After the 4 trials, there were altogether 308 (4x11x7) intra-speaker distance entries and 3080 (4x11x7x10) inter-speaker distance entries for each of the 10 digits. Each of these distance entries were classified into groups with distance range 0.2 and the number of entries for each groups were counted. Both DTW and LTW have been used for time alignment during the statistic calculation. The results of the distributions were plotted and shown in Fig.3-4(a) to 3-4(j) for DTW, Fig.3-5(a) to 3-5(j) for LTW, on the 10 digits respectively.

Besides the distributions on each single digit, the intra-speaker and inter-speaker distance distributions for D_{FA} averaged over M words have also been performed. Using the four trial sets stated before, all the possible combinations of M distinct digits from the vocabulary have been tried. For M=1, 2 and 3, all the possible combinations out of the 7 testing utterances of each word for a fixed digit pattern were tried while for M=4 and 5, only 500 random combinations of utterances in

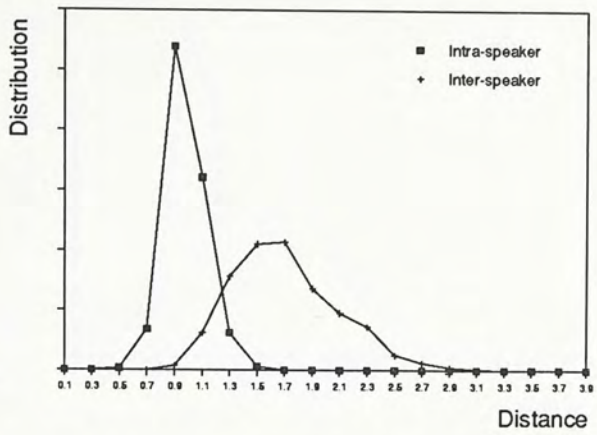
each trial has been used for the sake of simplicity. Table 3-2 lists the combination and the total sample size for each distribution statistic calculation. Similarly, the measured distance D_{FA}, both intra-speaker and inter-speaker, were classified and counted. The distribution for M=1,2,...,5 were plotted and shown in Fig.3-6(a) to 3-6(e) for DTW and Fig 3-7(a) to 3-7(e) for LTW.

No. of digit used (M)	No. of digit combinations	No. of actual / maxiu-mum utterance combina-tions	Total intra- / inter-speaker distance sample
1	10 ($_{10}C_1$)	7 / 7	3080 / 30800
2	45 ($_{10}C_2$)	49 / 49	97020 / 970200
3	120 ($_{10}C_3$)	343 / 343	1811040 / 18110400
4	210 ($_{10}C_4$)	500 / 2401	$462x10^4$ / $462x10^5$
5	252 ($_{10}C_5$)	500 / 16807	$5544x10^3$ / $5544x10^4$

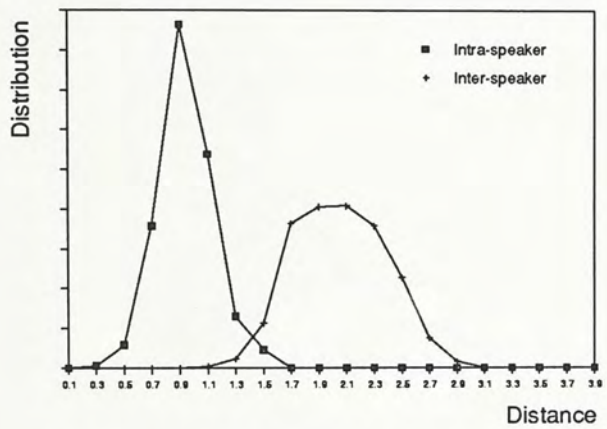
Table 3-2 Intra-speaker and Inter-speaker distance statistics sample size for different no. of digit used.



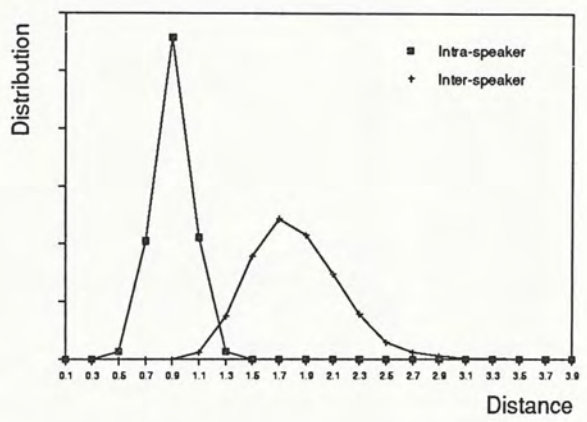
(a) Digit "1"



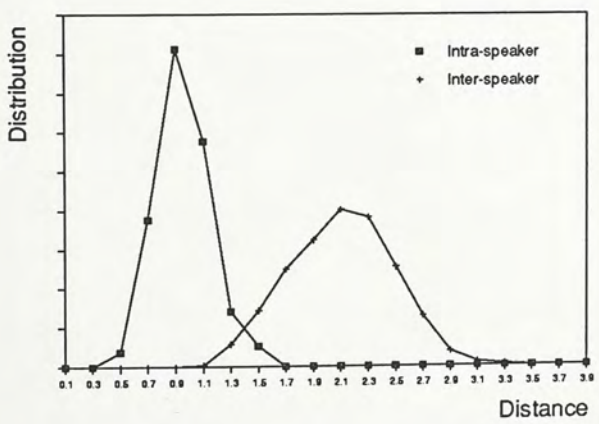
(b) Digit "2"



(c) Digit "3"

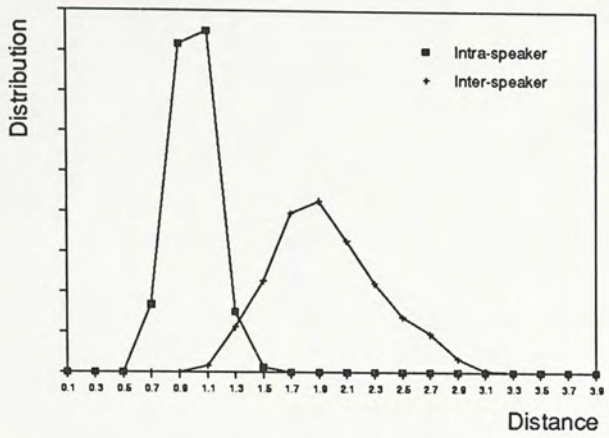


(d) Digit "4"

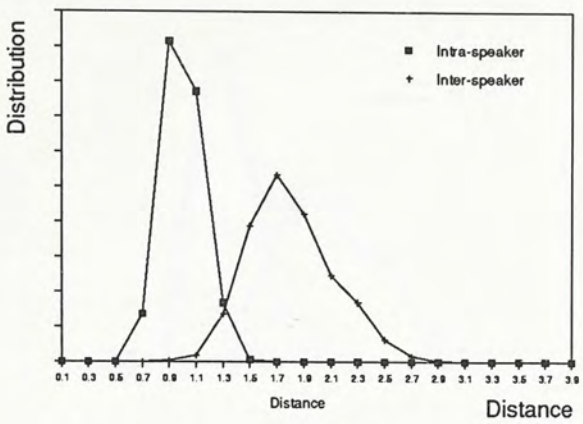


(e) Digit "5"

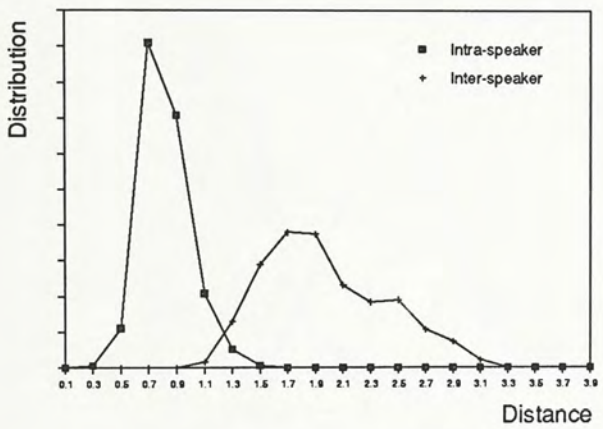
Figure 3-4 Intra-speaker and Inter-speaker distance (DTW) distribution for the 10 Cantonese digits.



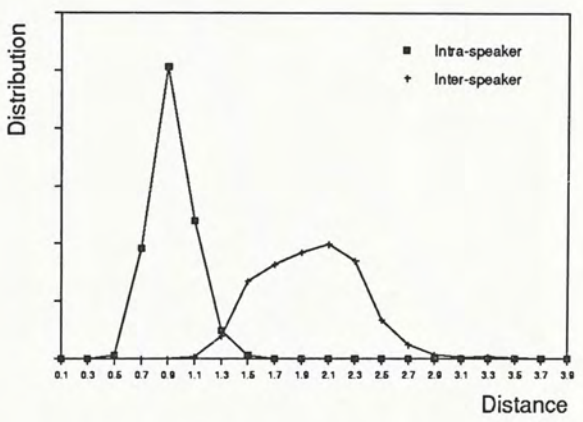
(f) Digit "6"



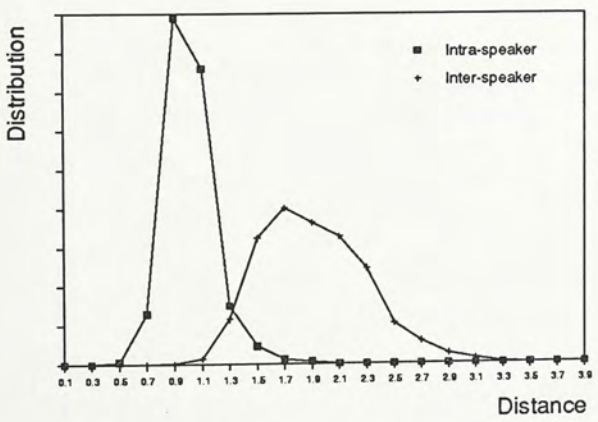
(g) Digit "7"



(h) Digit "8"

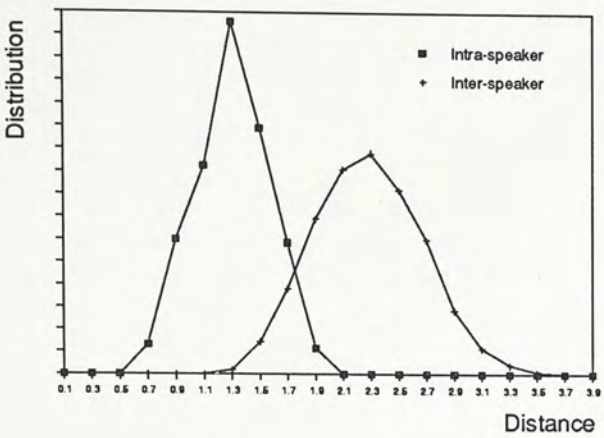


(i) Digit "9"

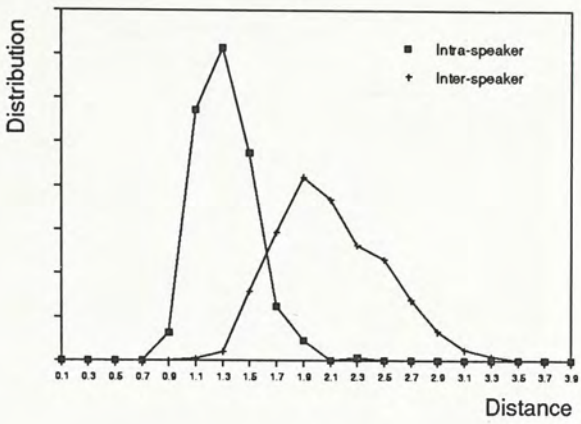


(j) Digit "10"

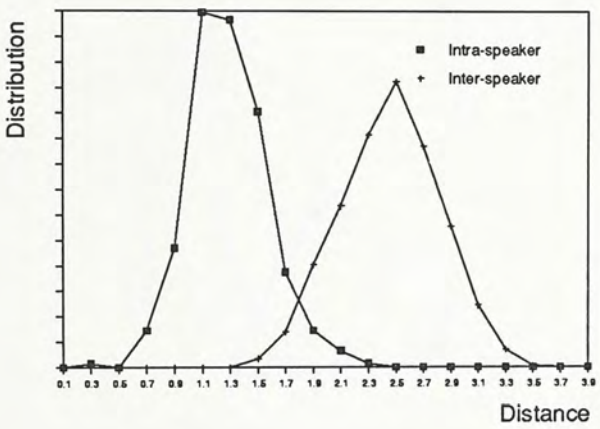
Figure 3-4 Intra-speaker and Inter-speaker distance (DTW) distribution for the 10 Cantonese digits.



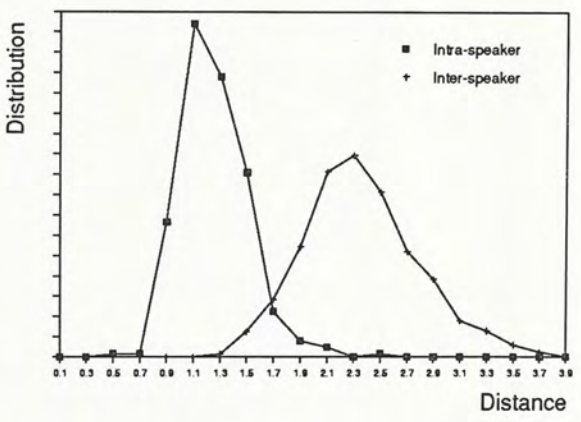
(a) Digit "1"



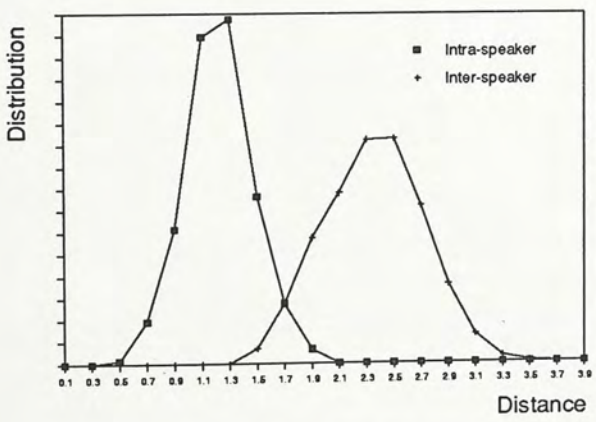
(b) Digit "2"



(c) Digit "3"

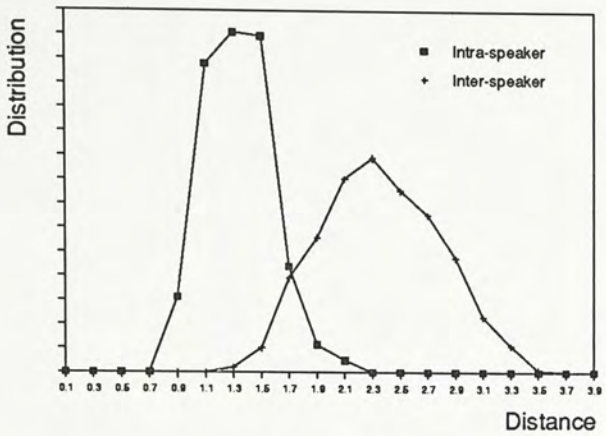


(d) Digit "4"

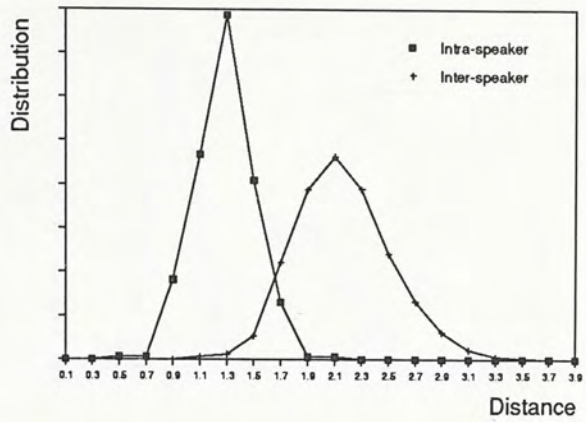


(e) Digit "5"

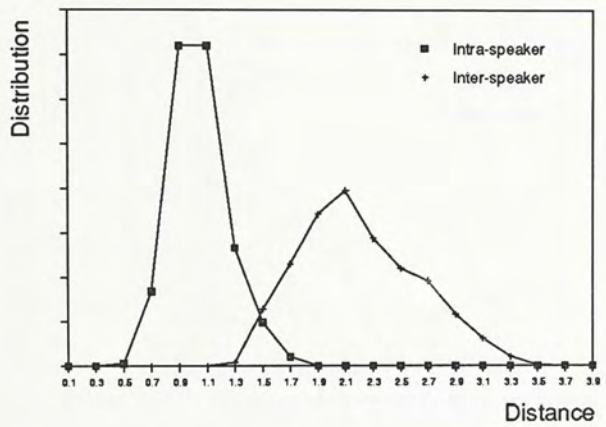
Figure 3-5 Intra-speaker and Inter-speaker distance (LTW) distribution for the 10 Cantonese digits.



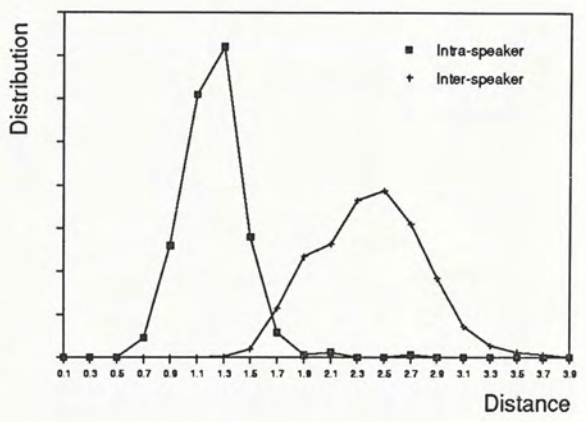
(f) Digit "6"



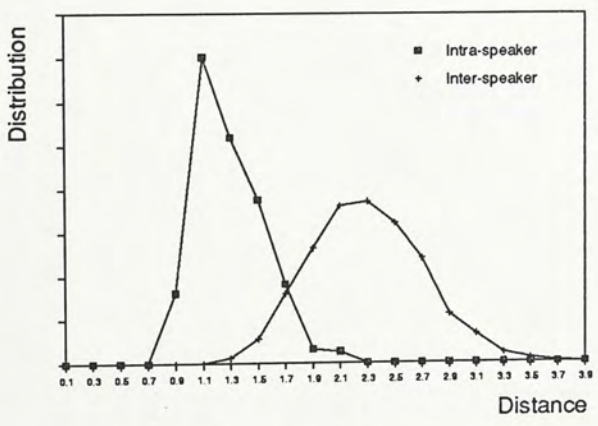
(g) Digit "7"



(h) Digit "8"

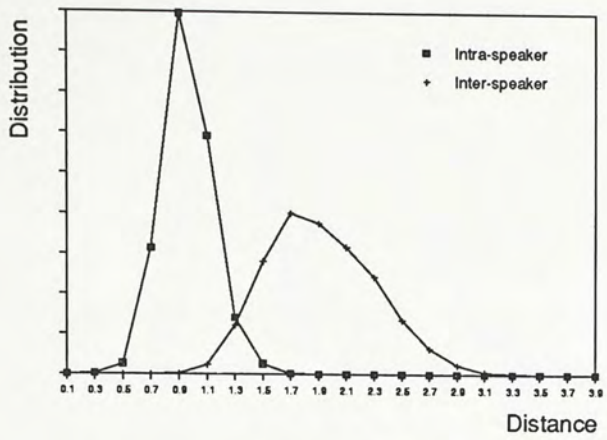


(i) Digit "9"

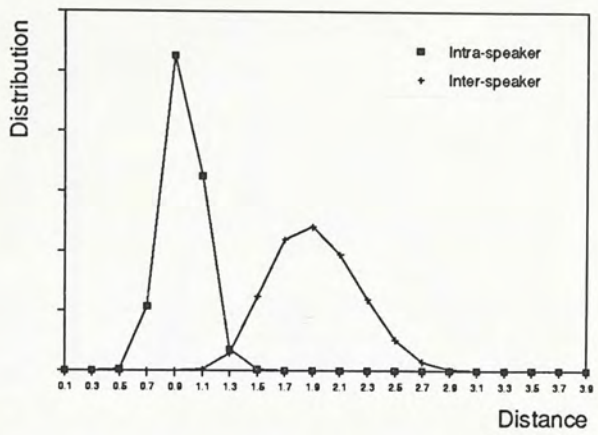


(j) Digit "10"

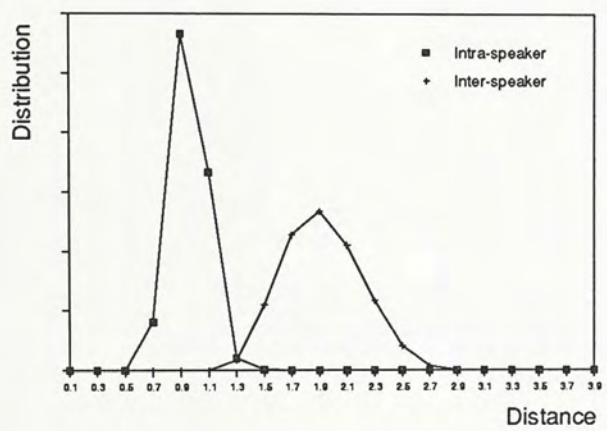
Figure 3-5 Intra-speaker and Inter-speaker distance (LTW) distribution for the 10 Cantonese digits.



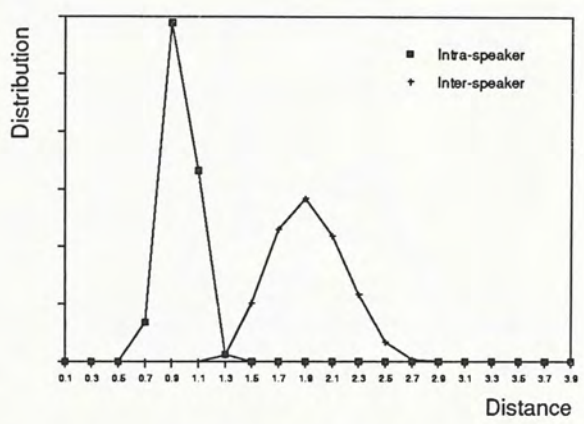
(a) Using 1 digit



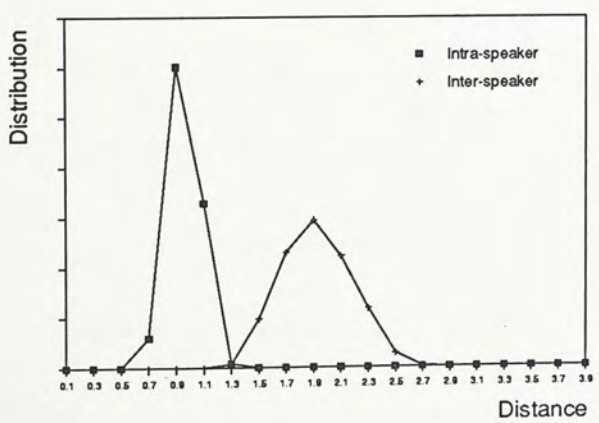
(b) Using 2 digits



(c) Using 3 digits

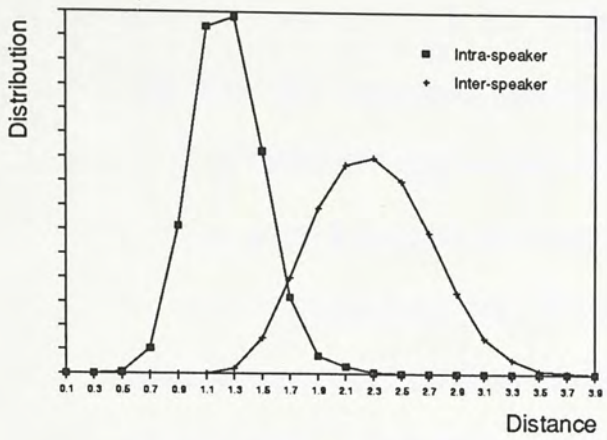


(d) Using 4 digits

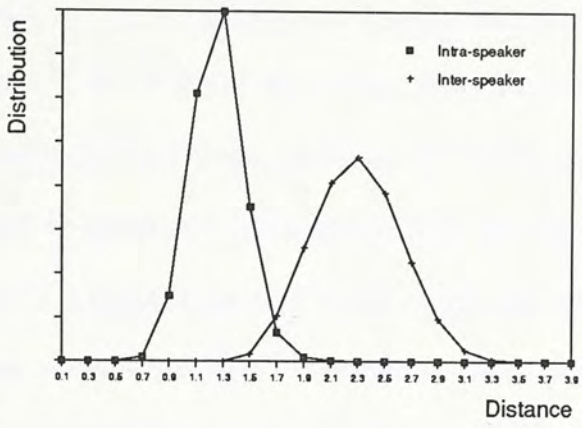


(e) Using 5 digits

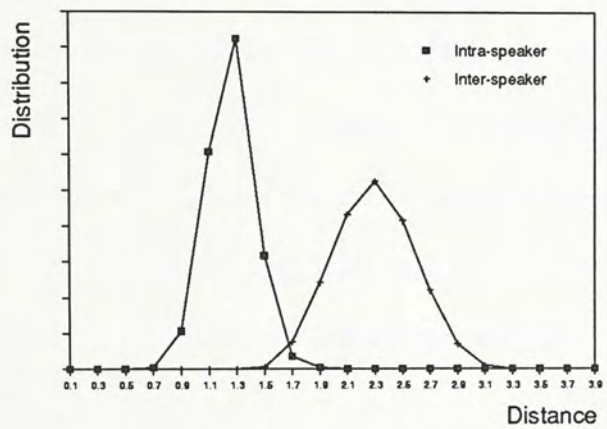
Figure 3-6 Intra-speaker and Inter-speaker distance (DTW) distribution employing different number of digits as speaker identity.



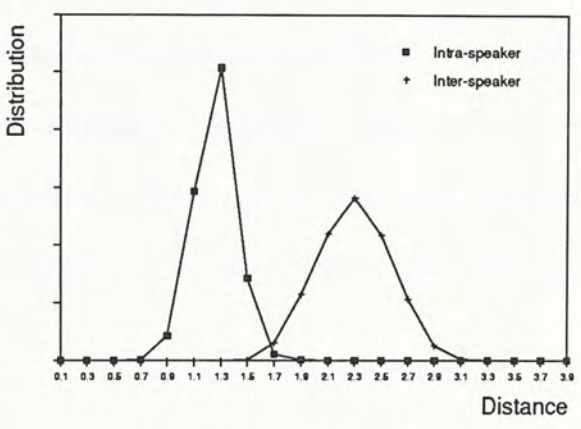
(a) Using 1 digit



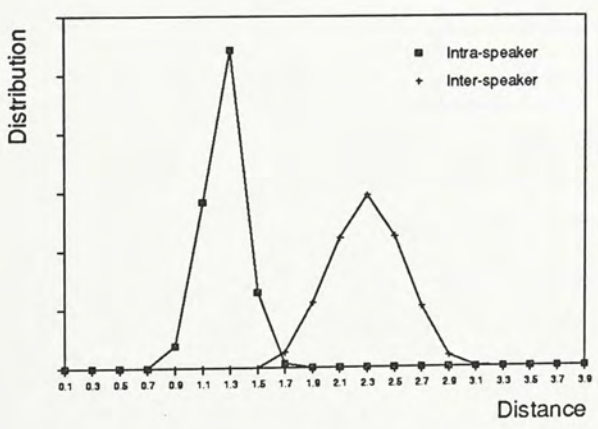
(b) Using 2 digits



(c) Using 3 digits



(d) Using 4 digits



(e) Using 5 digits

Figure 3-7 Intra-speaker and Inter-speaker distance (LTW) distribution employing different number of digits as speaker identity.

Finally, an evaluation for the proposed speaker verification system has been done on the data base described previously in a VAX 8200 computer. Another four different sets of reference utterances were selected and were recorded in Table 3-3. The procedures were identical to those used in statistics calculation. But this time each verification of input sequence of words of a certain speaker to the same speaker was considered as a true claim verification while that to a different speaker was considered as a mimic. The sample size for true claim and false claim verification were therefore equal to that for intra-speaker and inter-speaker respectively in the statistics calculation. Verification results for M=1,2,3,4 and 5 under different distance threshold were tabulated in Table 3-4 and Table 3-5 respectively for DTW and LTW. The results against D_{TH} for different values of M were plotted in Fig.3-8 for DTW and 3-9 for LTW. Finally, the results at two distance threshold, 1.25 and 1.3 for DTW, 1.6 and 1.7 for LTW, against the number of digits used in the input sequence were plotted in Fig.3-10 and 3-11 correspondingly. The observation on the results will be given in the following section.

Trial no.	Reference set utterances
1	U1, U2, U3, U4, U5
2	U8, U9, U10, U11, U12
3	U3, U5, U7, U9, U11
4	U4, U6, U8, U10 ,U12

Table 3-3 Reference sets for verification.

True Reject Error % False Accept Error % Total Error %	Distance Threshold				
No. of digit used (M)	1.20	1.25	1.30	1.35	1.40
1	7.7922 1.1883 8.9805	4.9351 2.1299 7.0650	3.1169 3.5162 6.6331	2.2727 5.1981 7.4708	1.5260 7.5000 9.0260
2	3.7559 0.1340 3.8899	1.8594 0.3470 2.2064	0.8565 0.7930 1.6495	0.3556 1.6349 1.9905	0.1453 2.9792 3.1245
3	1.9386 0.0214 1.9600	0.7178 0.0910 0.8088	0.2290 0.2892 0.5182	0.0685 0.7486 0.8171	0.0187 1.6565 1.6752
4	1.0293 0.0027 1.0320	0.2865 0.0232 0.3097	0.0655 0.1141 0.1796	0.0129 0.3873 0.4002	0.0027 1.0373 1.0400
5	0.6150 0.0003 0.6153	0.1232 0.0050 0.1282	0.0192 0.0431 0.0623	0.0028 0.2060 0.2088	0.0002 0.6841 0.6843

Table 3-4 Verification results under DTW.

True Reject Error % False Accept Error % Total Error %	Distance Threshold				
No. of digit used (M)	1.50	1.60	1.70	1.80	1.90
1	14.7078 1.3506 16.0584	8.6688 3.4091 12.0779	4.9026 6.7532 11.6558	2.3701 11.5519 13.9220	1.2013 17.8474 19.0487
2	9.0622 0.2013 9.2675	3.5457 0.8624 4.4081	1.2441 2.6378 3.8819	0.4710 6.0809 6.5519	0.2144 11.5866 11.8010
3	5.9026 0.0399 5.9425	1.7226 0.3041 2.0267	0.5042 1.3945 1.8987	0.1607 4.1530 4.3137	0.0480 9.9067 9.1447
4	4.0402 0.0079 4.0481	0.9564 0.1164 1.0728	0.2313 0.8251 1.0564	0.0509 3.1684 3.2193	0.0098 7.8397 7.8495
5	2.9505 0.0012 2.9517	0.5720 0.0418 0.6138	0.1004 0.4962 0.5966	0.0164 2.5430 2.5594	0.0037 7.0613 7.0650

Table 3-5 Verification results under LTW.

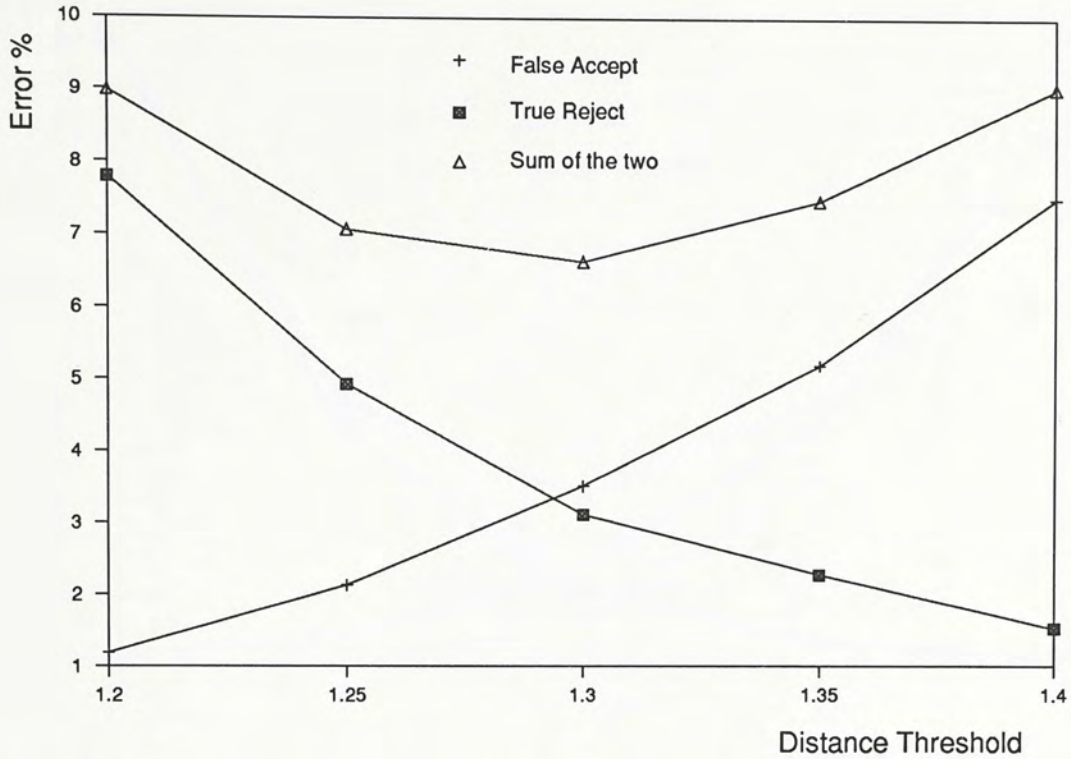


Figure 3-8(a) Verification error vs distance threshold, using 1 digit with DTW.

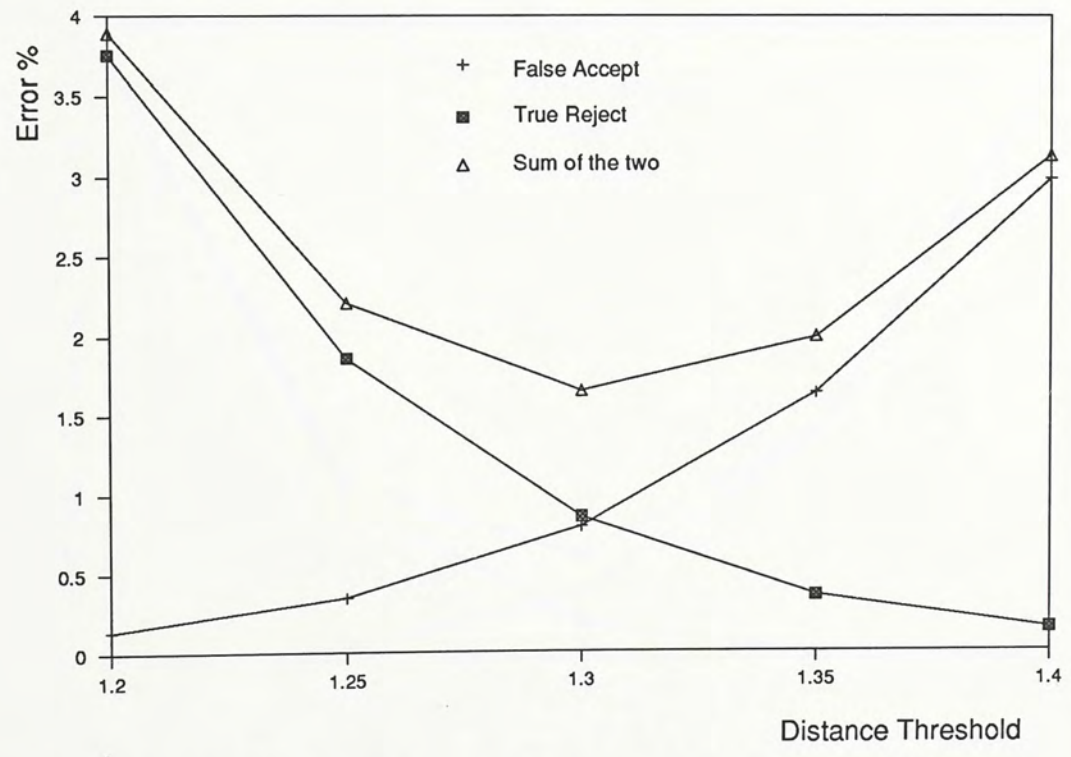


Figure 3-8(b) Verification error vs distance threshold, using 2 digits with DTW.

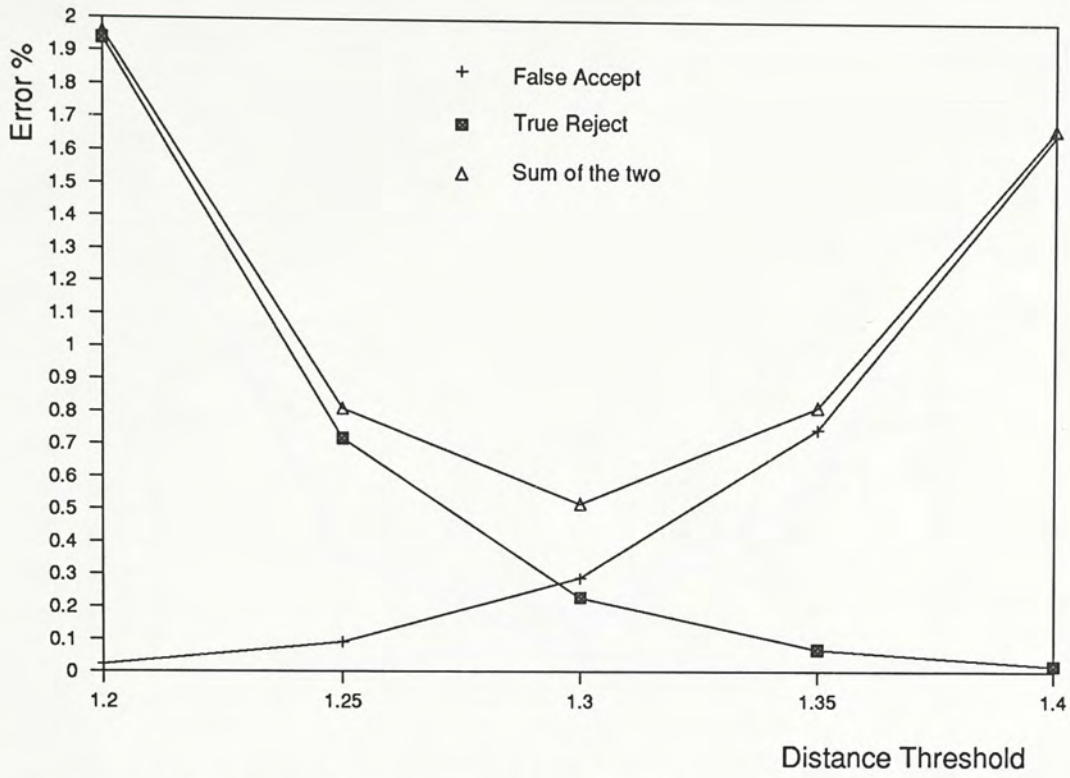


Figure 3-8(c) Verification error vs distance threshold, using 3 digits with DTW.

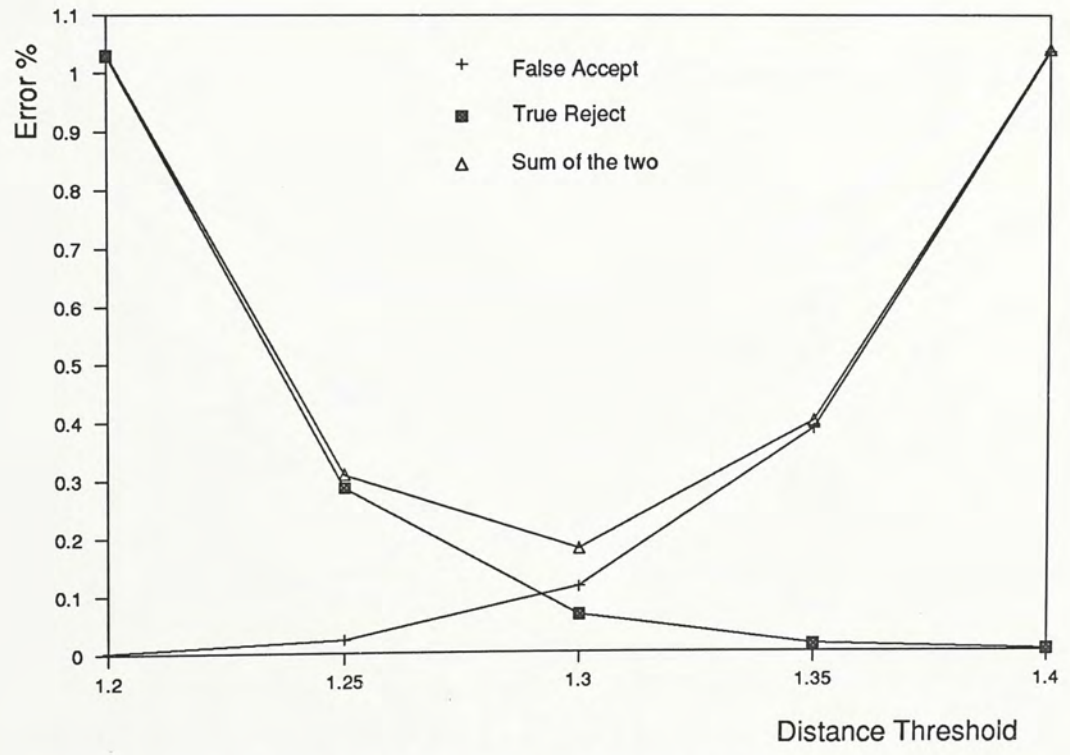


Figure 3-8(d) Verification error vs distance threshold, using 4 digits with DTW.

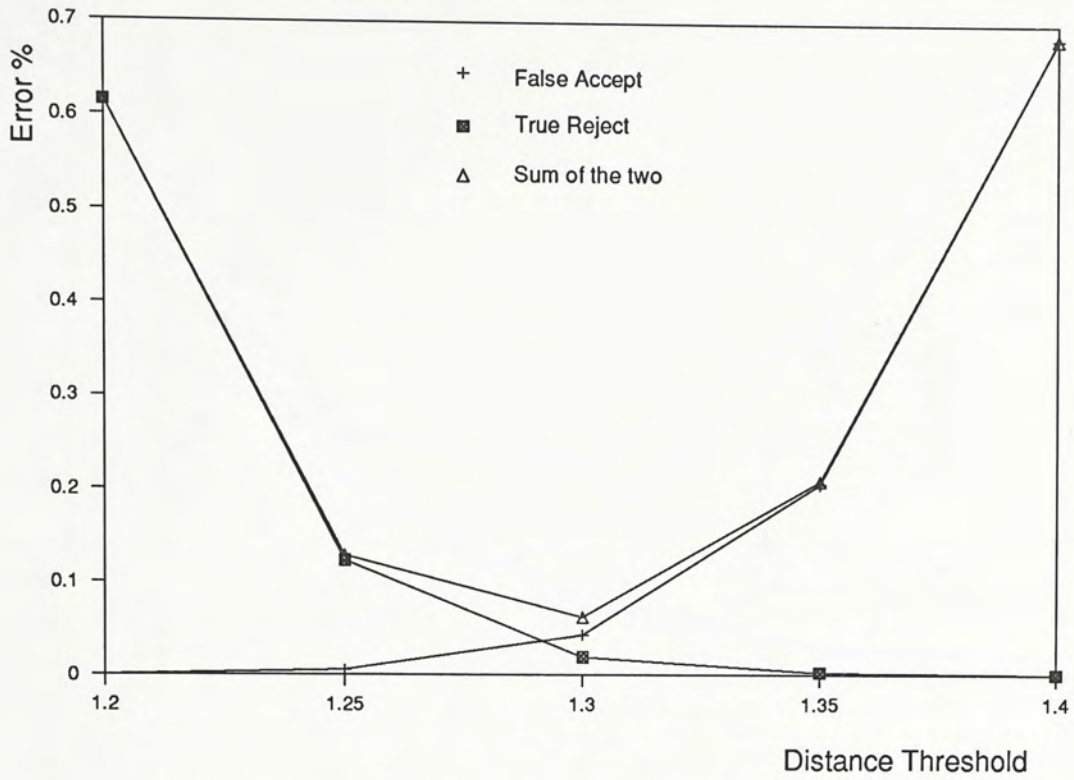


Figure 3-8(e) Verification error vs distance threshold, using 5 digits with DTW.

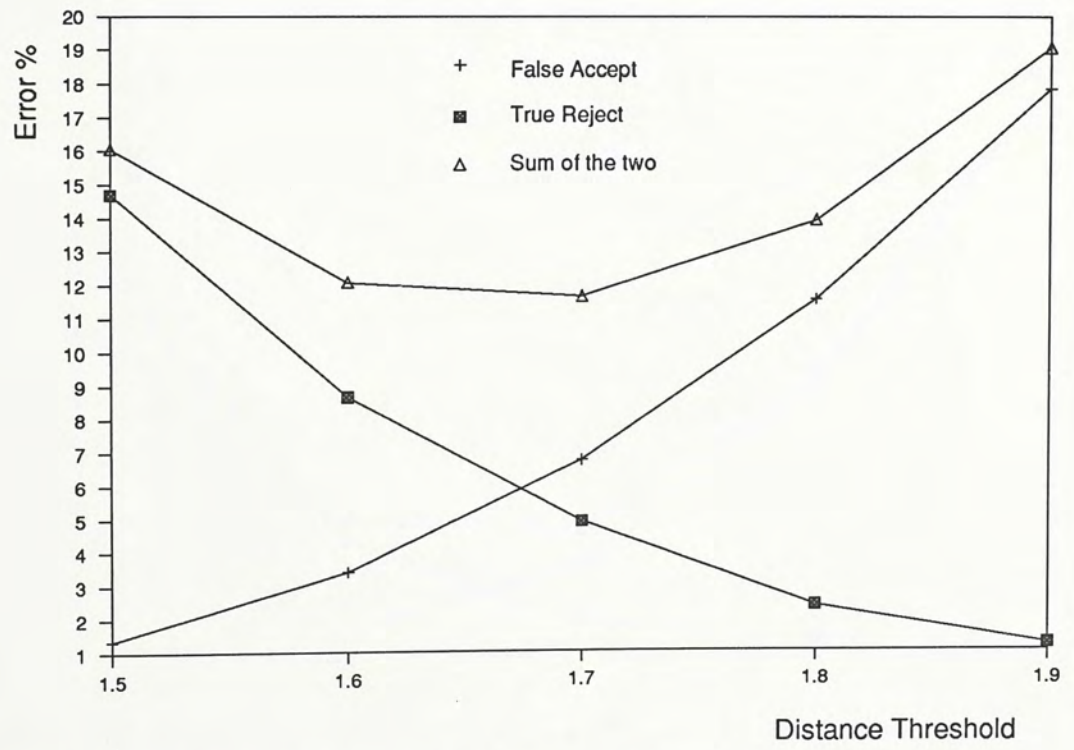


Figure 3-9(a) Verification error vs distance threshold, using 1 digit with LTW.

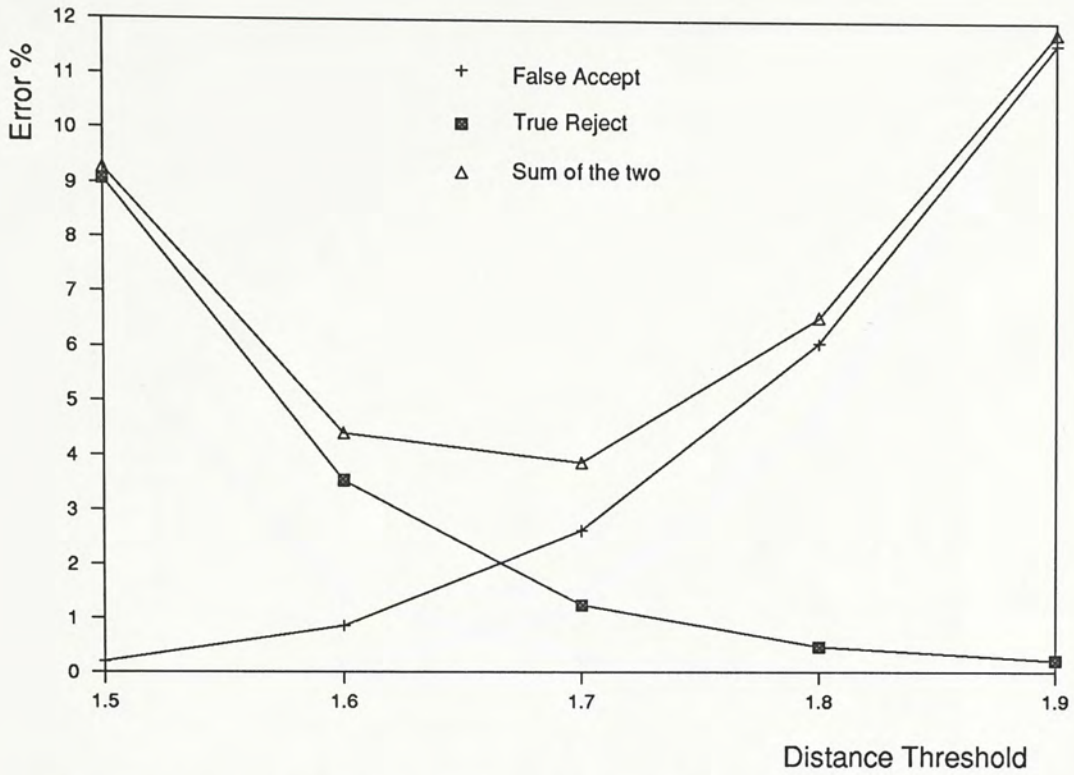


Figure 3-9(b) Verification error vs distance threshold, using 2 digits with LTW.

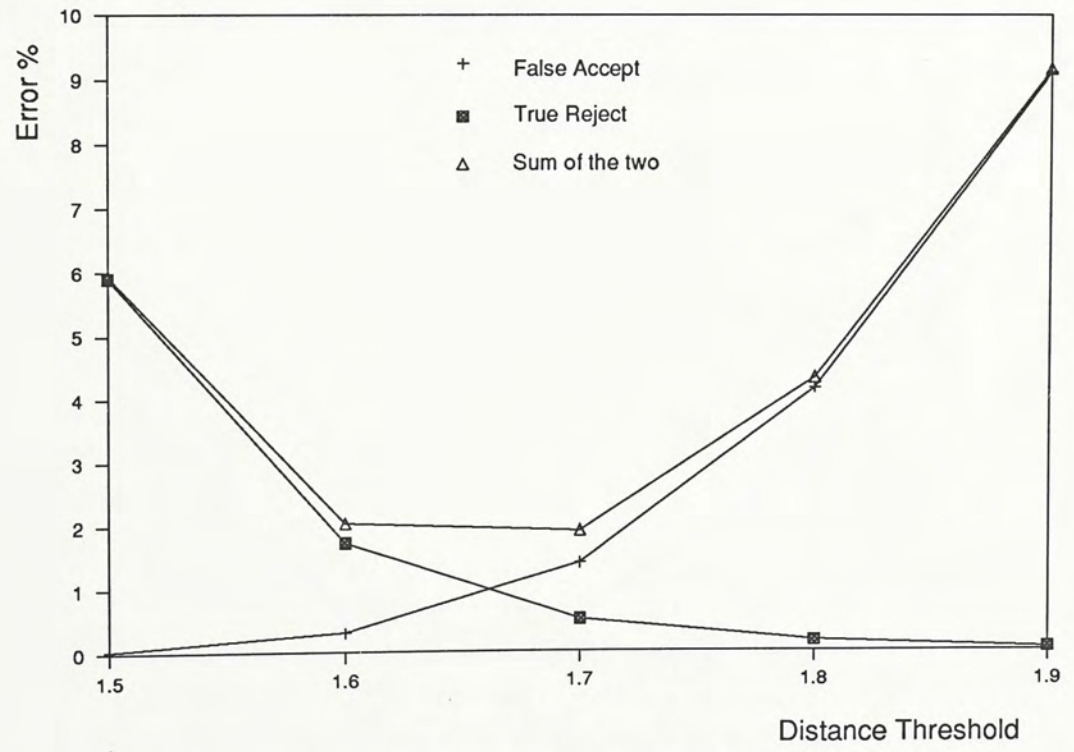


Figure 3-9(c) Verification error vs distance threshold, using 3 digits with LTW.

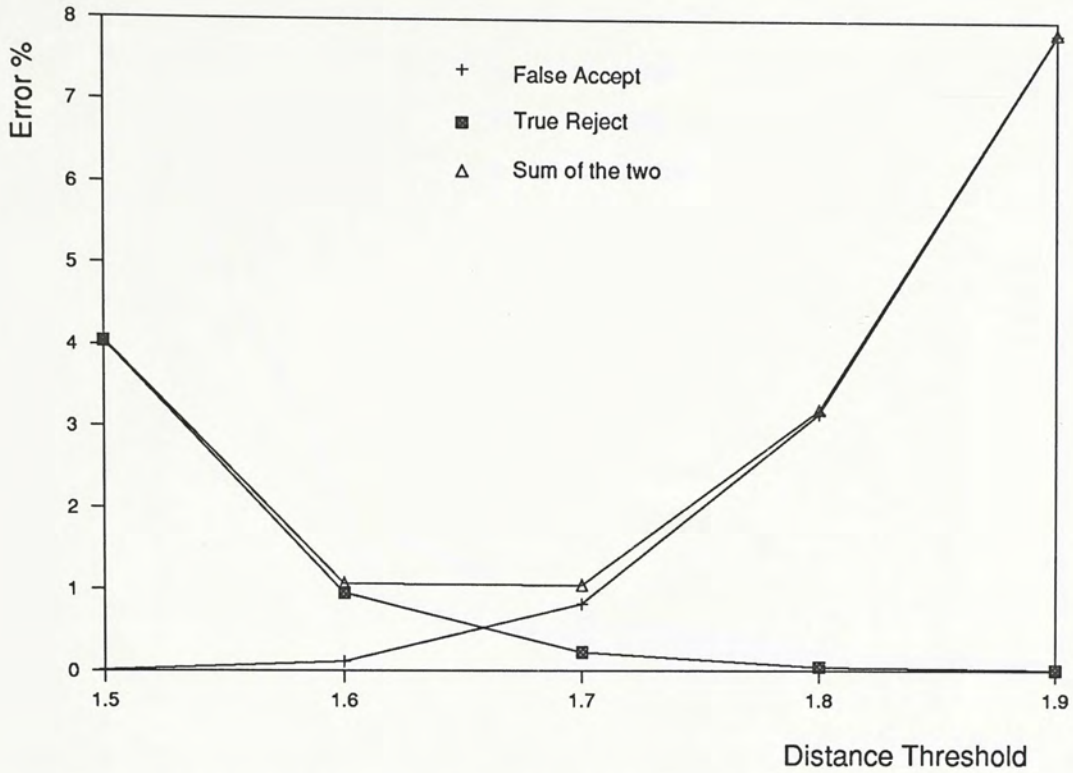


Figure 3-9(d) Verification error vs distance threshold, using 4 digits with LTW.

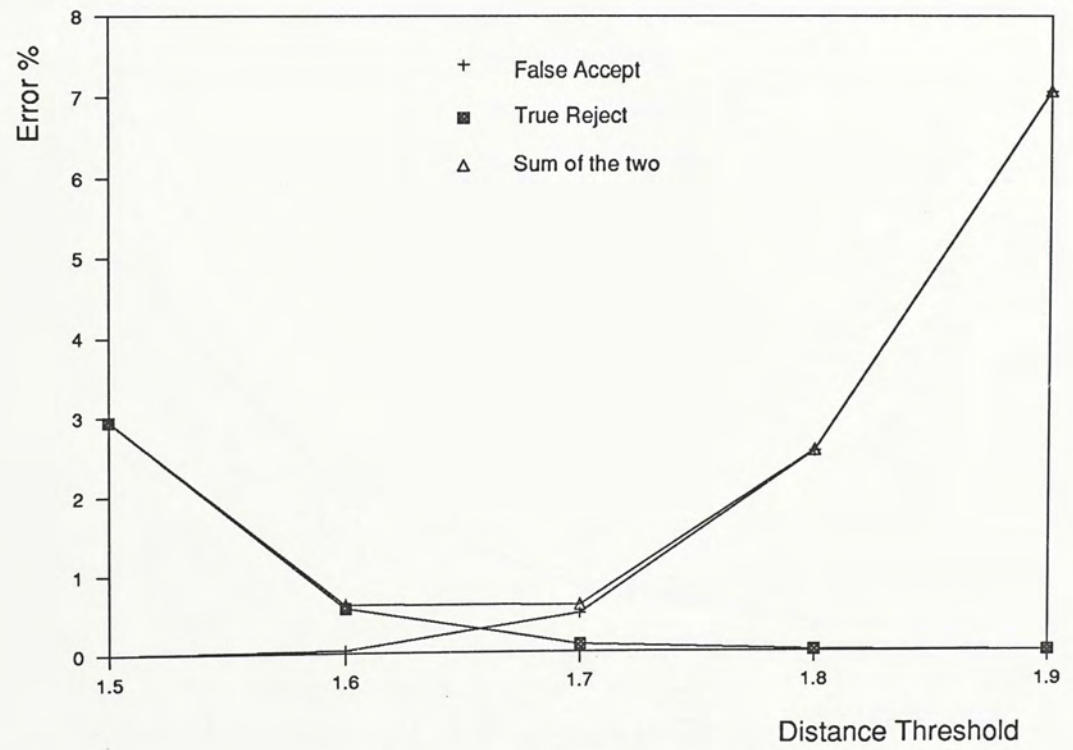


Figure 3-9(e) Verification error vs distance threshold, using 5 digits with LTW.

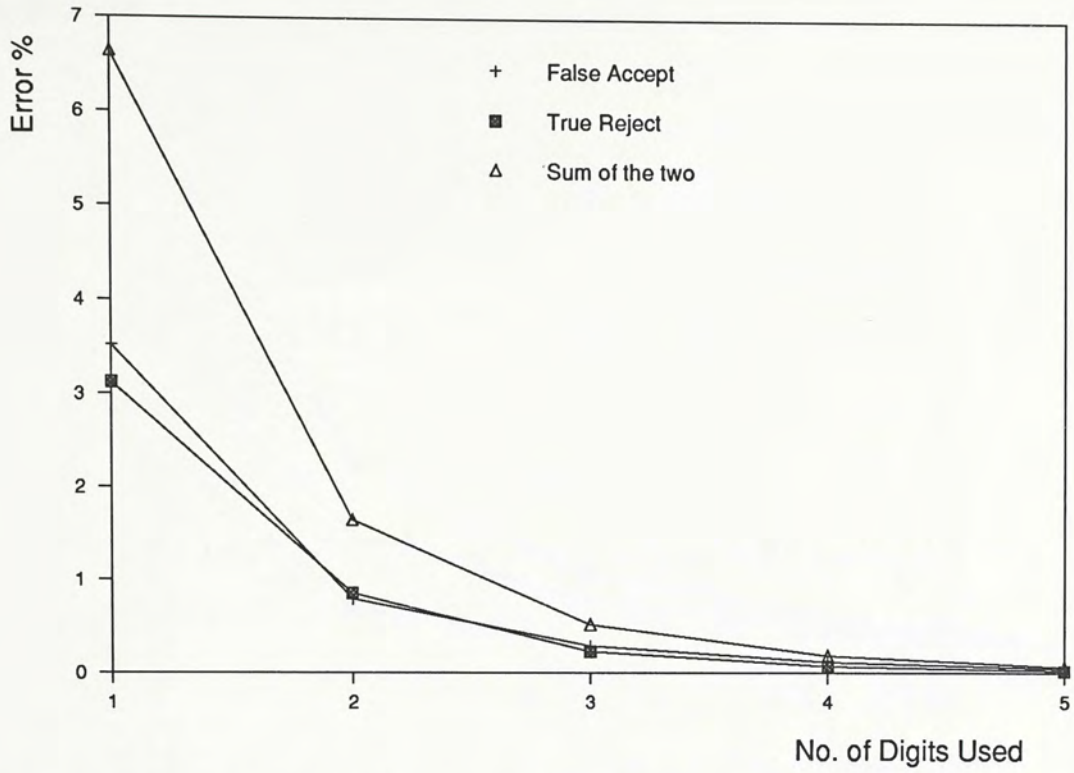


Figure 3-10(a) Verification error vs No. of digits used at distance threshold=1.3 (minimum total error point), under DTW.

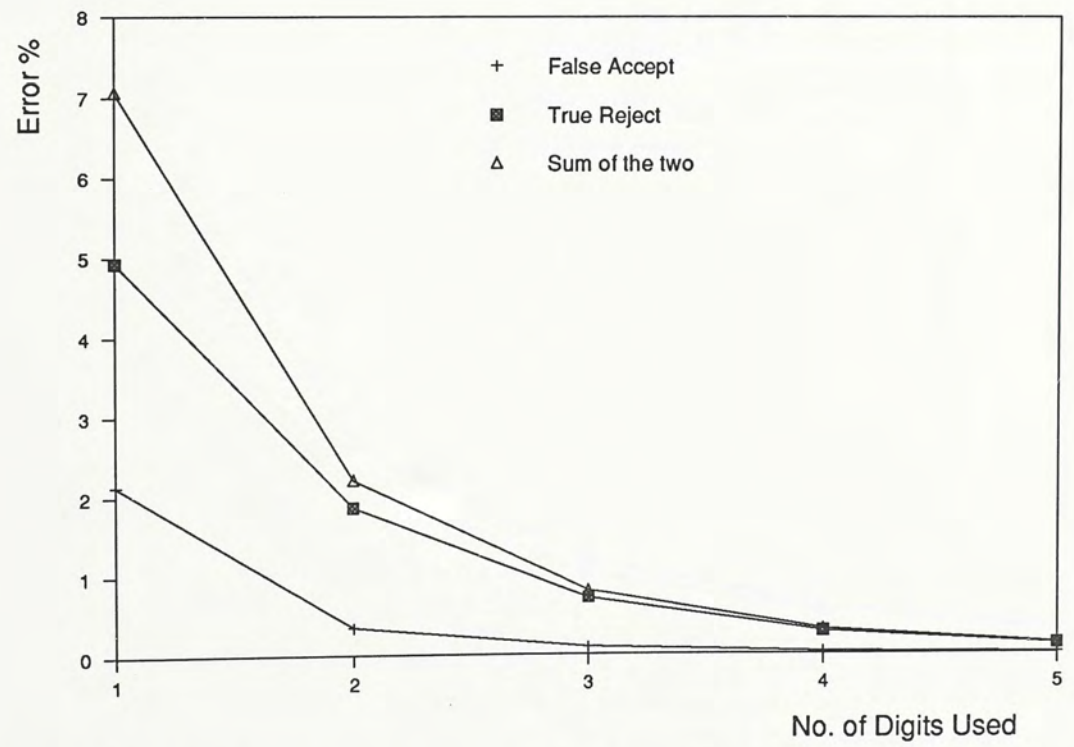


Figure 3-10(b) Verification error vs No. of digits used at distance threshold=1.25 (practical choice), under DTW.

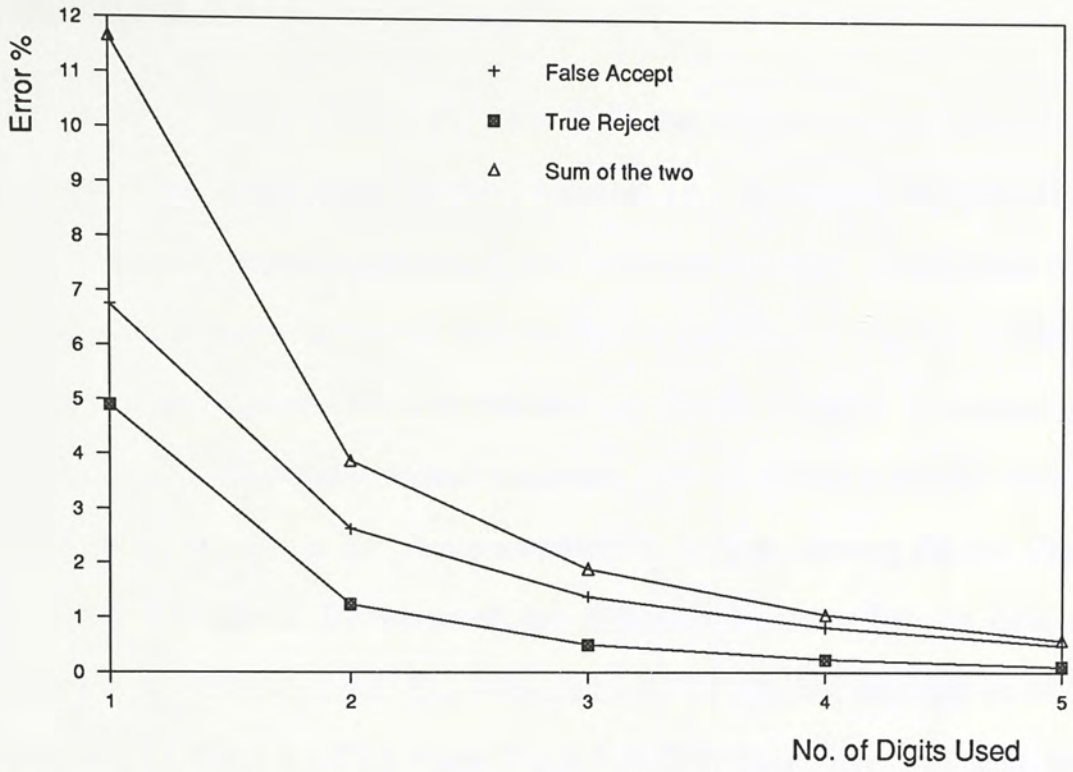


Figure 3-11(a) Verification error vs No. of digits used at distance threshold=1.7 (minimum total error point), under LTW.

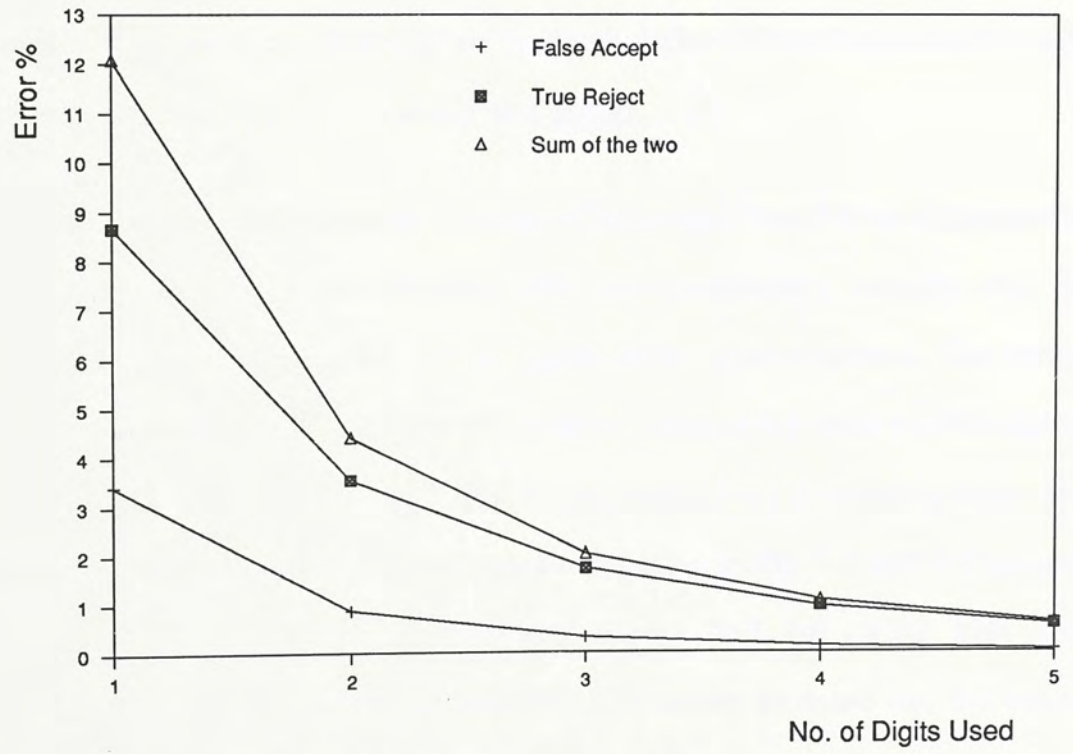


Figure 3-11(b) Verification error vs No. of digits used at distance threshold=1.60 (practical choice), under LTW.

3.4 Observation

In the derivation of equations (3.3) to (3.6), the assumption that the distributions for the intra-speaker and inter-speaker distance, i.e., the closest distance D(m), on each of the single words are identical and follow a Gaussian distribution has been made. From these varies curves shown in Fig.3-4 (DTW) and Fig.3-5 (LTW), it can been seen that although the distributions are not exactly identical between digits due to the small size of the experimental database, they all follow a similar trend which provide a solid ground for the above assumption, at least, among the ten Cantonese digits in the experiment. Furthermore, the distribution curves for the intra-speaker and inter-speaker distances, i.e., D_{FA} averaged over M smallest distance on M distinct digits shown in Fig.3-6 (DTW) and Fig.3-7 (LTW) also have a similar shape as Fig.3-3 which give us a definite means to determine D_{TH} for the purpose of speaker verification. The derivation, though not rigorous, points out that more information can be extracted from a sequence of discrete words so as to distinguish a talker from his voice. The system can therefore be operated under different values of M according to the required verification accuracy and speed.

From the verification results, despite of the two different time alignment method used, the verification errors change with two parameters, namely, the distance threshold, D_{TH}, and the number, M, of digits in the input sequence. The verification results at different values of M shown in Table 3-4 and 3-5 were plotted against D_{TH} on Fig.3-8 and 3-9 for DTW and LTW respectively. As a measuring index, the total of the two errors was calculated and plotted together in Fig 3-8 and 3-9 accordingly. From these figures with different values of M, we find that all the false accepting errors increase with D_{TH} while the true rejecting errors decrease and the total errors have a bowl shape with the minimum error point occur at a D_{TH}=1.3 for DTW and 1.7 for LTW. This minimum value of the total error point is close to the intersecting point of the false accept error and true reject error curves, i.e. the equal error point.

Though for a practical SV system, the operating point should not be necessarily set at the minimum total error point but a lower false accept error point (for the cost of rejecting a true claim is much lower than accepting a false claim), the position of the minimum error point locates the nearby region of possible setting of D_{TH}. In my selection, a value 1.25 and 1.6 is chosen for the operating D_{TH} for DTW and LTW respectively in the proposed system. The verification results at the minimum total error point and the operating point are then plotted against the number M of digits used in the input sequence of test utterance on Fig.3-10 and 3-11.

We can see from Fig.3-10 and 3-11 that all the three errors decrease with the number, M, of digits in the input sequence. This phenomenon agrees with the hypothesis described previously. One important point is that even though the results for DTW are better than those for LTW under the same value of M, however, the performance for LTW is able to catch up with that for DTW if a greater value of M, say 2 in my case, is used. Selecting M to be 5 in the operating D_{TH}, i.e. 1.6, an verification accuracy of 99.3862% is obtained in using LTW. This result is comparable to, and even better than, the verification accuracy 99.1912% obtained in using DTW with M=3 at the operating point. However, by the records obtained during the verification experiment, the ratio of the average computational time involved by using LTW to that using DTW for one verification test with M=1 is only 1:40 approximately. This indicates clearly that employing LTW in speaker verification on a word-by-word basis not only can allow a higher verification score by increasing the number of words used in the input test utterance, the time for verification can, at the same time, be reduced greatly so that real time application will be practically possible.

Chapter 4

Speaker Identification System

In the speaker verification (SV) system described in chapter 3, the speakers' characteristics are carried intrinsically by the acoustic events of their utterances while the identity claim are extracted from the contents of the sentence and have to be remembered by the users. However, if the identity code were forgotten, verification becomes impossible. To enhance system applicability and full speech automation, a speaker identification (SI) process has been devised and implemented which can be called upon in case the system user cannot remember his identity or personal code. In fact, a SI system can be made stand-alone for any specific application, in particular for the purpose of reconnaisance. The suggested SI system in this project is mainly used to identify the user as one of the eligible candidates so that personal code could be recalled, and hence it might be considered as an extension of the previous SV system.

The energy-time profile (ETP) is used again as speech parameter in the SI system to represent phonetic characteristics of individual user and therefore all the preprocessing and parameter extraction procedures are identical to those in the SV system. The block diagram of the SI system is shown in Fig.4-1. The system is supposed to be operating in an "open set" environment, i.e. anyone, even not one of the candidates (those called the registered users and are known to the system) can approach and request the system for an identification. For a system with S candidates or registered users, including the decision that the user is not one of the S members, there are totally S+1 possible decision outcomes. Each user approaching the system is requested to utter a sequence of words by a monitoring unit which informs the system about this sequence simultaneously so that a proper matching of the input

words to the references can be achieved. This requested sequence is designed to be composed of 5 distinct Cantonese words which are randomly chosen and ordered from the system word library, i.e., the 10 Cantonese digits in this case.

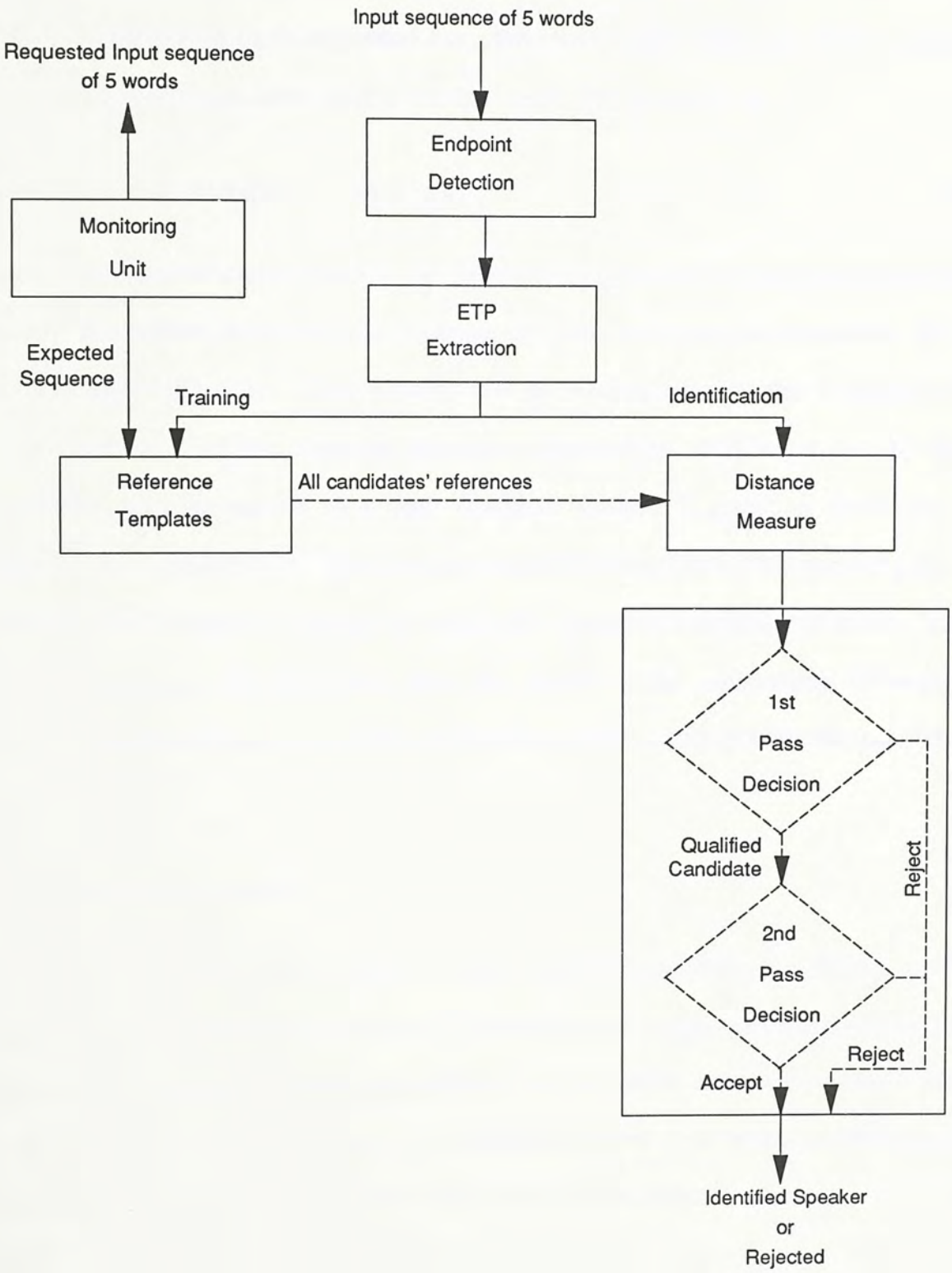


Figure 4-1 A speaker identification system using ETP

4.1 Decision Making

Similar to the SV system, each of the discrete words will be subjected to a distance measure with the reference templates and the smallest distances will be computed as shown in Fig.3-2 in page 33. Now, if there are again N references for each word uttered by each registered user, then the smallest distance, $D^q(m)$, obtained when being compared with user q for the word "m" is given by

$$D^q(m) = \min[d_i^q(m), \quad i=1,2,...,N] \tag{4.1}$$

Since in the identification process, no identity is known beforehand and therefore a total of S smallest distances are obtained, one for each possible candidate, for the word designated by "m". This process will be repeated for all the 5 words in the input utterance and the distance measures obtained in each word for all the S candidates will be passed to a first decision process in order to determine the existence of a qualified candidate under a set of criteria. During the second pass, the conclusion of whether to accept or reject this qualified candidate, if exists, as the identified speaker will be drawn from the results of the comparison between the obtained smallest distances for this qualified candidate and a predetermined distance threshold D_{TH}.

4.1.1 First Pass Decision

In this decision stage, either a unique qualified candidate who fulfills a set of decision criteria is selected among the S candidates, or a reject decision will be made which indicates an insufficient resemblance of the user's utterance to any of those from the S candidates. If none of the registered users is selected as the qualified candidate, the identification process will cease with a reject outcome.

For each of the 5 words in the input utterance, S smallest distances are obtained for the S candidates. These S smallest distances are then arranged in ascending order and the candidates in the position with 1st, 2nd and 3rd minimum of the smallest distance are given the credit A, B and C respectively. This procedure is repeated in all of the 5 words and so 5 A's, 5 B's and 5 C's are given to the corresponding candidates. A table of credit, with entries of the code of those candidates who have received credit(s) during the process, is then formed. Some typical records are shown in Table 4-1. The symbol 'Spk "s"' in a certain entry indicates the speaker designated by "s" has obtained the corresponding credit for the respective word. The qualified candidate is then selected from one of the credited candidates in the table if "he" satisfies any one of the following 3 conditions:

(1) possessing altogether 5 credit A's
(2) possessing altogether 4 credit A's and 1 credit B or credit C
(3) possessing altogether 3 credit A's and 2 credit B's

Obviously, the above empirical criteria of selection are based on a majority rule so that the outcome is unique. In the 4 cases shown in Table 4-1(a) to 4-1(d), speaker "3" satisfies the conditions and is therefore selected as the qualified candidate and a second test will be therefore adminstered to "him" for the acceptance of "he" to be the identified speaker. However, in cases shown in Table 4-1(e) and 4-1(f), none of the candidates has obtained sufficient credits to be the qualified candidates and a reject decision will finally be made.

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "3"	Spk "3"	Spk "3"	Spk "3"	Spk "3"
Credit B	Spk "2"	Spk "1"	Spk "2"	Spk "5"	Spk "2"
Credit C	Spk "1"	Spk "4"	Spk "1"	Spk "1"	Spk "1"

Table 4-1(a) Credit table, case 1

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "3"	Spk "3"	Spk "3"	Spk "2"	Spk "3"
Credit B	Spk "2"	Spk "1"	Spk "2"	Spk "3"	Spk "2"
Credit C	Spk "4"	Spk "2"	Spk "5"	Spk "4"	Spk "1"

Table 4-1(b) Credit table, case 2

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "3"	Spk "4"	Spk "3"	Spk "3"	Spk "3"
Credit B	Spk "4"	Spk "2"	Spk "1"	Spk "4"	Spk "4"
Credit C	Spk "2"	Spk "3"	Spk "4"	Spk "1"	Spk "5"

Table 4-1(c) Credit table, case 3

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "2"	Spk "3"	Spk "3"	Spk "3"	Spk "2"
Credit B	Spk "3"	Spk "2"	Spk "2"	Spk "2"	Spk "3"
Credit C	Spk "4"	Spk "1"	Spk "4"	Spk "1"	Spk "1"

Table 4-1(d) Credit table, case 4

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "1"	Spk "3"	Spk "3"	Spk "3"	Spk "3"
Credit B	Spk "2"	Spk "5"	Spk "4"	Spk "1"	Spk "1"
Credit C	Spk "4"	Spk "2"	Spk "1"	Spk "2"	Spk "5"

Table 4-1(e) Credit table, case 5

	1st word	2nd word	3rd word	4th word	5th word
Credit A	Spk "3"	Spk "5"	Spk "3"	Spk "2"	Spk "3"
Credit B	Spk "5"	Spk "3"	Spk "2"	Spk "5"	Spk "2"
Credit C	Spk "2"	Spk "2"	Spk "5"	Spk "3"	Spk "5"

Table 4-1(f) Credit table, case 6

4.1.2 Second Pass Decision

After selecting the qualified candidate in the previous stage, decision is to be made in this stage of whether to accept or reject this most probable candidate as the identified speaker. With the observation in the intra-speaker and inter-speaker distance distributions in the previous chapter, the 5 smallest distances, obtained for the qualified speaker (the speaker "3" in the example shown) are compared to a preset distance threshold D_{TH}. To accept the qualified candidate as the final identified speaker, the average of the 5 smallest distances must be smaller than D_{TH}. In addition, at least 3 or more of the 5 smallest distances for the qualified candidate must also be smaller than D_{TH}. The qualified candidate will not be accepted as the identified speaker should one of these two criteria not being satisfied. This will obviously eliminate the selection of a non-registered user even though his acoustic features resemble that of a specific speaker but not exact enough.

4.2 System Evaluation and Results

The same database, which had been used for evaluation of the SV system, was used again to study the performance of the above identificaition system. There are three types of error that might arise in a SI system, namely, the legal user rejection error, the illegal accepting error and the incorrect matching error. The first two are mistakes made by either rejecting a legal user or accepting an illegal user whilst the third is due to misidentifying a registered speaker to a wrong one. The evaluation was performed over the 4 training and testing utterance sets (recorded in Table 3-3 in page 46) as in the case for the SV system. However, throughout the tests, one out of the eleven speakers was considered as an illegal user, or "outsider", while the remaining speakers acted as candidates who had registered to use the system and their utterance had been stored as reference templates. The purpose for such arrangement was to investigate the performance of the system under the request of an illegal user for an identification. This situation, being essential and inevitable, exists in every practical speaker identification system and must be considered. Each of the eleven speakers took turns to be the "outsider" while the other ten were treated as a system qualifier. For each outsider, 50 random combinations, chosen from the 12 utterances for each word, were tested for each of the 252 different patterns (${}_{10}C_5$) consisting 5 distinct Cantonese digits. On the other hand, for each system user, 50 other random combinations, chosen from the 7 utterances for each word, were again tested for each of the 5-digit patterns. Consequently, the total sample space for the identification of sytem users and illegal users were 4x252x50x11x10=5544000 and 4x252x50x11=554400 respectively. These sample sizes, for a 10 user system with one outsider, are large enough to give us unbiased testing results. The results were tabulated in Table 4-2. The sytem user rejection error, the illegal user accepting error, together with the total of these two errors were plotted against D_{TH} and were shown in Fig.4-2 and 4-3 for DTW and LTW respectively. The incorrect matching error,

due to its comparatively small values, were not plotted together. Finally, tables showing to whom the outsider was incorrectly identified at a distance threshold=1.3 for DTW and 1.6 for LTW were shown in Table 4-3.

	Distance Threshold				
Error % of	1.20	1.25	1.30	1.35	1.40
Wrong match of system users	0.0000	0.0000	0.0001	0.0001	0.0001
Reject of system users	1.9240	1.1832	1.0450	1.0341	1.0216
Accept of illegal users	0.0014	0.0186	0.1448	0.5372	1.5438

Table 4-2(a) Identification results using DTW

	Distance Threshold				
Error % of	1.5	1.6	1.7	1.8	1.9
Wrong match of system users	0.0000	0.0001	0.0003	0.0004	0.0005
Reject of system users	6.7080	3.6526	2.9677	2.8324	2.8124
Accept of illegal users	0.0049	0.1378	1.2314	4.4547	7.6414

Table 4-2(b) Identification results using LTW

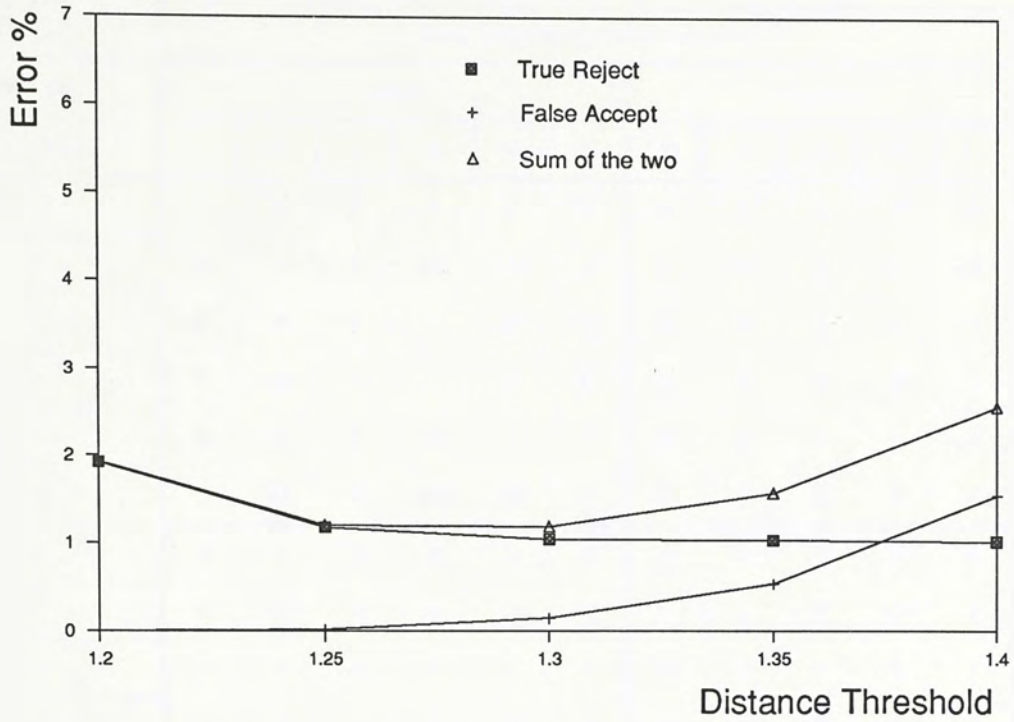


Figure 4-2 Identification error against distance threshold (DTW)

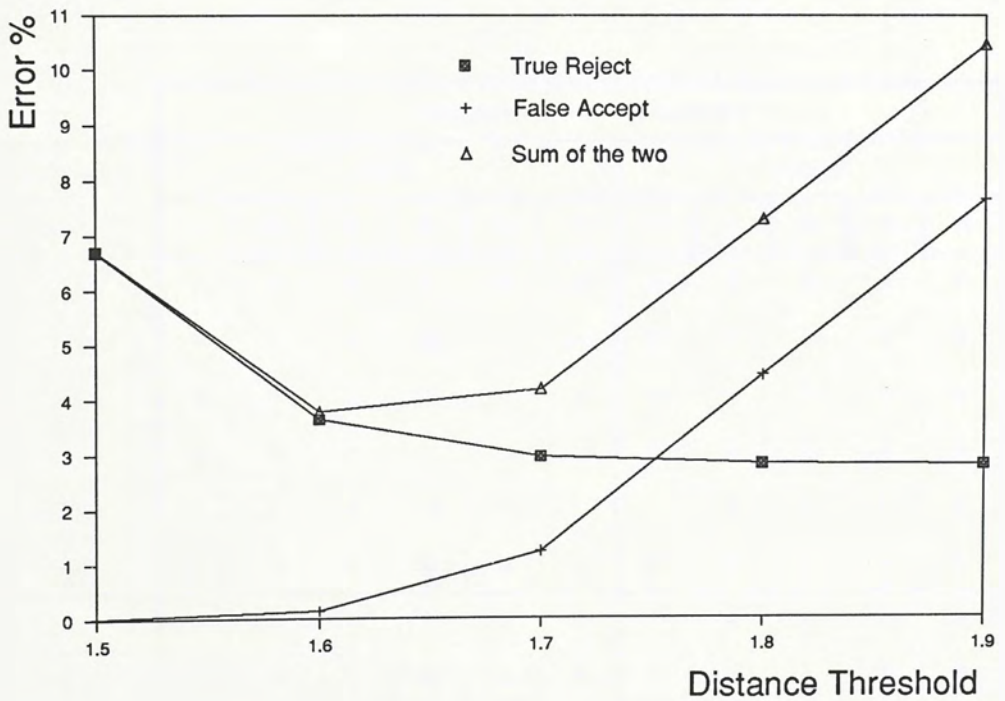


Figure 4-3 Identification error against distance threshold (LTW)

"Outsider" being speaker	Being incorrectly identified as speaker										
	(Male)						(Female)				
	1	2	3	4	5	6	7	8	9	10	11
1	/	0	0	0	0	0	0	0	0	0	0
2	0	/	0	93	0	8	0	0	0	0	0
3	0	0	/	0	0	0	0	0	0	0	0
4	0	12	0	/	10	81	0	0	0	0	0
5	0	0	0	40	/	0	0	0	0	0	0
6	2	84	0	222	0	/	0	0	0	0	0
7	0	0	0	0	0	0	/	0	0	0	0
8	0	0	0	0	0	0	0	/	0	117	0
9	0	0	0	0	0	0	0	0	/	0	5
10	0	0	0	0	0	0	0	129	0	/	0
11	0	0	0	0	0	0	0	0	0	0	/

Table 4-3(a) Incorrect identity matching of outsiders (DTW)

"Outsider" being speaker	Being incorrectly identified as speaker										
	(Male)						(Female)				
	1	2	3	4	5	6	7	8	9	10	11
1	/	0	0	0	0	11	0	0	0	0	0
2	0	/	0	16	0	142	0	0	0	0	0
3	0	0	/	0	0	0	0	0	0	0	0
4	0	3	0	/	15	112	0	0	0	0	0
5	0	0	0	120	/	1	0	0	0	0	0
6	11	62	0	35	0	/	0	0	0	0	0
7	0	0	0	0	0	0	/	0	0	0	0
8	0	0	0	0	0	0	0	/	0	168	0
9	0	0	0	0	0	0	0	0	/	2	6
10	0	0	0	0	0	0	0	50	2	/	0
11	0	0	0	0	0	0	0	0	8	0	/

Table 4-3(b) Incorrect identity matching of outsiders (LTW)

4.3 Observation

From the results shown in Table 4-2, it can be seen that the wrong matching error of system users is relatively low. Actually, this phenomenon should not be surprising as all of us speaks somewhat differently. In fact, it acts as an indication on the effectiveness of the ETP in distinguishing speakers from their voice. The errors in rejecting system users and accepting illegal users on the other hand, is much much higher than this mis-matching error in all the tests. These two significant errors, together with their sum, will therefore become an objective measure of system performance. Their corresponding values against D_{TH} were shown in Fig.4-2 and 4-3.

It can be seen from Fig.4-2 and 4-3 that the rejecting error of system users tends to decrease with D_{TH} initially but converge approximately to a constant at a greater value of D_{TH}. Since the change of D_{TH} only affects the decision in the 2nd pass, the constant error in this steady state region would thus be the total reject percentage of system user during the 1st pass decision. On the other hand, the accepting error of illegal user inceases with D_{TH} and consequently, the total of the two errors comes to a minimum at a certain value of D_{TH}. For a practical SI system, the accepting error should be kept much lower than the rejecting error as the cost of accepting illegal user is much higher. Therefore, the minimum total error point, at which reasonable values for the accepting and rejecting error(for DTW, 0.1448 and 1.0450 respectively; for LTW, 0.1378 and 3.6526 respectively) were found, becomes a very good guide for the determination of the system distance threshold. In this experiment, a distance threshold of 1.3 for DTW and 1.6 for LTW were chosen. At these operating points, the obtained identification accuracies are 98.8102% and 96.2096% by using DTW and LTW. Obviously, the identification error using DTW is lower than that using LTW, but will be entirely contributed by the incorrect rejection of legal users which might be bearable in most cases.

From Table 4-3, one can notice that identification errors occur entirely among speakers of the same sex. This is in fact a very common phenomenon in most of the identification systems even using different kinds of parameters to represent speakers' phonetic characteristics. Moreover, identification error seems to be happened in a fairly symmetric way, i.e., if a certain speaker is identified to be another specific speaker then the reverse is usually true as well.

Chapter 5

Speech Recognition of Discrete Cantonese Words on a Probabilistic Criterion

In order to extract the identity code from the user's input utterance, a speech recognition algorithm of Cantonese characters become an essential part in the SV system described in Chapter 3. Moreover, the speech recogniser, if installed, can also be applied for the translation of users' requests in the form of system commands to facilitate the automatic man-machine communication by voice.

Although the goal of machine recognition of continuous speech remains elusive, a greater degree of success has been achieved in recognition of discrete word from a fixed vocabulary. Indeed, many isolated word recognition (IWR) system have been built and used in a wide variety of applications. However, a large amout of computation is needed in most template-based recognizers to achieve time alignment using DTW [29]. Speech recogniser based on hidden Markov models, on the other hand, have less computation but more complicated parameter estimation procedures for model generation [30]. Large storage, intensive computation, together with complex system configuration have made hardware implementation of a high speed, low cost speech recognition system very difficult, if not impossible.

In this project, besides the SV and SI systems described previously, an efficient talker independent IWR algorithm is studied for mono-syllabic languages, specially for Cantonese. ETPs are used once more to carry speech features for the recognition of discrete word. Again, instead of using DTW, another form of LTW is employed which is simpler and comparable to DTW for small variation of speech periods [19]. A probabilistic approach is applied to measure the degree of similarity between an input utterance and the reference templates [28]. On matching an unknown token with the references, the one with largest probability of resemblance was taken as the

recognised word. The block diagram of the IWR system is shown in Fig.5-1.

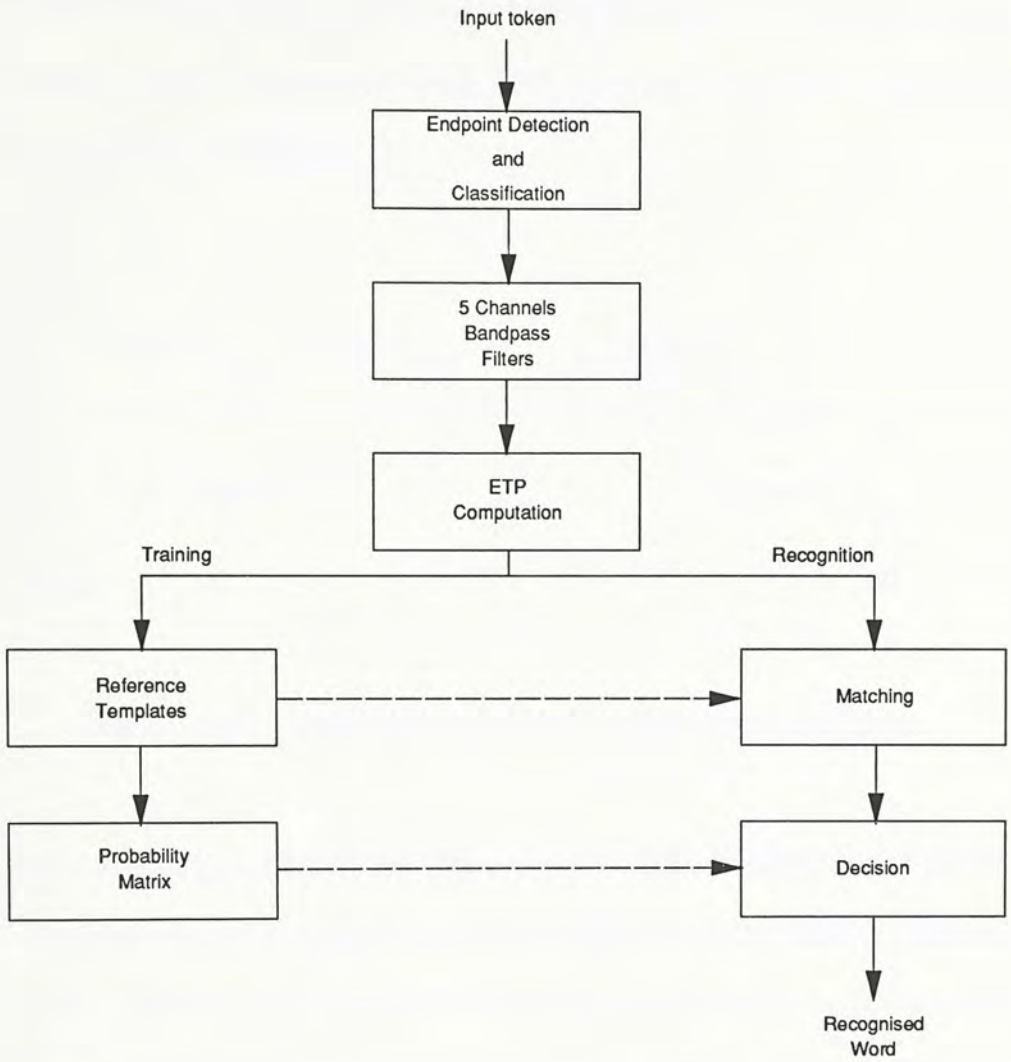


Figure 5-1 Speech Recognition system using ETP on a Probabilistic Criterion

5.1 Feature Extraction

In accompany to the 5 filter band signals from which the ETPs are computed, the segmental energies of the wide-band signal is added to the ETP matrix in order to represent the speech contents in a better way. Once again LTW technique is employed to achieve the goal for fast recognition using economic hardware. In the previous chapters, LTW is simply implemented by linear interpolating the feature parameters (i.e. the ETPs, extracted from a time segment of constant length) to a

fixed number, say 16 in the previous case, as described in equations (2.19)-(2.25). An alternative way for the implementation of LTW is by dividing each utterance, which contains completely a discrete word after endpoint detection and preprocessing, into 16 equal duration segments with 50% overlapping. The way of locating the segments is shown in Fig.5-2.

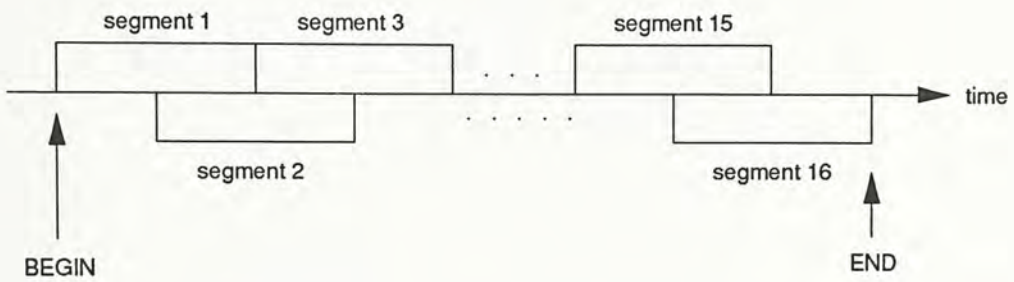


Figure 5-2 Utterance segmentation with equal durations.

The segmental energies, calculated as the sum of the squared values of all the sampled data within each segment, are then evaluated. However, for different utterances, their length will be different and hence the durations for each utterance's segment will be unequal and finally the absolute energy so calculated will be no longer useful for comparison. The segmental energies are, therefore, normalized by the maximum segmental energy, E_{0M}, of the wide band signal of the utterance to achieve meaningful comparison between utterances.

As discussed previously that, due to the energy level variation between the high and low energy portion of a discrete word, a distance measure given by equation (2.13) is adopted to alleviate this problem. An alternative solution is by transforming the normalized energies logarithmically and measuring the distance by traditional Eucludean distance formula. This method has been found effective for speech recognition. Consequently, the utterance containing a discrete word will be represented

by an ETP matrix having 16 energy time vectors with contains 6 logarithmic normalized energy elements. An ETP matrix is shown below and its transformed normalized energy elements are given by equation (5.1).

$$\begin{pmatrix} \hat{E}_0(1) & \hat{E}_0(2) & \hat{E}_0(3) & \dots & \hat{E}_0(16) \\ \hat{E}_1(1) & \hat{E}_1(2) & \hat{E}_1(3) & \dots & \hat{E}_1(16) \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \hat{E}_5(1) & \hat{E}_5(2) & \hat{E}_5(3) & \dots & \hat{E}_5(16) \end{pmatrix}$$

where

$$\hat{E}_q(k) = \text{Log}_{10}\left[\frac{E_q(k)}{E_{0M}}\right] \qquad q=0,1,...,5, \quad k=1,2,...,16 \tag{5.1}$$

The ETP matrix so created for the input token can be used to compare with the reference patterns to find the correct word by a minimum distance measure. The distance d(X,Y) between the ETP matrix of an input utterance X to that of the reference utterance Y is then given by

$$d(X,Y) = \sum_{k=1}^{16} \sum_{q=0}^{5} (\hat{E}_q(k)_X - \hat{E}_q(k)_Y)^2 \tag{5.2}$$

Therefore, upon recognition of an input token containing a discrete word, the word whose reference having the minimum distance measured to the input token during template matching will be selected as the recognised word if they resemble close enough.

5.2 System Training

Instead of using template matching based on a minimum distance measure, a probability measure is introduced so that the recognition decision is made upon a

statistical probability criterion. The reference prototypes are generated by clustering the feature parameters extracted from a large training set on a temporal basis. The training set contains utterances uttered by a number of speakers on each word of the vocabulary. Each time frame will then be represented by a number of energy vectors from each of these training utterances. The number of utterances for each word that closely matched to a specific cluster centre is recorded and this is related statistically as the probability of finding the corresponding word in that particular template. On receipt of an input token, template matching is first performed using a minimum distance measure according to equation (5.2) and recognition is then accomplished by identifying the word which has the largest probability of resemblance.

For each ETP matrix, the temporal energy vector will be treated as an input pattern in its entirety and no time warping of frame based features is required. For the ith segment, reference prototypes are created from the corresponding transformed normalized ETP vectors of the training utterance using the modified K-Means (MKM) clustering algorithm [31] and the condition under which clusters are split will depend on the largest intracluster distance. This has the advantage of permitting the isolation of outliers while still maintaining the property that within each cluster the word patterns are highly similar. The flowchart for the MKM clustering algorithm is shown in Fig.5-3.

However, if the vocabulary size is T and if S tokens for each word are used for training, then a distance matrix of dimension (SxT) by (SxT) will be required in the clustering. A substantial amount of computation will then be needed if the order of SxT is large, say, 1000 or above, and this would probably exceed the capability of most microcomputers. Another way of performing clustering for template generation is, therefore, adopted.

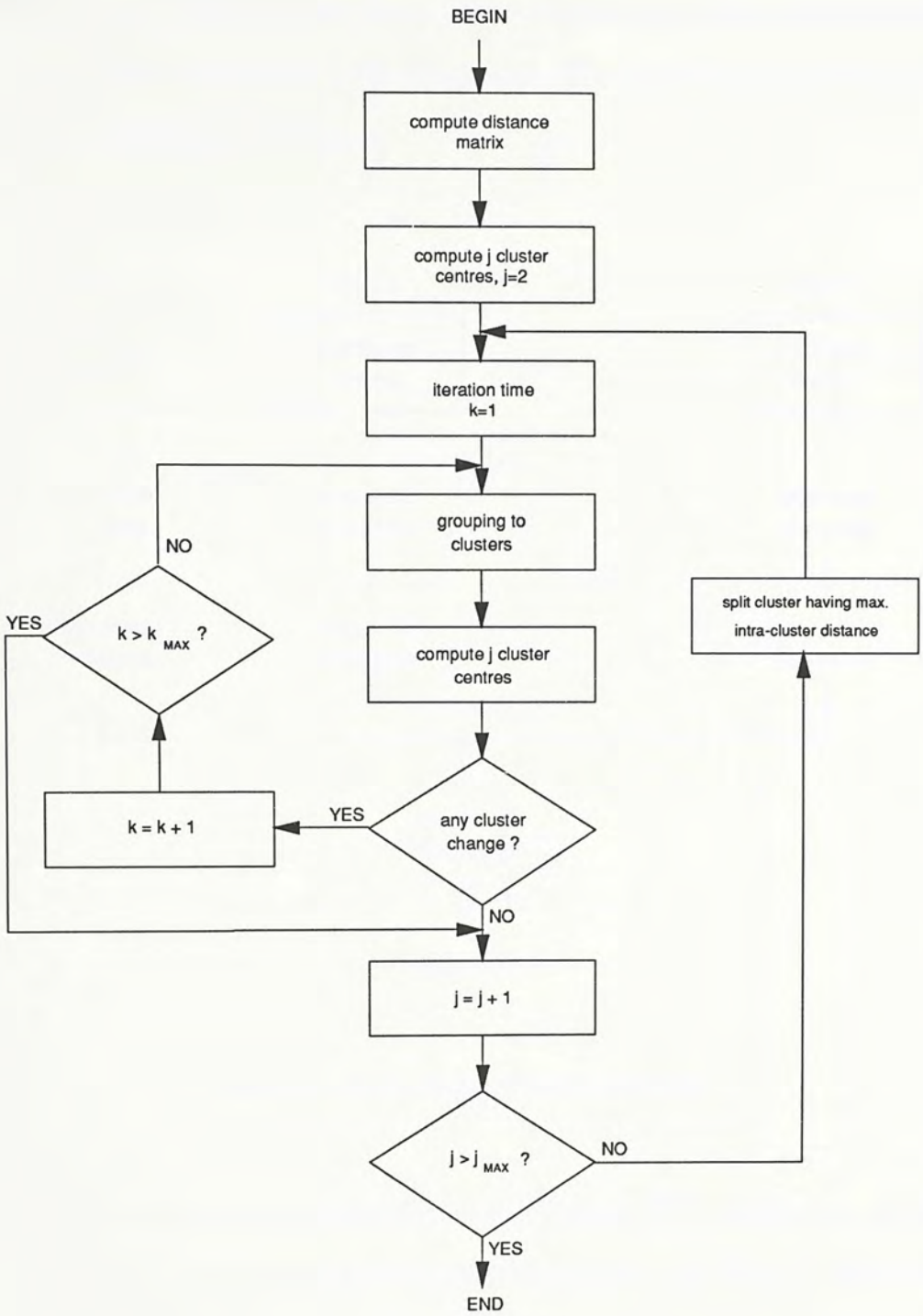


Figure 5-3 Modified K-Mean Clustering Algorithm.

Instead of using altogether S tokens for each word in the clustering process, a group of, say, R temporary prototypes are first produced by the use of the MKM clustering algorithm for each word from the respective S training tokens, where $R \ll S$. That is, a total number of RxT references are now formed for the T words in the

vocabulary. These prototypes are then clustered again in a second stage to form, say, Q final class representatives -- the templates. The modified way of clustering is illustrated schematically in Fig.5-4.

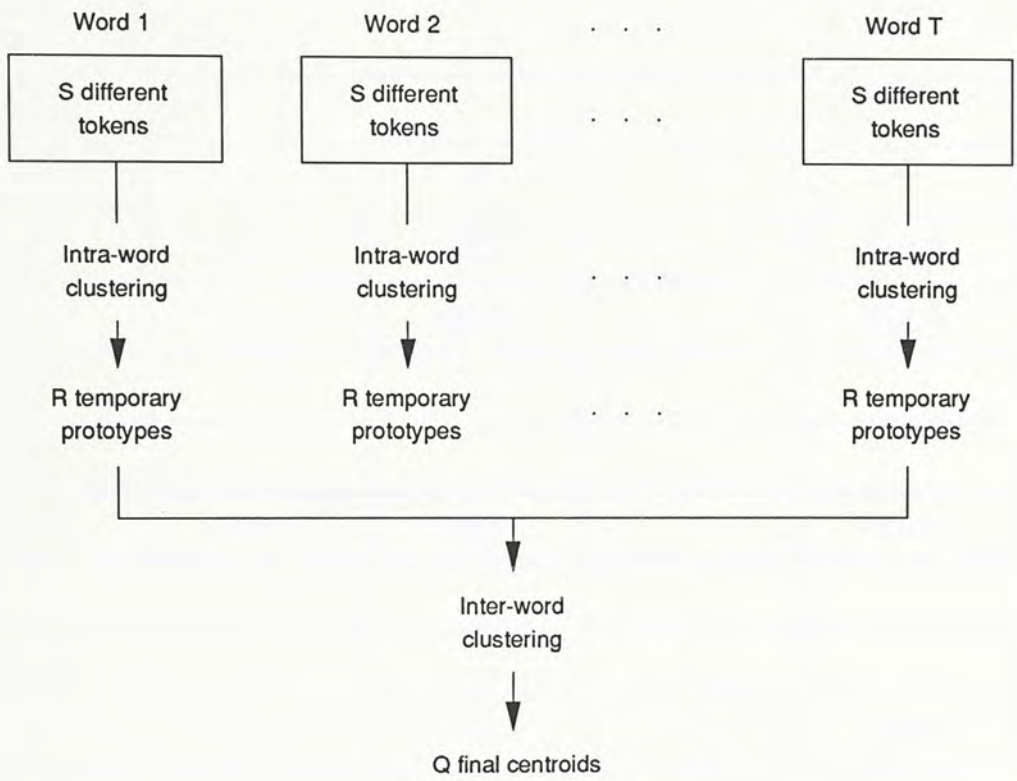


Figure 5-4 2-stage system training.

Let the number of training tokens for word j whose ETP vectors at the ith frame are best matched to that of a particular cluster centre k be $N_i(j,k)$. The probability of occurrence for this word j to appear in the kth cluster at segment i is then defined as

$$P_i(j,k) = \frac{N_i(j,k)}{\sum_{l=1}^{Q} N_i(j,l)} = \frac{N_i(j,k)}{S} \qquad (5.3)$$

This procedure is then repeated for all the 16 time frames and finally a 3-dimensional probability table is, therefore, created with its element given by $P_i(j,k)$,

$1 \le i \le 16$, $1 \le j \le T$ and $1 \le k \le Q$, and this will be used in the final matching decision of the recognition process. A part of the probability table, containing the probabilities of occurance for each word at each cluster in the ith frame is shown in Table 5-1.

	Word			
Cluster	1	2	. . .	T
1	$P_i(1,1)$	$P_i(2,1)$	. . .	$P_i(T,1)$
2	$P_i(1,2)$	$P_i(2,2)$	. . .	$P_i(T,2)$
.	.	.	.	.
.	.	.	.	.
k_i	$P_i(1,k_i)$	$P_i(2,k_i)$	. . .	$P_i(T,k_i)$
.	.	.	.	.
.	.	.	.	.
Q	$P_i(1,Q)$	$P_i(2,Q)$	. . .	$P_i(T,Q)$

Table 5-1 Probability table in time frame i.

5.3 Decision Making

In recognition, an unknown utterance is first subjected to preprocessing including endpoint detection, classification, bandpass filtering and segmentation. The ETP vectors are then extracted for all the 16 time frames. These feature vectors will be matched to its corresponding reference templates on a temporal basis. At segment i, let k_i be the selected reference vector with the minimum distance $D(k_i)$and the probability $P_i(j,k_i)$, for which the vector is closely resembled to that of word j at frame i, can then be retrieved from the probability table (Table 5-1). However, if the reference vector k_i' having a second minimum distance $D(k_i')$ close enough to

the minimum, the probability will be taken as the average of the two probabilities $P_i(j,k_i)$ and $P_i(j,k_i')$ retrieved for these two reference vectors respectively. These procedures are repeated until the last segment and the total probability that the unknown token resembles the word j will then be given by

$$P(j) = \prod_{i=1}^{16} P_i(j) \qquad 1 \le j \le T \tag{5.4}$$

and

$$P_i(j) = f \cdot P_i(j,k_i) + g \cdot P_i(j,k_i') \tag{5.5}$$

$$\text{where } f = \begin{cases} 1 \\ \text{and} \\ 0.5 \end{cases} \quad g = \begin{cases} 0 & \text{if } D(k_i') > \varepsilon \cdot D(k_i) \\ \\ 0.5 & \text{if } D(k_i') < \varepsilon \cdot D(k_i) \end{cases}$$

In addition, in order to reduce word confusion, a fricative/voice classification [32] has been used in which the vocabulary is divided into two groups depending on the phonetic labelling of their initial regions. If the input token is classified into a particular group, the possibility for those words in the other group will be neglected and only the probability of those words in the particular group will be computed. The one with the highest probability of resemblance is taken as the recognised word unless there is little differentiation in probability between the most probable and the next probable. A difference measure, R(m,n), for the probability of resemblance is, thus, introduced to determine the degree of closeness between the highest probability P(m) and the next highest P(n) obtained for word m and n respectively. This is defined as

$$R(m,n) = \frac{P(m) - P(n)}{P(n)} \tag{5.6}$$

Now if, $R(m,n) > \delta$, where δ is a predetermined threshold, the utterance is taken as the word m. Otherwise, a 2nd level decision will be performed. In this case, the variance σ_m^2 and σ_n^2, of the probability distribution of the sixteen segments for word m and n respectively are calculated according to the formula

$$\sigma_u^2 = \sum_{i=1}^{16} \left\{ P_i(u) - \sum_{i=1}^{16} \frac{P_i(u)}{16} \right\}^2 \qquad \text{where u=m,n.} \tag{5.7}$$

The recognition decision rule in this stage is as follow:

(1)	$\sigma_m^2 - \sigma_n^2 \leq \beta$	recognised word is m
(2)	otherwise	rejected

The value of the thresholds ε, δ and β are all determined experimentally and will in general affect the overall recognition accuracy of the system.

5.4 System Evaluation and Results

The system has preliminary been evaluated using the ten Cantonese digits. Two sets of input utterances have been used in various recognition tests. They are:

Data Set A: A total of 1000 tokens for the ten Cantonese digits were produced by 20 different speakers including male and female. Each speaker was requested to utter each digit 5 times and hence there was one hundred tokens for each word.

Data Set B: A total of 1500 tokens for the ten Cantonese digits were produced by another 15 speakers who did not participate in preparing Data Set A. In this case, each speaker was requested to utter each digit 10 times.

All the utterances were recorded through a microphone to a cassette recorder. The signal was then bandpass filtered to telephone bandwidth (100 Hz to 3.3 kHz) and digitized at 8 kHz with 12-bit resolution. ETP vectors of these utterances were extracted according to the procedures described previously. The two sets of data were exploited in the following experiments for both trained and untrained speaker recognition.

Experiment 1: Recognition for trained speakers

part (i) Half of the utterances from both Data Set A and B were used for training while the rest were used as input for recognition. That is T=10 and S=125.

part (ii) The utterances for training and testing as in part (i) were interchanged.

Experiment 2: Recognition for untrained speakers

part (i) Data Set A was used for training while Data Set B was used as input tokens for recognition. That is T=10 and S=100.

part (ii) Again, the tokens for training and testing as in part 2(i) were interchanged. But in this case, S=150.

Many different values for R and Q have been tried and finally, R=8 and Q=30 was chosen in our tests to give a compromise in recognition results and computation complexity. When the probabilities were multiplied together, it was noted that these probabilities could become very small and to avoid data underflow, a proper scaling had to be incorporated. In our case, each probability was in fact scaled up 100 time. In addition, if any of the probability of resemblance for a particular word at a time frame was zero, then this word would never be recognized no matter how high the probabilities were in the other segments. To alleviate this problem, we have set a zero probability to a fairly small number, typically, 5×10^{-6} and so far, no difficulties in matching due to the bias against a word that has only a small probability in any

one of the 16 probabilities were encountered. For the system thresholds ε, δ and β, they were generally set to 1.2, 100 and 0 respectively to give good recognition scores. The results for the mentioned experiments were tabulated in Table 5-2 while the confusion matrices for the corresponding experiments were tabulated in Table 5-3 (a) to (d).

Experiment	Score %	Error %	Reject %
1(i)	98.16	1.52	0.32
1(ii)	97.60	2.08	0.30
2(i)	95.00	4.60	0.40
2(ii)	94.60	5.10	0.30

Table 5-2 Speech Recognition Results

Test digit	Recognised as digit										Reject
	1	2	3	4	5	6	7	8	9	10	
1	145	0	0	0	0	0	0	5	0	0	0
2	0	146	0	0	4	0	0	0	0	0	0
3	0	0	140	0	0	0	1	0	0	8	1
4	0	0	0	150	0	0	0	0	0	0	0
5	0	3	0	0	144	3	0	0	0	0	0
6	4	0	0	0	0	140	0	6	0	0	0
7	0	0	2	0	0	0	130	0	0	14	4
8	0	0	0	0	0	2	0	148	0	0	0
9	0	0	0	0	0	4	0	0	146	0	0
10	0	0	0	0	0	0	13	0	0	136	1

Table 5-3(a) Confusion matrix for Experiment 1(i).

Test digit	Recognised as digit										Reject
	1	2	3	4	5	6	7	8	9	10	
1	97	0	0	0	0	1	0	2	0	0	0
2	0	100	0	0	0	0	0	0	0	0	0
3	0	0	98	0	0	0	1	0	0	1	0
4	0	0	0	100	0	0	0	0	0	0	0
5	0	11	0	0	89	0	0	0	0	0	0
6	0	0	0	0	0	95	0	0	5	0	0
7	0	0	5	0	0	0	75	0	0	18	2
8	0	0	0	0	0	0	0	100	0	0	0
9	0	0	0	0	0	0	0	0	100	0	0
10	0	0	1	0	0	0	6	0	0	92	1

Table 5-3(b) Confusion matrix for Experiment 1(ii).

Test digit	Recognised as digit										Reject
	1	2	3	4	5	6	7	8	9	10	
1	125	0	0	0	0	0	0	0	0	0	0
2	0	125	0	0	0	0	0	0	0	0	0
3	0	0	120	0	0	0	2	0	0	2	1
4	0	0	0	125	0	0	0	0	0	0	0
5	0	3	0	0	122	0	0	0	0	0	0
6	0	0	0	0	0	123	0	1	0	0	1
7	0	0	2	0	0	0	115	0	0	8	0
8	0	0	0	0	0	0	0	125	0	0	0
9	0	0	0	0	0	0	0	0	125	0	0
10	0	0	2	0	0	0	6	0	0	115	2

Table 5-3(c) Confusion matrix for Experiment 2(i).

Test digit	Recognised as digit										Reject
	1	2	3	4	5	6	7	8	9	10	
1	125	0	0	0	0	0	0	0	0	0	0
2	0	125	0	0	0	0	0	0	0	0	0
3	0	0	123	0	0	0	2	0	0	0	0
4	0	0	0	125	0	0	0	0	0	0	0
5	0	0	0	0	125	0	0	0	0	0	0
6	0	0	0	0	0	121	0	1	2	0	1
7	0	0	2	0	0	0	118	0	0	3	2
8	0	0	0	0	0	0	0	125	0	0	0
9	0	0	0	0	0	1	0	0	124	0	0
10	0	0	1	0	0	0	7	0	0	116	1

Table 5-3(d) Confusion matrix for Experiment 2(ii).

5.5 Observation

The average recognition rate for trained and untrained speakers were found to be 97.88% and 94.8% respectively. The accuracy of this speech recognizer was roughly 3.8% better than the one that simply uses Euclidean distance for matching. When using a statistical probability criterion for final matching decision, it is effectively comparing the similarities and dissimilarities between an unknown token with each of the reference words. Furthermore, the dynamic properties of the feature parameters are preserved by clustering the ETP vectors on a segmental basis. There is only marginal improvement in accuracy by increasing the number of templates for each word at each time frame, but the extra computation demanded for training and testing is certainly not justified. The introduction of a 2-pass decision based on the distribution of the probability of resemblance has reduced the number of errors in our tests. Because of its simplicity, the recognition algorithm can be implemented on a microcomputer. Although the evaluation uses a limited vocabulary of ten digits only, the results are promising. The only disadvantage of the system is probably the relatively lengthy training process.

Chapter 6

Conclusion and Discussion

Energy-time profile (ETP), extracted from the outputs of 5 consecutive bandpass filters, has been used simultaneously to carry speaker characteristics and speech contents from the tokens uttered by Cantonese speakers. In the speaker verification and identification experiments, ETP, when used and defined in a proper way, has been proved to be an effective parameter to distinguish speakers from their voice. In recognizing a speaker, instead of using the entire utterance for comparison, a word-by-word matching approach has been adopted which allows the LTW to be applied while high accuracy can be achieved at the same time.

In the speaker verification system, the verification results indicate that the methodology used is sound, whatever the two different time warping techniques are employed. As expected, the results obtained by using DTW is, obviously, better than that by using LTW when the number of digits used in the test utterances are the same. However, as the verification accuracies in both cases increase rapidly with the number of digits used in the test utterance and consequently, it is possible for the performance using LTW to catch up with that using DTW by employing more digits in the input utterances. In my case, an extra of 2 digits in the utterance is sufficient for the LTW to perform better than for the DTW with the distance threshold chosen either at the minimum total error point or at a practical value. Under a practical consideration, because of the word-by-word sequential matching approach used in the proposed speaker verification system, increasing the number of digits in the input test utterance only increase the recognition time in an additive way so that the ultimate verification time required for LTW will be still much less than that for DTW. In my case, more than 10 times faster in the recognition speed can be achieved by employing LTW instead of DTW but with a comparable performance. Even

though in parallel processing, which can be made use of to achieve high speed recognition under the word-by-word approach, the increase in the number of uttering digits will only require a comparatively greater number of "discrete word processing units" in using the LTW. This cost, however, is worth paying to change for the a high speed and accurate verification system. Consequently, dividing the whole utterance into units of discrete word while decision is made upon the accumulated results of these units, elimination of DTW for speaker verification becomes possible for mono-syllabic language speakers while high accuracy can be maintained.

Moreover, as the verification accuracy can be increased by simply including more words in the testing utterances, it can be, therefore, applied in a wide range of applications in which different accuracy standards are required. This allowed flexibility should be made use of carefully with the considerations on the system specification, implementation cost and other limitations in the system design. Nevertheless, an average verification score up to 99.39% has obtained in my experiment with 5 distinct digits in the input utterance using only LTW. This performance is already sufficient in various kinds of real time application.

Similarly, the above advantages will also be expected in the speaker identification system, though these have not been studied in the project. In the proposed speaker identification system, a 5-digit sentencs is suggested as the testing utterance together with a set of decision rules. Both the results obtained by using DTW and LTW have been tabulated in Table 4-2. Obviously and as expected, the result on using DTW is better than that using LTW. At the point of minimum total error (distance threshold = 1.3 and 1.6 respectively for DTW and LTW), an identification accuracy of 96.21% was obtained on using LTW which is less than 98.81% obtained on using DTW. But the increase in error for LTW over that for DTW falls entirely on the rejection of system users. Although such error will cause inconvenience during identification and usually a second attempt will be necessary, it is nevertheless worth

accepting to exchange for the large amount of computational time saved which is crucial in low cost real time recognition. It is therefore feasible to make use of LTW in identifying speakers without introduction of high cost error in the identification process.

The setting of decision rules in the identification process are based on a majority rule and are finally fixed on an ad hoc basis. Observe Fig. 4-2 and 4-3 which show the identification error against the distance threshold for DTW and LTW respectively, the true reject error converge to a constant value at a greater preset distance threshold. This implies that there exists a minimum percentage in the rejecting of legal users due to the introduction of 1st pass criterion for the selection of the "qualified candidate". On the other hand, this also allows a certain amount of illegal users to enter into the 2nd pass decision which is in fact indicated by the drastic increase of false accept error with the distance threshold. The proposed criterion for the 1st pass decision has been set in a reasonable compromise on the above two conditions. Finally, the rules for the 2nd pass decision which is actually a verification of the selected identity but under a relatively tight standard, is introduced to keep the false acceptance error as small as possible while maintaining the true reject error in a reasonable small value. Though the proposed decision rules can be adjusted and modified for different applications or when the number of words used in the test utterance is changed, it provides a very good guideline for the decision and in fact the experimental results are satisfactory.

Though the word-by-word matching approach which has allowed the implementation of LTW on Cantonese to achieve high score in speaker recognition, it demands the successful separation of the input utterance into units of discrete word. Uttering the input token digit by digit is one of the method to fulfill this requirement but is somewhat unnatural in real application. As stated before, separating the Cantonese words from a sentence is not difficult due to its simple energy variation,

it is still possible that the speaking behaviour of the speaker on a single digit will alter between different digit sequences. However, to study this correlation of the speaking behaviour on the digit order, a huge amount of trials will be required and is left for the study in the real time system. To allow reasonable recognition experiments, only isolated digits uttered by speakers has been recorded and from these recorded tokens, combinations of various digit patterns has been selected as the input test utterance. However, serious degradation in the system performance due to the existence of this possible correlation is not expected.

Finally, ETP is used for the speech recognition of isolated Cantonese words. This has already been used in [32] with a simple template matching approach and proved to be effective. However, a further improvement is obtained by using the proposed probabilistic criterion. Even though the proposed algorithm is not emerged from any well structured derivation, statistical phenomenon of spoken words which distribute in a clustering manner has provided a solid ground for this method and is experimentally found to be practically effective. From the results shown in Table 5-2, the introduction of the 2-stage clustering technique has been proved to be useful for system training. Although the clustering process is somewhat laborious, it can be performed off line in a microcomputer without much difficulty. In fact, the system evaluation, especially the system training process, has been made possible by the use of this 2-stage clustering technique to operate in a micro-computer in which the system memory is usually small. Though the average score for the open-test (tester's utterance has not been used in the training process) is only 94.8%, an average score of 97.88% is obtained for the semi-open test (the tester's utterances has been used as the training data) which is very useful for the proposed speaker verification system to extract the speaker's claimed identity because the user is expected to be one of the legal candidates who have participated in the system training.

Instead of its effectiveness in the extraction of speaker's characteristics and speech features, ETP has little tolerance to the intensity distortion caused by the transmission or backgound noise. In order to reduce experimental complexity, simulation has been performed only on comparatively clean speech of discrete Cantonese digits while speech from a noisy environment or transmitted through telephone channels have not been tried in the experiment. However, even though the effect on the system performance due to the introduction of noise has not been studied together in this project, the experimental results has promised good system performance for speaker recognition using ETP under a reasonable quiet condition and at a certain level of loudness of testing speech.

To summarize, an automatic speaker verification system using ETP, in which speaker identity is derived from the contents of the input token, is described for the recognition of Cantonese speakers. The system, in fact, can be further applied with any mono-syllabic language such as Mandarine and many other Chinese dialects having similar phonetic sturcture with Cantonese. Moreover, by employing the simple algorithm on a word-by-word approach, LTW can be applied on a discrete word basis to achieve fast speaker recognition with high accuracy so that real time implementation is possible using only low cost hardware.

References

[1] F.K. Soong, A.E. Rosenberg, L.R. Rabiner and B.H. Juang, " A Vector Quantization Approach to Speaker Recognition," ICASSP-85, pp. 387-390, 1985.

[2] S. Furui, " Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 254-272, Apr. 1981.

[3] J. Makhoul, " Linear Prediction: A Tutorial Review, " Proc. IEEE, vol. 63, pp. 561-580, Apr. 1975.

[4] A.E. Rosenberg and M.R. Sambur, " New Techniques for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 169-175, Apr. 1975.

[5] R.C. Lummis, " Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., vol. AU-19, pp. 80-89, Apr. 1973.

[6] B.S. Atal, " Automatic Recognition of Speakers from Their Voices," Proc. IEEE, vol. 64, pp. 460-475, Apr. 1976.

[7] J.D. Markel, B.T. Oshika and A.H. Gray, Jr., " Long-Term Feature Average for Speaker Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 330-337, Aug. 1977.

[8] J.D. Markel and S.B. Davis, " Text-Independent Speaker Identification From A Large Linguistically Unconstrained Time-Spaced Data Base," Proc. ICASSP-78, pp. 287-290, 1978.

[9] L.L. Pfeifer, " New Techniques for Text-Independent Speaker Identification," Proc. ICASSP-78, pp. 283-286, 1978.

[10] H. Matsumoto and T. Nimura, " Text-Independent Speaker Identification Based on Piecewise Canonical Discriminant Analysis," ICASSP-78, pp. 291-294, 1978.

[11] A.E. Rosenberg, " Automatic Speaker Verification: A Review," Proc. IEEE, vol. 64, pp. 475-487, Apr. 1976.

[12] G.R. Doddington, " Speaker Recognition - Identifying People from their Voices," Proc. IEEE, vol. 73, pp. 1651-1664, Nov. 1985.

[13] S. Furui, " Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 342-350, Jun. 1981.

[14] W.S. Mohn, Jr., " Two Statistical Feature Evaluation Techniwues Applied to Speaker Identification," IEEE Trans. Comput., vol. C-20, pp. 979-987, Sept. 1971.

[15] D.K. Burton, " Text-Dependent Speaker Verification Using Vector Quantization Source Coding," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, pp. 133-143, Feb. 1987.

[16] F.K. Soong and A.E. Rosenberg, " On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-36, pp. 871-879, Jun. 1988.

[17] S.L. Wong, " A Chinese Syllabary Pronounced According to the Dialect of Canton,"

[18] C.Y.Y. Fok, " A Perceptual Study of Tones in Cantonese," Centre of Asian Studies, HKU.

[19] A. Komatsu, A. Ichikawa, Nakata, Y.Asakawa and H. Matsuzaka, " Phoneme Recognition in Continuous Speech," Proc. ICASSP-82, pp. 883-886, 1982.

[20] W.M. Lai, " Efficient Algorithm for Speech Recognition of Cantonese," M. Phil. Thesis, CUHK, 1987.

[21] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, " An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp.777-785, August 1981.

[22] J.J. Dubnowski, D.W. Schafer and L.R. Rabiner, " Real Time Digital Hardware Pitch Detector," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, no. 1, pp.2-8, February, 1976.

[23] A.E. Rosenberg and M.R. Sambur, " New Techniques for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp.169-175, April, 1975.

[24] R.C. Lummis, " Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., vol. AU-21, pp.80-89, April, 1973.

[25] L.R. Rabiner and R.W. Schafer, " Digital Processing of Speech Signals," Prentice-Hall Inc., pp.120-126, 1978.

[26] L.R. Rabiner, A.E. Rosenberg and S.E. Levinson, " Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp.575-582, December, 1978.

[27] S. Furui, " Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp.254-272, April, 1981.

[28] P.C. Ching, C.K. Yu, K.M. Tse and Y.T. Chan, " Speech Recognition of Cantonese Based on a Probabilistic Approach," Proc. 1988 Int. Conf. on Computer Processing of Chinese and Oriental Languages, pp.561-564.

[29] H. Sakoe and S. Chiba, " Dynamic Programming Algorithm Optimaization for Spoken Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 43-49, Feb. 1978.

[30] J. Cox, " Hidden Markov Models for Automatic Speech Recognition: Theory and Application," Br. Telecom. Technol. J., vol 6, 1988, No. 2, pp.105-115.

[31] J.G. Wilpon and L.R. Rabiner, " A Modified K-Mean Clustering Algorithm for Use in Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-33, No. 3, pp. 587-594, June 1985.

[32] W.M. Lai, P.C. Ching and Y.T. Chan, " Discrete Word Recognition using Energy-time Profiles," Int. J. Electronics, 63, 1987, No.6, pp.857-865.