Large Vocabulary Cantonese Speech Recognition Using Neural Networks

Tsik Chung Wai Benjamin
（植頌偉）

Supervisors:
Dr. Chan Lai Wan
Dr. Ching Pak Chung

Dissertation submitted in partial fulfillment of the
requirements for the degree of
Master of Philosophy
in
Computer Science Department
The Chinese University of Hong Kong

Hong Kong
September, 1994

Acknowledgment

I am grateful to my supervisors, Dr. L. W. Chan and Dr. P. C. Ching, for their patience, guidance, advice and encouragement. I would also like to give my special thanks to Mr. Tan Lee, Mr. Brian Mak and Mr. Alfred Ng for providing comments and ideas on this research project and helping in setting up the recording procedure of the speech data.

Abstract

Cantonese is the second most widely spoken Chinese dialect after Mandarin, and it is the mother tongue of about 40 million peoples all over the world. There is importance to develop an automatic Cantonese speech recognition system. However, previous researches on Cantonese speech recognition mainly aimed on tone classification and were in a limited vocabulary size.

An efficient *large vocabulary* isolated Cantonese syllable recognition system has been proposed to deal with all known Cantonese syllables. For a large vocabulary of 1470 different syllables in Cantonese, problem decomposition is employed for its efficient and effective recognition. A hierarchical neural networks model is designed for the recognition system, fitting the *monosyllabic* and *tonal* nature of Cantonese speech. The isolated Cantonese syllables are segmented into their constituting phonemic segments: *initials, syllabic segments* and *endings*. The feature parameters for phoneme and tone recognition are first passed into 4 MLP neural networks in the primary level of the hierarchical model. The outputs of these classifiers are then fed into a syllable classifier in the secondary level of the model. The syllable classifier is implemented in two different approaches: concatenation with correction and Fuzzy ART.

The performance of the system under speaker-dependent setting has been evaluated by computer simulation. The classification accuracy of the system on *initials, syllabic segments*, *endings*, tones and syllables are 58.6%, 76.5%, 74.8%, 91.2% and 45.4% respectively for the Cantonese database used.

Table of Contents

1 Introduction

1.1 Automatic Speech Recognition

In daily life, most people communicate with others by means of natural speech. However, natural speech is rarely used in the current human-machine communication. As revealed from most of the science fictions, human beings always dream of taking control on machines, such as robots and computers, by voices rather than pressing buttons by hands. To approach such fantastic future, the first step is to teach the machines to learn to recognize the human speech.

It is very useful to have successful automatic speech recognition. Voice input to computers offers a number of advantages by providing a natural, fast, hands-free, eyes-free and location-free input medium (Lee, 1989). Besides of the simplified human-machine communication, automatic speech recognition is also useful for the hearing impaired and physically disabled.

Unfortunately, current researches still have little success on speaker-independent continuous speech recognition in which most ordinary people find no difficulties on them. It is proved to be a difficult task to duplicate the speech recognition ability with machines, though children learn to understand speech with little explicit supervision and adults take speech recognition ability for granted (Lippmann, 1989). This is mainly due to the following problems:

- the variability and overlap of information in the acoustic signal
- the need for high computational rates
- the multiplicity of language analyses
- the lack of any comprehensive theory of speech recognition

Although many speech recognition strategies have been proposed and implemented, a few systems demonstrated the feasibility of accurately recognizing human speech. Most existing speech recognizers performs well only in constrained tasks. Some of the constraints imposed on the well-developed speech recognition systems are:

- speaker-dependent instead of speaker-independent
- isolated words instead of connected speech or continuous speech
- small vocabulary size instead of large vocabulary size

The vocabulary size of the speech recognition systems can be classified as the following table (Rabiner, 1992):

Classes of Speech Recognition System	Vocabulary Size on the order of
Very Small Vocabulary	10
Moderate Vocabulary	100
Large Vocabulary	1000
Very Large Vocabulary	10000

Table 1.1 Classes of Vocabulary Size in Speech Recognition Systems

1.2 Cantonese Speech Recognition

Cantonese is the second most widely spoken Chinese dialect after Mandarin (the official spoken Chinese language). It is the mother tongue of about 40 million people as it is widely spoken by Chinese people in Southern China, Hong Kong, Macao and overseas Chinese in Southeast Asia, Northern America, United Kingdom and Australia. Over 90% of the Hong Kong citizens speak Cantonese in their daily lives. Since Hong Kong is an important Asian international trade center, there are needs on high-technology application support, such as Cantonese speech input to machines. As one of the 6 million Hong Kong citizens, the author is motivated to perform research on Cantonese speech recognition very much.

Most of the previous works on automatic speech recognition were done on the tasks for Western languages, such as English, French, etc. The special features of Chinese dialects, such as the *tonal* and *monosyllabic* nature, are not considered in most of the automatic speech recognition systems.

Previous researches on automatic Cantonese speech recognition mainly aimed on tone classification and were in a limited vocabulary size (Cheng, 1991; Ng, 1992; Lee *et al*, 1993). It is because Cantonese speech recognition is still not a well-developed research area, whereas the tone classification is the fundamental task of speech recognition on *tonal* languages.

1.3 Neural Networks

Artificial neural networks are the models that attempt to achieve good performance via dense interconnection of simple computational elements. The structure of neural networks is based on our present understanding of biological nervous systems. In human brain, there are approximately hundred billion (10^{11}) neurons, which are the basic processing units of brain. Each of these neurons is connected to about ten thousand (10^4) others. There are also artificial neurons that model the biological neurons in the neural networks.

Study on artificial neural networks has a long history of more than 40 years. However, in the past decade, the field has gained new interest as new networks topology and algorithms were developed. One of the most important development is the discovery of the general delta rule for error backpropagation (Rumelhart *et al*, 1986). This made the supervised training to find the internal parameters of multi-layer perceptron (MLP) possible.

Recent interest is also driven by the realization that enormous amounts of processing will be required by human-like performance in the areas of speech and image recognition. Artificial neural networks can provide processing capacity using many simple processing units operating in parallel.

There are numerous types of artificial neural networks that have been well-developed, including multi-layer perceptron (MLP), Kohonen self-organization maps, Hopfield Networks, Adaptive Resonance Theory (ART), etc. (Lippmann, 1987)

Artificial neural networks become popular classifiers for they have shown successful in various applications, including those for speech recognition (Lippmann, 1987, 1989).

1.4 About this Thesis

An efficient *large vocabulary* isolated Cantonese syllable recognition system is proposed. This system aims at recognizing all Cantonese syllables[1]. The vocabulary size is on the order of 1000, so there is not any constraint of small vocabulary size imposed. Phoneme-based approach is used because there are only limited number of phonemes found in Cantonese. We try to divide the syllables into their corresponding phonemes and tones according to the phonology of Cantonese. A hierarchical neural networks system is designed to recognize the phonemes and tones, and to recognize the syllables by integrating the recognition results of the phonemes and tones. Since each speaker has to provide a large set of speech data, we have collected speech data for speaker-dependent recognition only. So, experiments are done at speaker-dependent setting only.

In this thesis, the phonology of Cantonese is introduced in the next chapter. The review on automatic speech recognition systems is given in chapter 3. Chapter 4 discusses about the Cantonese speech corpus and the speech database acquisition. The feature parameters extraction process of the speech signal is described in chapter 5 in details. Chapter 6 shows the design of our speech recognition system. The computer experiments and the discussions on their results are given in chapter 7. Chapter 8 suggests some further works. The last chapter gives the conclusions.

---

1 There exists some Cantonese syllables collected in some old syllabaries, such as Wong (1941), which are not known by the people nowadays anymore. These "out-dated" syllables are excluded in our system.

2 The Phonology of Cantonese

There are over 10000 characters in Chinese language. In the standard traditional Chinese character set used by computers, there are 13053 characters defined. According to Wong (1941), all Chinese characters are pronounced as around 1800 different syllables in Cantonese dialect.

Cantonese is a *monosyllabic* language, which means that every written Chinese character is pronounced as a single syllable. Cantonese is also a *tonal* language, which means that syllables pronounced with different tones (at different pitches) give different meanings and Chinese characters in most cases. For examples, the syllable /mai/ can mean either "buy" (買) or "sell" (賣), according to the tone it is pronounced with.

According to the phonology of Cantonese (Hashimoto, 1972), a Cantonese syllable consists of two parts: syllabic structure and tone. For example, for the syllable corresponding to the character "班", the syllabic structure is /ban/ and the tone is tone 1. The syllabic structure consists of two parts: an optional *initial* (the phoneme which begins the syllable) and a *final* (the rhyme which ends the syllable). A *final* can further be divided into a *syllabic segment* and an optional *ending*. In the previous example (/ban/), the *initial* is /b/, the *final* is /an/, the *syllabic segment* is /a/ and the *ending* is /n/. The composition of a Cantonese syllable is shown in Figure 2.1.

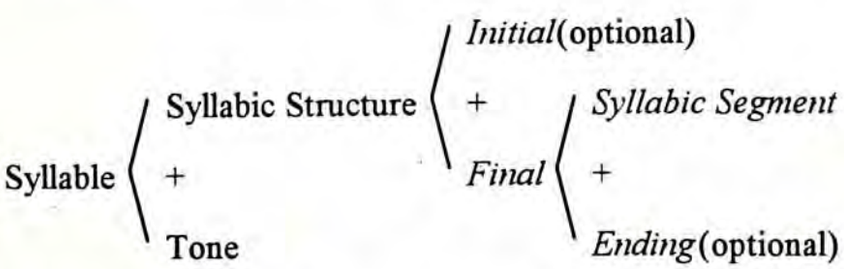

Figure 2.1 The Composition of a Cantonese Syllable

2.1 The Syllabic Structure of Cantonese Syllable

There are totally 19 *initials* and 53 *finals* (composed of 10 *syllabic segments* and 8 *endings*) in Cantonese. The *initial* can be null, a consonant or a glide. The *syllabic segment* can be a vowel or a syllabic nasal consonant. The *ending* can be either null, a glide (forming diphthong with vowel from *syllabic segment*), a nasal or a plosive stop consonant. A summary of the basic phonetic units of Cantonese is given in Table 2.1.

Classes (Number)	Sub-Classes	Phonetic Units
Initials (19)	Aspirated Plosives	p, t, k, kw
	Non-aspirated Plosives	b, d, g, gw
	Affricatives	ts, dz
	Fricatives	s, f, h
	Lateral	l
	Nasals	m, n, ŋ
	Glides	j, w
Syllabic Segments (10)	Vowels	a, ɐ , ɛ/e, i, ɔ/o, oe, u, y
	Syllabic Nasals	m̩, ŋ̩
Endings (8)	Glides	i/y, u
	Nasals	m, n, ŋ
	Plosive Stops	p, t, k
Tones (9)	Non-entering	tones 1, 2, 3, 4, 5, 6
	Entering	tones 7, 8, 9

Table 2.1 Basic Phonetic Units of Cantonese

One of the Cantonese phonetic concatenation rules is that only 53 combinations of the *syllabic segments* and *endings* are valid to form *finals*. These 53 valid *finals* are shown in Table 2.2.

Finals			null	*Endings*								
				Glides		Nasals			Plosives		
				i / y	u	m	n	ŋ	p	t	k
Syllabic Segments	Vowels	a	a	ai	au	am	an	aŋ	ap	at	ak
		ɐ		ɐi	ɐu	ɐm	ɐn	ɐŋ	ɐp	ɐt	ɐk
		ɛ/e	ɛ	ei				ɛŋ			ɛk
		i	i		iu	im	in	iŋ	ip	it	ik
		ɔ/o	ɔ	ɔi	ou		ɔn	ɔŋ		ɔt	ɔk
		oe	oe	oey			oen	oeŋ		oet	oek
		u	u	ui			un	uŋ		ut	uk
		y	y				yn			yt	
	Syllabic Nasals	m̩	m̩								
		ŋ̩	ŋ̩								

Table 2.2 The 53 Valid *Finals* in Cantonese

2.2 The Tone System of Cantonese

A summary on the Cantonese tone system is shown in Table 2.3.

Tone Numbers	Traditional Class Names	Pitch Values
1	Upper Level (陰平)	5,5 or 5,3
2	Upper Rising (陰上)	3,5
3	Upper Going (陰去)	3,3
4	Lower Level (陽平)	2,1 or 1,1
5	Lower Rising (陽上)	1,3
6	Lower Going (陽去)	2,2
7 (1)	Upper Entering (陰入)	5,5
8 (3)	Middle Entering (中入)	3,3
9 (6)	Lower Entering (陽入)	2,2

Table 2.3 Cantonese Tone System

In the above table, the tone numbers and the Chinese traditional class names are adopted from Chow & Yiu (1988). The English translation of the traditional class names are adopted from Wong (1941). The pitch values are adopted from Ho (1987). The tone numbers in brackets are for the 6-tone system.

The two digits for the pitch values denote the beginning and the ending pitch values respectively. The digits 1, 2, 3, 4, 5 are corresponding to "do", "re", "mi", "fa" and "so" (the first, second, third, fourth and fifth notes in the musical octave) in music. According to Ho (1987), these pitch values, which are used by the traditional phonologists, give a general picture only. It does not mean that there are no differences among the pitch ranges and keys when the people speak. For examples,

some people pronounced the syllable in tones 3 and 8 with the pitch values 4,4. However, the values can give the relative pitch levels of the tones spoken by all the people. For examples, tone 3 must be higher than tone 6; tone 3 must have the same pitch values as tone 8 and tone 6 must have the same pitch values as tone 9.

There are two main kinds of viewpoints on the number of tones in Cantonese. The traditional one suggested that there are 9 tones in Cantonese and the modern one suggested that there are only 6 tones in Cantonese. The difference is due to the viewpoints on "entering tones" (入聲). In the traditional Chinese phonology, tone does not only relate to pitch levels, but also relates to *ending* phonemes. For examples, "幻" (/wan/ in tone 6) and "滑" (/wat/ in tone 9) are regarded as having the same base syllable but different tones, though their pitch levels are the same. Traditional Chinese phonologists regard that their differences are in tones rather than in *finals*, because their *endings* /n/ and /t/ have common places of articulation. Similar cases exist on between /m/ and /p/ and also between /ŋ/ and /k/.

So, in the 9-tone system, these 9 tones can be divided into two main groups: 6 non-entering tones and 3 entering tones. All syllables in entering tones must have plosive stops *endings* (/p/, /t/ or /k/) and all syllables in non-entering tones must have either null *ending*, glide *endings* (/i/y/ or /u/) or nasal *endings* (/m/, /n/ or /ŋ/). Since the syllables in entering tones have stop *endings*, they can be characterized by their relatively shorter duration.

On the other hand, in the 6-tone system, there is no division between entering tones and non-entering tones. The original 3 entering tones (upper entering, middle entering and lower entering) are combined with 3 other non-entering tones (upper

level, upper going and lower going) to form tones 1, 3 and 6 respectively. This is possible only when the pair of tones have same pitch value.

As the differences between syllables in non-entering tones and in entering tones are quite significant, the traditional 9-tone system is adopted in our design. Also, the rule that "syllables are pronounced in entering tones if and only if they have plosive stop *endings*" can be taken as one of the Cantonese phonetic concatenation rule for further usage in the recognition task.

3 Review of Automatic Speech Recognition Systems

Artificial Intelligence (AI) is the study of how to make computers acquire the human ability. One of the important human ability is perception. Perception is regarded as a difficult task because it involves noisy analog signals. Among the various ways of perception, hearing (including speech recognition) and vision are studied extensively in artificial intelligence research. We shall review some of the existing speech recognition systems in this chapter.

3.1 Hidden Markov Model Approach

The current best performing speech recognition algorithms use Hidden Markov Model (HMM) techniques (Lippmann, 1989). HMM is one of the statistically modeling methods. In speech recognition, HMM tries to capture the relationships between a speech unit (e.g. a word) and its acoustic signal stochastically.

The HMM approach provides a framework which includes an efficient decoding algorithm for use in recognition and an automatic supervised training algorithm. Let us take isolated word speech recognition as an example to briefly describe what have to be done in the HMM approach. For each word in the vocabulary, we must build an HMM. Then, for each unknown word which is to be recognized, we calculate the model likelihood for all possible models, and then select

the word whose model likelihood is the highest as the recognized word (Rabiner, 1989).

Since HMM theory does not specify the structure of implementation hardware, high computation and memory are required for large vocabulary, continuous speech recognition using current algorithms.

3.2 Neural Networks Approach

The performance of current speech recognition systems is far below that of humans. As the model based on the biological nervous system, neural networks offer the potential of providing massive parallelism, adaptation, and new algorithmic approaches to the problems in speech recognition (Lippmann, 1989). There are several neural network approaches to deal with the speech recognition problems. Some of them are discussed below.

3.2.1 Multi-Layer Perceptrons (MLP)

MLP classifiers have been applied to speech recognition problems more often than other neural network classifiers. There are excellent speaker-dependent MLP

recognizers using small sets of words and digits. They have the similar recognition accuracy as that of HMM recognizers (Lippmann, 1989).

As shown in Figure 3.1, MLP contains one input layer, one output layer and one to many hidden layers. To deal with the speech recognition problem, the feature parameters of the speech signal are applied to the input layer. The output layer is expected to give the recognition result: one node for each of the speech recognition goals. Hidden layers are required for good performance. They provide better performance, faster training and higher probability of convergence as compared with single-layer perceptron. An analysis indicated that hidden nodes often become feature detectors and differentiate between subsets of sound types such as consonants versus vowels (Lippmann, 1989).

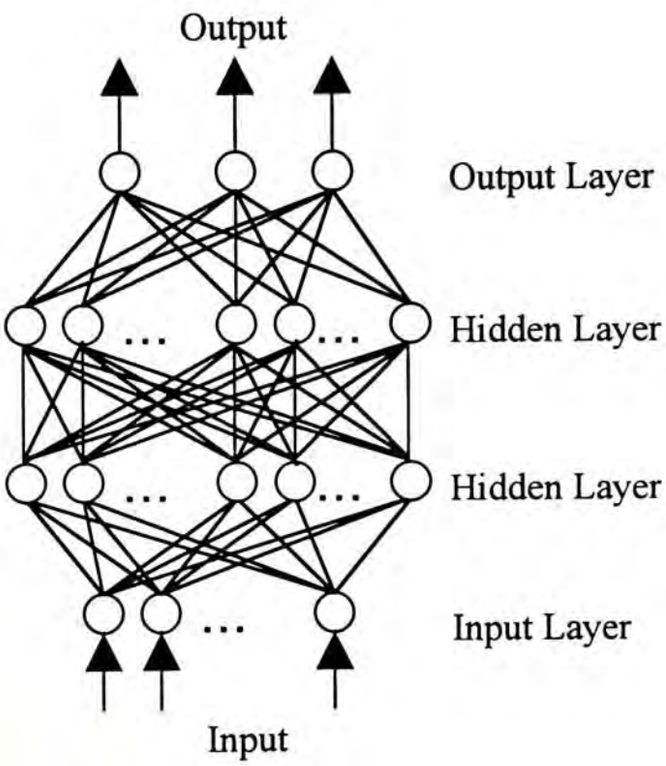

Figure 3.1 Multi-Layer Perceptron

3.2.2 Time-Delay Neural Networks (TDNN)

Waibel *et al* (1989) had proposed a Time-Delay Neural Network (TDNN) approach to speech recognition. The time-delay arrangement enables the network to discover acoustic-phonetic features and the temporal relationships between them, independent of position in time, and hence is not blurred by temporal shifts in the input.

The basic unit in most of the neural networks computes the weighted sum of its inputs and then passes the sum through a nonlinear function operator. In TDNN, the basic unit is modified by introducing delays D_1 through D_N as shown in Figure 3.2.

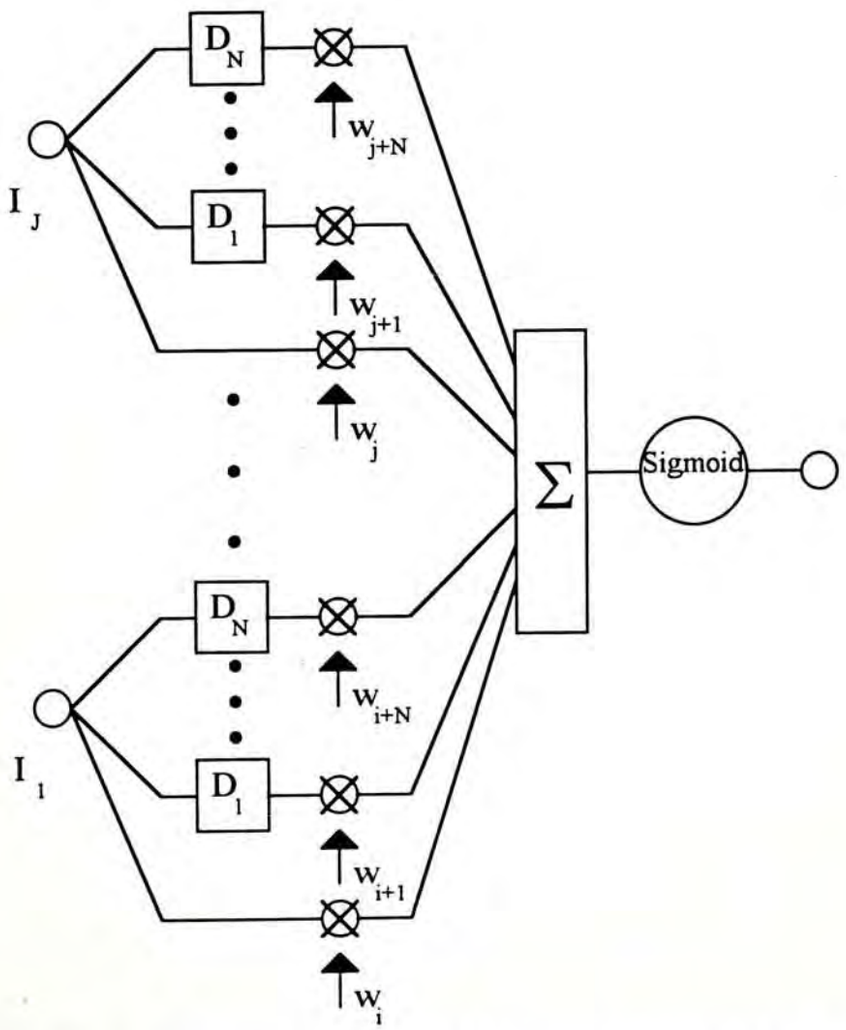

Figure 3.2 A Time-Delay Neural Network (TDNN) basic unit

The inputs are now multiplied by several weights, one for each delay and one for the undelayed input before the sum is computed. In this way, the unit has the ability to relate and compare current input to the past history of events.

The overall architecture of a TDNN is shown in Figure 3.3. Note that there are fixed delays among the input frames. The nodes are not fully connected between every two layers.

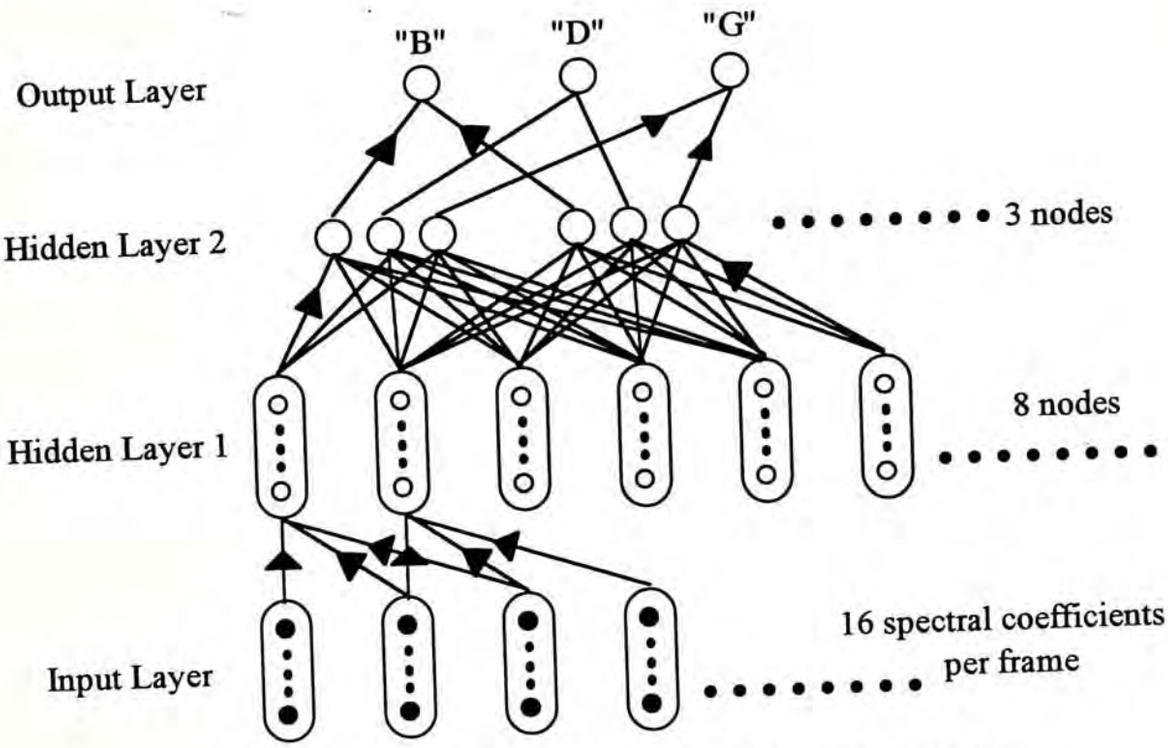

Figure 3.3 Time-Delay Neural Network (TDNN)

The backpropagation learning algorithm is used to train the TDNN. However, the weights of the corresponding connections in the time-shifted copies must be constrained to be the same. To achieve this, the regular backpropagation forward and backward passes are applied to all time-shifted copies as if they were separate events. Rather than changing the weights on time-shifted connections separately, each weight on corresponding connections is updated by the same value, which is the average of all corresponding time-delayed weight changes.

3.2.3 Recurrent Neural Networks

Recurrent neural networks are neural networks with recurrent connections as shown in Figure 3.4. They have not been used as extensively for speech recognition problems as feed-forward (i.e. MLP) neural networks because they are more difficult to train, analyze and design. Many types of recurrent neural networks have been proposed that can be trained with modified forms of backpropagation training algorithm (Lippmann, 1989). There are interests on using recurrent neural networks to deal with the speech recognition problems because the recurrent connections make the neural networks capable to accept time varying inputs.

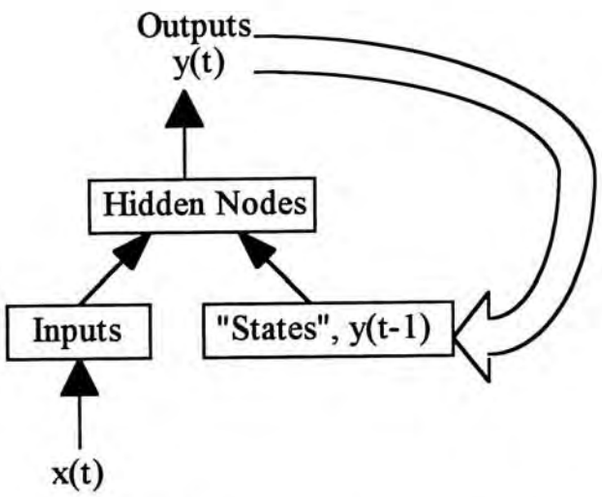

Figure 3.4 A Recurrent Neural Network

3.3 Integrated Approach

There is a trend to integrate multi-layer perceptron classifiers with conventional HMM recognizers to deal with the speech recognition problems. This may lead to improved recognition accuracy by combining the good discrimination of neural network classifiers with the automatic scoring and training algorithms used in HMM recognizers (Lippmann, 1989).

A multi-layer perceptron can be used to decrease the error rate of an HMM recognizer (Huang & Lippmann, 1988). The Viterbi backtraces from the HMM recognizer can be used to segment input speech frames and average HMM probability scores for segments are provided as inputs to the MLP classifier.

3.4 Mandarin and Cantonese Speech Recognition Systems

Some works on Mandarin and Cantonese speech recognition have been investigated (Yang *et al*, 1988; Liu *et al*, 1989; Chang *et al*, 1990; Wang *et al*, 1991; Cheng, 1991; Ng, 1992; Lee *et al*, 1993). The recognizer used, the vocabulary size and the accuracy for both tone and syllable recognition are summarized as follows in Table 3.1.

	Speech	Recog-nizer	Voca-bulary Size	Tone Recognition Accuracy	Syllable Recognition Accuracy
Yang *et al*, 1988	Mandarin	HMM	72	S.D.:- 98.33% S.I.:- 96.53%	---
Liu *et al*, 1989			72	S.I.:- 97.9%	---
Chang *et al*, 1990		MLP	148	S.I.:- 93.8%	---
Wang *et al*, 1991			1300	S.D.:- 97.92%	S.D.:- 90.14%
Cheng, 1991	Cantonese		234	S.D.:- 87% M.S.:- 77% S.I.:- 73%	---
Ng, 1992		HMM + MLP	108	M.S.:- 94.2%	M.S.:- 92.0%
Lee *et al*, 1993		MLP	234	S.D.:- 89.0% M.S.:- 87.6%	---

S.D. - speaker-dependent M.S. - multi-speaker S.I. - speaker-independent

Table 3.1 Summary of Previous Works on Mandarin and Cantonese Speech Recognition

Note that the tone and syllable recognition may be done in different modes, including speaker-dependent, multi-speaker and speaker-independent. Speaker-dependent mode refers to that both training and testing data are obtained from the same single speaker. Multi-speaker mode refers to that both training and testing use data obtained from the same group of speakers. Speaker-independent mode refers to that training uses one group of speakers and testing uses a separate group with no common members.

As the official spoken language of Chinese, development of Mandarin speech recognition is much matured than Cantonese. Chang *et al* (1990) announced that the tone recognition itself as a problem is not trivial, though there are only 4 tones in Mandarin. So, much efforts should be paid on this problem.

Cantonese tone system is more complicated than the Mandarin one because Cantonese tone system uses both the variation of pitch to characterize different tones while the Mandarin one uses only the variation of pitch to distinguish among the tones (Cheng, 1991). Also, there are 9 tones in Cantonese, which is more than twice of that in Mandarin (which has only 4 tones), the difficulties of Cantonese tone classification are thus increased accordingly. This can be shown by the difference in the recognition accuracy between Mandarin and Cantonese tone recognition.

Note that all the vocabulary size of the speech recognition systems except Wang *et al*'s (i.e. including all the Cantonese systems) are just on the order of 100, which is regarded as moderate vocabulary in Table 1.1. A larger vocabulary size will increase the difficulties of the recognition problems accordingly.

4 The Speech Corpus and Database

Since there was no established Cantonese speech database for automatic speech recognition, a new speech corpus has been designed and the corresponding database has been established.

4.1 Design of the Speech Corpus

To develop a large vocabulary isolated Cantonese syllable recognition system, the aim of the design of the new speech corpus is to cover all known Cantonese syllables. However, the number of all Cantonese syllables varies according to different Cantonese syllabaries (Wong, 1941; Hashimoto, 1972; NTTCLRC, 1980; Chow & Yiu, 1987). It is because the vocabulary of the Cantonese dialect is varying from time to time and also from place to place. Many syllables may exist in the colloquial speech but no corresponding written Chinese characters can be found. Also, some of the distinctions among the syllables are not known by most of the Cantonese-speaking people nowadays. So, a set of 1470 different syllables is selected as the full set of Cantonese syllables in this project on the following basis:

- They are known by most of the Cantonese-speaking people nowadays (especially those who live in Hong Kong).

- There are known written Chinese characters corresponding to them.

These 1470 selected Cantonese syllables fully cover all Cantonese phonemes (19 *initials* + 1 null *initial* and 53 *finals* (with 10 *syllabic segments* and 8 *endings* + 1 null *ending*)). If the tone is neglected, there are still 572 distinct syllables in the corpus.

To deal with the case of polyphonics in Cantonese (i.e. multi-pronunciation of written Chinese characters), a simple Chinese word or phrase consisting two to four written Chinese characters is associated with each syllable in full set list to remind the speaker what the desired syllables are. The order of the syllables in the list is the same as that in NTTCLRC's (1980) syllabary. This Cantonese syllable full set list is provided in Appendix A.

4.2 Speech Database Acquisition

All recording work was done on computer directly in a quiet room with echo suppression. The speech signals were first filtered by a lowpass filter with 4 kHz bandwidth, quantized by a 14-bit A/D converter at 10 kHz sampling rate, and stored as files in disk directly. No recording tapes were used.

One male speaker born and educated in Hong Kong was asked to provided all the training and testing database. Each utterance was uttered in isolated syllable. Two trials of recording, each providing a set of 1470 utterances according to the speech corpus, were done. To ensure the speaker to give the correct pronunciation of the syllable, especially for the cases of polyphonics in Cantonese, the speaker was asked to speak out the provided Chinese word or phrase that associated to the syllable before the recording of each isolated syllable.

The 1470 utterances are divided into 7 groups. Each group consists of 210 utterances. Short break time is allowed after recording of each group of utterances. Normalization of tone feature parameters will be done on each group of utterances to avoid pitch range and speaking speech variation due to long time recording or recording at different time. The normalization procedure is described later in section 5.5.

5 Feature Parameters Extraction

The feature extraction process described below can be summarized as in Figure 5.1.

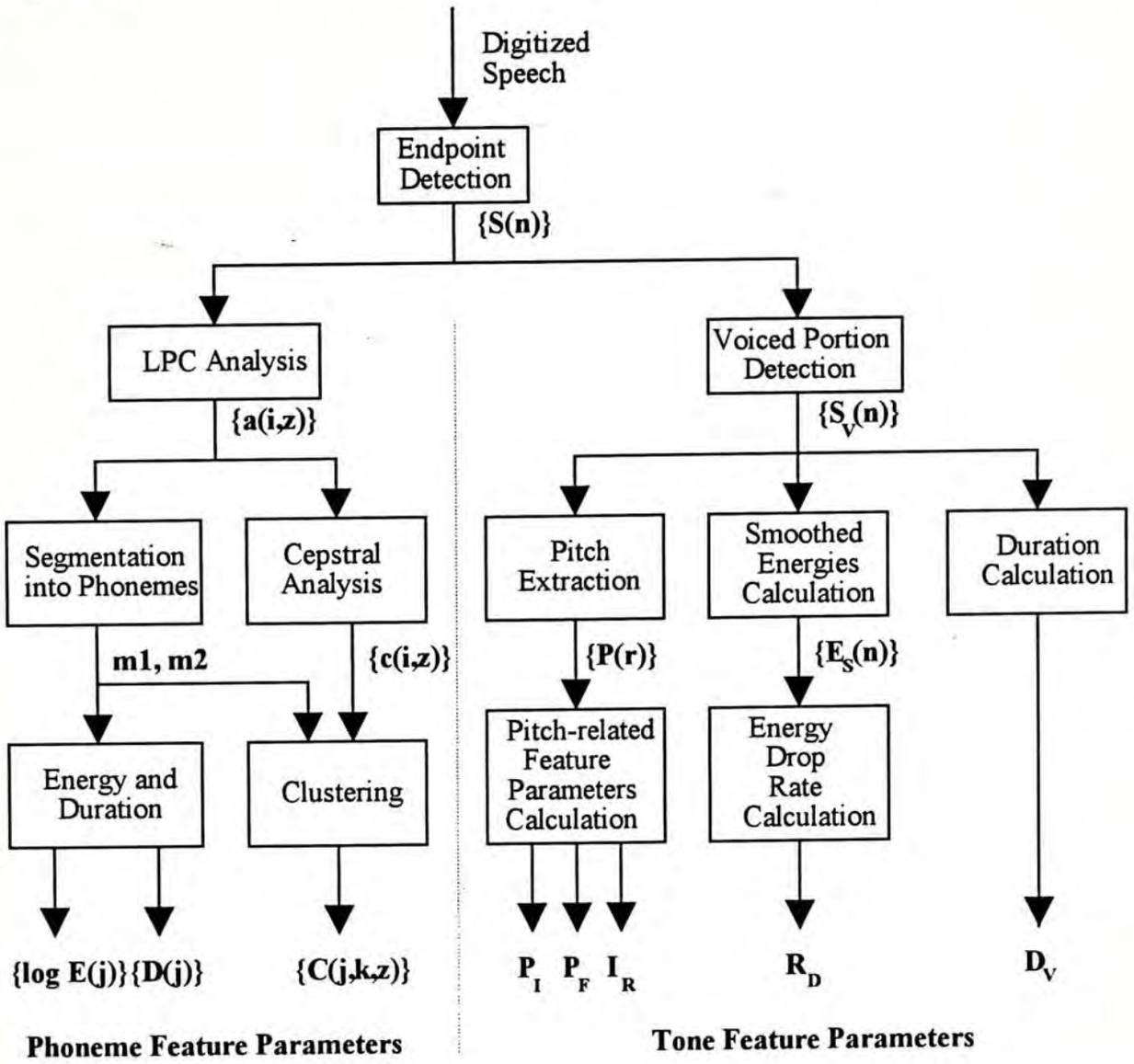

Figure 5.1 Extraction of Feature Parameters

5.1 Endpoint Detection

To locate the endpoints (the start point and the end point) of the syllable from the speech signals, we have adopted the modified version of the voiced portion detection algorithm from Lee *et al* (1993). The difference between our endpoint detection algorithm and Lee *et al*'s voiced portion endpoint detection algorithm is that different start points are desired. Since Lee *et al* worked on tone classification, they only tried to locate the start point of the voiced portion of the syllable and the leading unvoiced portion of the syllable was ignored. However, we have to keep the unvoiced portion of the syllable for syllable recognition and thus try to locate the start point of the syllable including the unvoiced portion.

Our endpoint detection algorithm is described below. The speech signals are first divided into several time frames. The frame length is 10ms and the time shift is 5ms (i.e. 50% overlapping). The frame energy and zero-crossing rate is calculated for each frame. The endpoints of the syllable in the speech signal are mainly determined by the frame energies. However, to deal with the low-energy unvoiced *initials*, the frame zero-crossing rates are also considered. From the middle of the signal, search backward until the frame energy is just less than a threshold and the frame zero-crossing rate is just below another threshold, then the start point of the syllable is located. The end point of the syllable is located similarly by searching forward until the frame energy is just less than the threshold. With the two endpoints, the speech sequence {**S(n)**} without leading and trailing silence or noise is given.

5.2 Speech Processing

The speech data are pre-emphasized using a first-order filter with transform function $1 - 0.95z^{-1}$ for spectral flattening and then applied with a Hamming window. The frame length for analysis is 25.6ms and the frame shift is 10ms. From each frame of speech signals (say i-th frame), a set of 12-order LPC (Linear Prediction Coding) coefficients **{a(i,z)}** is computed using autocorrelation method (Markel & Gray, Jr., 1976). Finally, the corresponding LPC smoothed log-amplitude spectra and the LPC-derived cepstral coefficients **{c(i,z)}** are obtained. The log-amplitude spectra are then used in the speech segmentation, while the cepstral coefficients are used as input parameters for phoneme recognition.

5.3 Speech Segmentation

The aim of speech segmentation is to divide the syllable into its corresponding phonemic segments. In this case, the phonemes can be used as the classification unit rather than using the whole syllable.

The idea of the implicit segmentation method proposed by van Hemert (1991) is adopted in a modified way. It makes use of the normalized correlation coefficient between the LPC smoothed log-amplitude spectra of every two frames in the utterance as a measure of frame similarity. Then the phoneme boundaries are identified as indicated by the low values in correlation measure as shown in Figure 5.2.

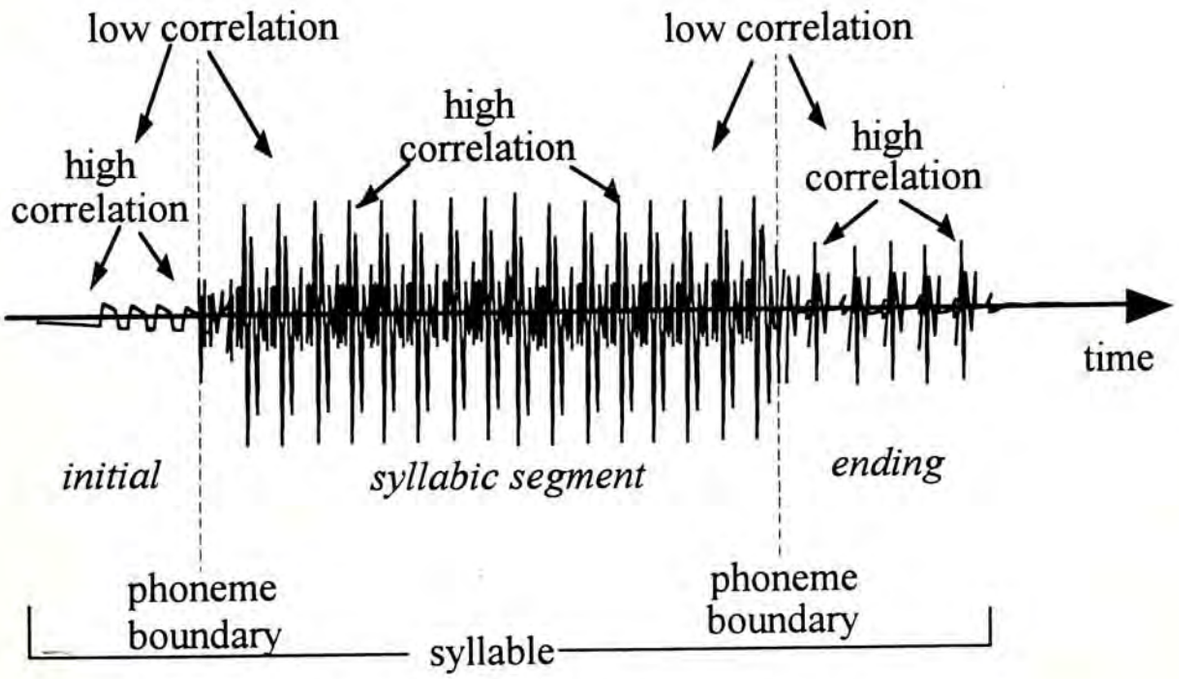

Figure 5.2 Segmentation of Syllable into Phonemes

In this method, the correlation coefficient of i-th frame and j-th frame ($C_{i,j}$) is calculated for every i and j. Then we locate the two frame indexes m (say m1 and m2, with m1 < m2) that the expression $\max(C_{m-2,m}, C_{m-2,m+1}, C_{m-1,m}, C_{m-1,m+1})$ gives

the two smallest values. **m1** and **m2** are taken as the segmentation marks, and then the syllable would be segmented into three parts: (F_0.....F_{m1-1}), (F_{m1}.....F_{m2-1}), (F_{m2}.....F_n). We assume these three parts are corresponding to the *initial*, *syllabic segment* and *ending* respectively, e.g. the three parts segmented from /but/ are /b/, /u/ and /t/ accordingly.

Note that this segmentation algorithm always divides a syllable into three segments, however, as stated in chapter 2, a Chinese syllable may consist 1, 2 or 3 phonemes according to the cases of whether the *initial* or *ending* are missing. Therefore, we have to make some assumptions on the extra parts segmented on such cases. For the case of missing *initial*, we treat the first part as a new class of zero *initial* /0/, as given in the traditional Cantonese phonology. For the case of missing *ending*, we treat the third part just as the duplication of the second part (i.e. *syllabic segment*). For examples, /a/ becomes /0/+/a/+/a/, /uk/ becomes /0/+/u/+/k/ and /si/ becomes /s/+/i/+/i/.

5.4 Phoneme Feature Extraction

Each phonemic segment (including *initial*, *syllabic segment* and *ending*) provided by the segmentation of the syllable is further divided into two equal parts. For each of these two parts (say k-th part of j-th phonemic segment), a reference set of 12-order LPC-derived cepstral coefficients **{C(j,k,1), C(j,k,2), ..., C(j,k,12)}** is selected by calculating the cluster pseudocenter (Wilpon, 1985) from cepstral coefficients **{c(i,z)}** of that part of speech using quefrency weighted cepstral distance measure (Tohkura, 1987).

Also, the log-energy **log E(j)** and duration **D(j)** of the j-th phonemic segment are also calculated. So, there are 26 feature parameters **{C(j,1,1), C(j,1,2), ..., C(j,1,12), C(j,2,1), C(j,2,2), ..., C(j,2,12), log E(j), D(j)}** altogether being used for the j-th phoneme classifier.

5.5 Tone Feature Extraction

The tone feature extraction algorithm and the selection of the set of feature parameters for tone recognition of isolated Cantonese syllables below is adopted from Lee *et al* (1993).

A modified version of the 3-level center clipped pitch extraction algorithm proposed by Sondhi (1968) is employed. First of all, any unvoiced consonants in the beginning of the syllables are not taken into consideration. It is because for an isolated Cantonese syllable, the information concerning the tone is only carried by the voiced portion of the utterance. This is done by skipping the leading frames with frame zero-crossing rates greater than the threshold. The detected voiced portion of the speech signals is then divided into 16 equal frames with 50% overlapping. For each frame, the largest peak of the filtered and normalized correlation sequence of the 3-level center clipped speech signal is taken as the pitch period of the speech segment. Then, a sequence of pitch values **{P(1), P(2), ..., P(r), ..., P(16)}** can be given by calculating the reciprocal of the detected pitch periods of the 16 frames.

The initial pitch $\mathbf{P_I}$, final pitch $\mathbf{P_F}$, pitch rising index $\mathbf{I_R}$, duration of voiced portion $\mathbf{D_V}$ and energy drop rate $\mathbf{R_D}$ are used as the 5 feature parameters in determining the tone of the speech utterance. The first 3 parameters are pitch-related while the last 2 parameters are duration-related. The definitions of these feature parameters are as follows:

a. $$\mathbf{P_I} = \frac{P(3)+P(4)}{2} \tag{5.1}$$

b. $$\mathbf{P_F} = \frac{P(13)+P(14)}{2} \tag{5.2}$$

where **P(r)** is the detected pitch of the r-th frame.

c. $\mathbf{I_R} = k\frac{\text{Max}\{P(r)\} - \text{Min}\{P(r)\}}{\text{Max}\{P(r)\} + \text{Min}\{P(r)\}}, \quad 3 \le r \le 14$

$$\text{where } k = \begin{cases} 1 & \text{for } arg\,\text{Max}\{P(r)\} > arg\,\text{Min}\{P(r)\} \\ -1 & \text{for } arg\,\text{Max}\{P(r)\} \le arg\,\text{Min}\{P(r)\} \end{cases} \tag{5.3}$$

Note that the sign of $\mathbf{I_R}$ indicates whether the pitch level within an utterance rises or drops with time while the magnitude of $\mathbf{I_R}$ measures the degree of such rise or drop in the pitch level.

d. $\mathbf{D_V}$ = duration of detected voiced portion

e. $\mathbf{R_D}$ = reciprocal of the time for energy declining from 90% to 10% in smoothed frame energy at the rear part of the speech signal

Though the experiments are performed in a speaker-dependent setting, normalization of the pitch-related and duration-related parameters is still needed to deal with the temporal fluctuation of pitch-related and duration-related parameters. This is because the pitch level and speed of speaking of the same speaker may vary to some extent from time to time. Both the pitch level and the speed of speaking can be affected by emotional, stylistic and environmental factors.

Normalization is done on each group of 210 utterances mentioned in section 4.2. The normalization procedure of the pitch-related parameters is done by simply dividing the parameters by a normalization factor $\mathbf{P_S}$. In our case, $\mathbf{P_S}$ is defined as the mean of the initial pitch values of a subset of the 210 utterances. 24 utterances of tones 2, 4, 5, 6 (6 utterances for each tone) are pre-selected to form the subset. Note that only $\mathbf{P_I}$ and $\mathbf{P_F}$ are needed to be normalized. Although $\mathbf{I_R}$ is also a pitch-related

parameter, it has no need to be normalized. It is because $\mathbf{I_R}$ has a self-normalizing property.

The normalization procedure of the duration-related parameters is similar to that of the pitch-related parameters, i.e. dividing the parameters by normalization factors. The normalization factors for $\mathbf{D_V}$ and $\mathbf{R_D}$ are $\mathbf{D_{SV}}$ and $\mathbf{R_{SD}}$ respectively. $\mathbf{D_{SV}}$ and $\mathbf{R_{SD}}$ are defined as the mean of $\mathbf{D_V}$ and $\mathbf{R_D}$ values of the same pre-selected 24 utterances as for $\mathbf{P_S}$.

6 The Design of the System

In general, a speech recognition system can be divided into two basic functional units as shown in Figure 6.1. The first unit is the feature extraction unit and the second one is the classification unit. The feature extraction unit extracts a set of feature parameters from the speech signal, and the classification unit classifies the speech signal by the feature parameters extracted.

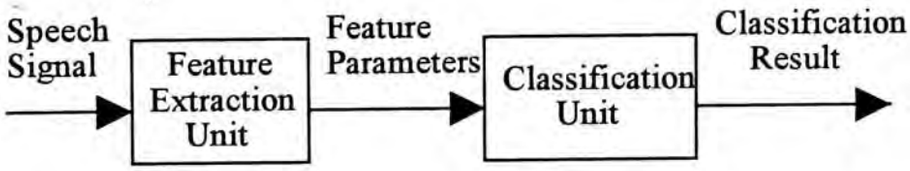

Figure 6.1 Diagram of a Speech Recognition System

The design of the feature extraction unit depends on what kinds of feature parameters are used in the classification unit. We shall concentrate on discussion of the design of the classification unit in this chapter. The design of the feature extraction unit has already been described in chapter 5.

In speech recognition, the category for classification can be various, such as a word, a syllable or a phoneme. Since Cantonese is a *monosyllabic tonal* language, *syllable with tone* (will be simply taken as *syllable* in the following of this thesis) is adopted as the basic unit for Cantonese speech recognition in our case.

6.1 Towards Large Vocabulary System

According to Wong (1941), there are about 1800 different syllables in the contemporary Cantonese dialect. In order to facilitate effective recognition of such a large vocabulary, simply treating the syllable as a single recognition unit is uneconomical. So, techniques of problem decomposition must be introduced.

In dealing with a syllable of a *monosyllabic tonal* language, one can simply divide the problem into sub-problems such as classifications of consonants, vowels and tones independently. However, Wang *et al* (1991) proposed a hierarchical neural network model for *large vocabulary* isolated Mandarin speech recognition, as Mandarin is also a *monosyllabic tonal* language. According to their viewpoint, there are inter-dependencies between consonants, vowels and tones which may not be ignored.

There are 4 tones, 37 vowels (*finals*) and 21 consonants (*initials*) for the constitution of 1300 Mandarin syllables. They first used a single tone network for tone recognition, and then four vowel networks, each for different tones, for vowel recognition. At last, 4x37 consonant networks, each for different tones and vowels, were used for consonant recognition. As a result, they used total 153 (= 1 + 4 + 4 x 37) neural networks to solve the problem. The hierarchical neural network model is shown in Figure 6.2.

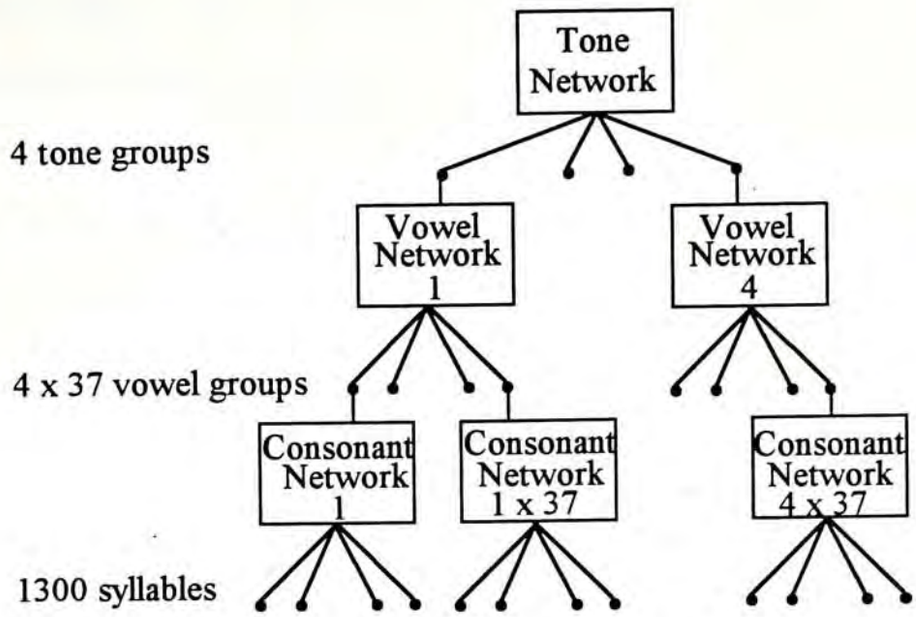

Figure 6.2 Wang *et al*'s Model for Mandarin Speech Recognition

From the viewpoint of Cantonese phonology given in the previous chapter, there exist basic units including 9 tones, 53 vowels (36 vowels for the 6 non-entering tones and 17 vowels for the 3 entering tones), and 19 consonants. If we adopt the same hierarchical neural network model as that of Wang *et al* to the Cantonese speech recognition, we have to use a total of 277 (= 1 + 9 + 6 x 36 + 3 x 17) neural networks for the problem, which is nearly twice as much as for Mandarin. Since this method takes too much efforts in training a large number of artificial neural networks, we would use another model in our recognition task.

6.2 Overview of the Isolated Cantonese Syllable Recognition System

We suggest to perform phoneme recognition as phonemes are given by the segmentation of syllable. Since Cantonese is a *tonal* language, tone classification is also needed as well as phoneme recognition. So, a model of hierarchical neural networks as shown in Figure 6.3 is adopted. The phonemes in different positions of the syllable (*initial*, *syllabic segment*, *ending*) and tone of each syllable are classified separately by the classifiers in the primary level. These recognition results are passed into the syllable classifier in the secondary level. Note that there is an *ending* corrector in the intermediate level between the *ending* classifier and the syllable classifier. The arrow from the tone classifier to the *ending* corrector denotes that the results of the *ending* classifier can be improved with the help of the results of the tone classifier. The method of *ending* correction by tone will be described in section 6.4.

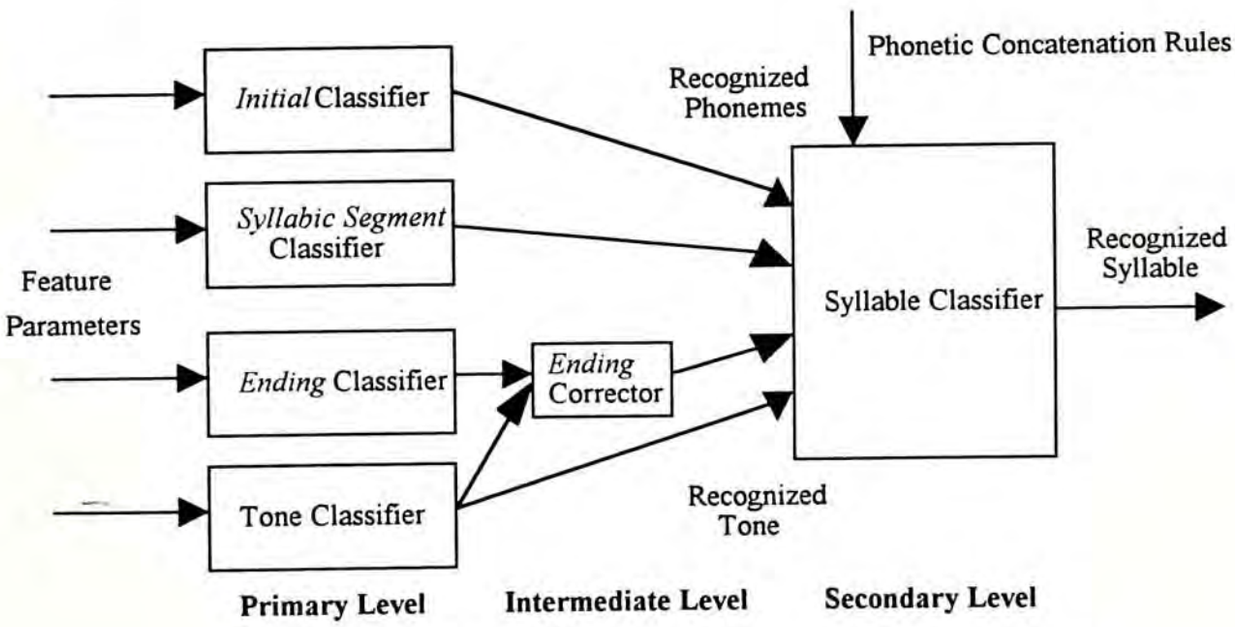

Figure 6.3 The Model of Hierarchical Neural Networks

The recognized syllable can be given by direct concatenation of the recognized phonemes and tone. However, invalid Cantonese syllables may be formed. The syllable classifier in the secondary level integrates the recognized phonemes and tone together to give the recognized syllable with the help of phonetic concatenation rules of Cantonese dialect. This syllable classifier can eliminate invalid Cantonese syllables by changing the wrong recognition results from one or more classifiers in the primary level.

Compared with the hierarchical model described in the last section, our hierarchical model is simpler than Wang *et al*'s. First, Wang *et al*'s model needs a lot of neural networks, a lot of training data and hence a lot of training time. Second, their model is much more precise as the classification of syllabic structures for different tones are done by different neural networks. However, once the syllable is classified wrongly in the first step, no correction can be done later. Our model seems to be more flexible by providing possibilities for correction of wrong classification in the primary level of the hierarchical model.

6.3 The Primary Level: Phoneme Classifiers and Tone Classifier

The well-known MLP (Multi-Layer Perceptron) neural network (Lippmann, 1987) is employed in each classifier of the primary level of the system. Every MLP neural network has one input layer and one output layer. The number of hidden layers in MLP neural network can be various. We use MLP neural network with only one hidden layer (i.e. two-layer perceptron), as shown in Figure 6.4, in our system.

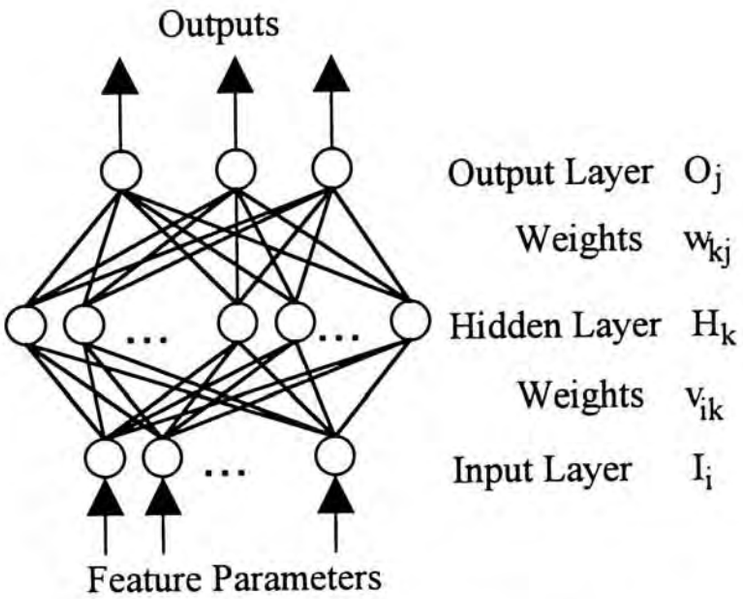

Figure 6.4 Two-Layer Perceptron

Different feature parameters of the syllable are fed into the 4 MLP neural networks correspondingly. These feature parameters are extracted according to the methods described in chapter 5. For every phoneme classifiers, the number of input parameters is 26. For the tone classifier, 5 input parameters are used. The number of nodes in the input layer is the same as the number of the input parameters. However, the number of nodes in the hidden layer is arbitrary assigned. The number of nodes in

the output layer the same as the number of classification goals of the MLP neural network.

Since we have adopted a simple speech segmentation algorithm, all the syllables are segmented into 3 speech segments. In the cases where the syllables have missing *initial* or *ending*, there are still speech segments assumed as the *initial* or *ending* of these syllables. Two new classification goals are added to the phoneme classifiers in order to deal with these *extra* speech segments. They are /0/ in *initial* classifier and /a/ɛ/ɔ/oe/ in *ending* classifier. There is no need to further divide the classification goal /a/ɛ/ɔ/oe/ in *ending* classifier into sub-goals, because it is just the duplication of the *syllable segment* and these phonemes have already been classified in *syllabic segment* classifier.

Moreover, each of the following pairs of phonemes: /l/ & /n/, /0/ & /ŋ/ in *initial* and /m̩/ & /ŋ̩/ in *syllabic segment* is merged into a single classification goal. It is because most people cannot distinguish between each pair of them (Ho, 1989). By treating each of the above pairs as the same phoneme, the number of different syllables in the full set becomes 1437, which is still 97.8% of the original size (1470). The classification goals of the classifiers are summarized in Table 6.1.

Classifiers	Number of Goals	Classification Goals
Initials	18	p, t, k, kw, b, d, g, gw, ts, dz, s, f, h, l/n, m, 0/ŋ, j, w
Syllabic Segment	9	a, ɐ, ɛ/e, i, ɔ/o, oe, u, y, m̩/ŋ̩
Ending	9	a/ɛ/ɔ/oe, i/y, u, m, n, ŋ, p, t, k
Tone	9	tones 1, 2, 3, 4, 5, 6, 7, 8, 9

Table 6.1 Classification Goals of the 4 Classifiers

The configurations of the 4 MLP neural networks are summarized in Table 6.2.

MLP Neural Networks	Number of Input Units	Number of Hidden Units	Number of Output Units
Initial Classifier	26	45	18
Syllabic Segment Classifier	26	35	9
Ending Classifier	26	35	9
Tone Classifier	5	25	9

Table 6.2 The Configurations of the Neural Networks

All the MLP neural networks are trained independently by the popular backpropagation training algorithm (Rumelhart *et al*, 1986). Backpropagation is gradient descent of mean-squared error as a function of the weights in the neural network. The training algorithm is an iterative algorithm which includes two passes: forward pass and backward pass.

During the forward pass, a sample is supplied to the input of the network. The outputs of all the neurons at each layer are computed. The computation starts at the input layer and forward to the output layer. The outputs of the neurons are given by the equations:

$$H_k = f(\sum_i v_{ji} I_i) \tag{6.1}$$

$$O_j = f(\sum_k w_{kj} H_k) \tag{6.2}$$

where I_i, H_k and O_j are the outputs of the neurons of input, hidden and output layer respectively. v_{ik} and w_{kj} are weights connecting inputs to hidden units and hidden to output units respectively. f(x) is the sigmoid function, where

$$f(x) = \frac{1}{1+e^{-x}} \tag{6.3}$$

The output of the network is then compared to the desired output and its error is calculated. The error function is defined as:

$$E = \frac{1}{2}\sum_{j}(T_j - O_j) \tag{6.4}$$

where T_j is the desired value of output O_j.

During the backward pass, the derivative δ of this error is then propagated back through the network (i.e. from the output layer back to the input layer). The derivatives of error δ_j for the output layer and δ_k^* for the hidden layer are given as:

$$\delta_j = O_j(1-O_j)(T_j - O_j) \tag{6.5}$$

$$\delta_k^* = H_k(1-H_k)\sum_{k}\delta_j w_{kj} \tag{6.6}$$

As to decrease the error, all the weights are updated according to the equations:

$$\Delta w_{kj}^{(new)} = \eta\delta_j H_k + \alpha\,\Delta w_{kj}^{(old)} \tag{6.7}$$

$$\Delta v_{ik}^{(new)} = \eta\delta_k^* I_i + \alpha\,\Delta v_{ik}^{(old)} \tag{6.8}$$

where Δw (or Δv) is the weight change, η is the learning rate and α is the momentum for adaptive learning.

These two passes are repeated many times for all training samples until the error is reduced below a threshold (i.e. the network converges to produce the desired output).

6.4 The Intermediate Level: *Ending* Corrector

Since there are inter-dependencies among the constituents of a syllable, we can first make use of the result from the tone classifier to correct the result from the *ending* classifier. This is the major function of the *ending* corrector in the intermediate level. This correction is based on the rule, which is stated in section 2.2, that syllables are pronounced in entering tones if and only if they have plosive stop *endings*. If the utterance is recognized to possess an entering tone, we reset the outputs of the nodes representing *endings* other than plosive stops to the desired output value of "0". This is because by so doing we can ensure that the recognized *ending* must be a plosive stop. Otherwise, we reset the outputs of the nodes representing plosive stop *endings* to the desired output value of "0", so that none of these nodes can become the winner node (i.e. no plosive stop *endings* for non-entering tones). Since the tone classifier provides a more reliable outcome than the *ending* classifier in recognition, it is therefore appropriate to use its result to rectify the outputs of the *ending* classifier.

6.5 The Secondary Level: Syllable Classifier

The recognition of syllables makes use of the classification results of 4 classifiers in the primary level. Without any syllable classifier, the simplest method to give the recognized syllable is to concatenate (i.e. concatenation without correction) the results of the phoneme classifiers and the tone classifier. A rightly recognized syllable can be given only if all the 4 results of the classifiers in the primary level are correct. If all the 4 classifiers in the primary level get a recognition rate of 80%, then the average syllable recognition rate can only be 40.96% (= $(80\%)^4$) and the worst-case syllable recognition rate is 20% (= 1 - 4 x (1 - 80%)). We can notice that without any further enhancement by a syllable classifier, it is very difficult to achieve a good syllable recognition. So, a syllable classifier is necessary for improving the recognition rate of the syllable and hence pays an important role in the whole system.

For the syllable classifier of the secondary level, two different approaches have been used. They are concatenation with correction and Fuzzy ART (Adaptive Resonance Theory). These approaches will be described in details below.

6.5.1 Concatenation with Correction Approach

For each of the MLP neural networks in the primary level, we select the output node with the greatest activation as the winning node. This is called the "winner-take-all" strategy. The classification goal represented by the winning node is taken as the classification result.

Next, we concatenate the classification results from the 4 MLP neural networks together to form a syllable. If a valid syllable (one of the 1437 syllables in the full set) is formed, we take it as the recognized syllable. Otherwise, we select the most possible one among the 1437 syllables as the recognized syllable. The method of selecting the most possible syllable is as follows:

1. For every MLP neural network, we normalize its outputs by the following steps, such that their sum is equal to unity:
   1.1 Add a small offset to each of the outputs.
   1.2 Divide all the outputs by the sum of all output values.
2. For every valid syllable in the full set, multiply the 4 output values of the output nodes (1 from each MLP neural network) representing the phonemes and tone of the syllable together.
3. Select the valid syllable that has the highest product in step 2.

Since we have considered the valid syllables in this approach, all the inappropriate syllables formed by simple concatenation will not be accepted as the recognized syllables. However, it has no improvement on those syllables which are wrongly classified as other valid syllables.

6.5.2 Fuzzy ART Approach

Fuzzy ART (Carpenter *et al*, 1991) is one of the Adaptive Resonance Theory (ART) model families. It incorporates computations from fuzzy set theory into the ART 1 neural network, which learns to categorize only binary input patterns. So, Fuzzy ART is capable to achieve rapid stable learning of recognition categories in response to arbitrary sequences of analog or binary input patterns. Its architecture is much more simple than that of ART 2, which is another ART model family that learns to categorize either analog or binary input patterns. The outputs of the MLP neural networks can be directly used as the inputs of the Fuzzy ART, because the elements of outputs of MLP neural networks are also in the interval [0, 1] as required by Fuzzy ART.

Fuzzy ART with complement coding is used as the syllable classifier. The outputs from the 4 MLP neural networks in the primary level are grouped together to form a vector **a** of size M. This vector is normalized by complement coding to form the input vector **I** of size 2M. The complement coding process is illustrated in Figure 6.5. The aim of this input normalization procedure is to deal with the category proliferation problem in analog ART systems.

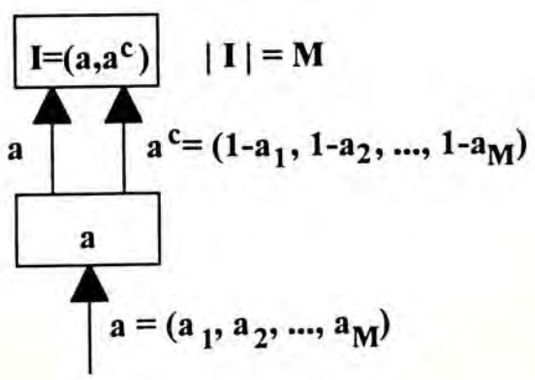

Figure 6.5 Complement Coding for Normalization of Input vectors

The architecture of the syllable classifier in Fuzzy ART is shown in Figure 6.6. There are one input layer and one output layer. Adaptive weights are fully connected between the two layers. The number of input nodes is twice the number of the input parameters if complement coding is used (i.e. 2M). Every output node corresponds to a category to be classified. So, the number of output nodes is equal to the maximum number of categories that can be discriminated. In our syllable classifier, the number of input nodes is 90 (= 2 x (18 + 9 + 9 + 9)) and the number of output nodes is 1437 (different syllables in the full set). There is also a vigilance test to check the similarity of the input vector and weight vector.

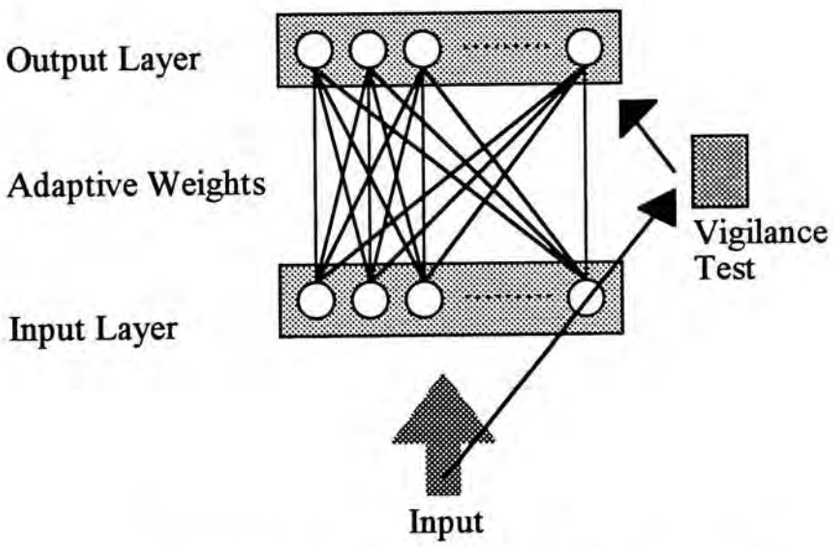

Figure 6.6 The Fuzzy ART Architecture

Initially, all adaptive weights are set to 1 and all categories are said to be *uncommitted.* After a category is selected for weights updating, it becomes *committed.* All the adaptive weights are monotone non-increasing and would converge. There are other parameters: choice parameter $\alpha > 0$; learning rate $\beta \in [0, 1]$; vigilance parameter $\rho \in [0, 1]$.

The output values of output node j for input vector $\mathbf{I}$ is defined by

$$O_j = \frac{|\mathbf{I} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|} \tag{6.9}$$

where $\mathbf{w}_j$ is the weight vector associated with output node j. The fuzzy AND operator $\wedge$ is defined by

$$(\mathbf{x} \wedge \mathbf{y})_i \equiv \min(x_i, y_i) \tag{6.10}$$

and the norm $|\cdot|$ is defined by

$$|\mathbf{x}| \equiv \sum_{i=1}^{2M} |x_i| \tag{6.11}$$

The output node J with the greatest output is selected as the category choice. A vigilance test,

$$\frac{|\mathbf{I} \wedge \mathbf{w}_J|}{|\mathbf{I}|} \geq \rho \tag{6.12}$$

will take place. If this category choice cannot pass the vigilance test, the output O_j is reset to -1. A new category choice will be chosen and this process continues until the category choice passes the vigilance test. Once the category choice passes the vigilance test, the weight vector $\mathbf{w}_J$ is updated according to the equation

$$\mathbf{w}_J^{(new)} = \beta(\mathbf{I} \wedge \mathbf{w}_J^{(old)}) + (1-\beta)\mathbf{w}_J^{(old)} \tag{6.13}$$

It is useful to combine fast initial learning with a slower rate of forgetting. So, the fast-commit slow-recode option is adopted, i.e. β is set to 1 when J is an uncommitted node and β is restored after the category is committed. This option retains the benefit of fast learning (i.e. quick adaptation to inputs that may occur only rarely) and prevents features that have already learned from being erroneously forgotten in response to noisy inputs.

So, in the training process, the desired outputs of the MLP neural networks for the 1437 syllables and the actual outputs of the MLP neural networks for the

training data are learned into the weights of the 1437 categories with the fast-commit slow-recode setting directly. In the recognition process, the actual outputs of the 4 MLP neural networks for the testing data are passed into the syllable classifier directly. No learning (i.e. weights updating) will take place at this time. The category choice which can pass the vigilance test is taken as the recognized syllable.

7 Computer Simulation

Computer experiments to validate the potential of the proposed isolated Cantonese syllable recognition system were done in the speaker-dependent setting. All the 1470 isolated Cantonese syllables of the full set were used.

7.1 Experimental Conditions

Since there were two trials in the Cantonese speech database, we use one trial as the training data and the another trial as the testing data.

The parameters used in the backpropagation training processes for the MLP neural networks in the primary level are as follows: learning rate $\eta = 0.1$; momentum $\alpha = 0.3$; desired output values T_j : "0" = 0.1; "1" = 0.9.

The parameter values in the Fuzzy ART neural network in the secondary level are as follows: choice parameter $\alpha = 0.05$; learning rate $\beta = 0.4$ for uncommitted node and 1 for committed node; vigilance parameter $\rho = 0.8$.

7.2 Experimental Results of the Primary Level Classifiers

Tables 7.1 to 7.8 are the confusion matrices for the 4 MLP classifiers in the primary level on the training and testing data.

Initial	p	t	k	kw	b	d	g	gw	ts	dz	s	f	h	l/n	m	0/ŋ	j	w	Total	Accuracy
p	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	100.0%
t	0	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	74	100.0%
k	0	1	49	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	51	96.1%
kw	0	0	1	12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	14	85.7%
b	1	0	0	0	74	0	0	0	0	0	0	0	1	0	0	0	0	0	76	97.4%
d	0	1	1	0	0	90	0	0	0	0	0	0	0	0	0	1	0	0	93	96.8%
g	0	0	1	0	0	0	102	0	0	0	0	0	0	0	0	0	0	0	103	99.0%
gw	0	0	2	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	28	92.9%
ts	0	0	0	0	0	0	0	0	114	0	0	0	0	0	0	0	0	0	114	100.0%
dz	0	0	1	0	0	0	0	0	0	121	0	0	1	0	0	0	0	0	123	98.4%
s	0	0	0	0	0	0	0	0	0	0	133	0	0	0	0	0	0	0	133	100.0%
f	0	0	0	0	0	0	0	0	0	0	0	53	0	0	0	0	0	0	53	100.0%
h	0	0	0	0	0	0	0	0	0	0	0	0	111	0	0	0	0	0	111	100.0%
l/n	0	0	0	0	0	0	0	0	0	0	0	0	0	135	0	0	0	0	135	100.0%
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	1	0	79	98.7%
0/ŋ	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	69	1	0	71	97.2%
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	0	89	100.0%
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	60	100.0%

Table 7.1 Confusion Matrix of *Initial* Classifier on Training Data

Initial	p	t	k	kw	b	d	g	gw	ts	dz	s	f	h	l/n	m	0/ŋ	j	w	Total	Accuracy
p	26	4	7	1	3	0	5	0	0	3	0	3	9	0	1	0	0	1	63	41.3%
t	2	38	5	0	0	0	2	0	9	1	3	1	5	0	1	1	6	0	74	51.4%
k	3	5	18	1	2	0	7	1	3	2	1	3	3	0	0	0	2	0	51	35.3%
kw	0	1	0	4	1	0	1	2	1	1	0	1	1	0	0	0	0	1	14	28.6%
b	4	3	3	0	51	1	3	0	1	2	0	1	2	2	1	0	2	0	76	67.1%
d	0	3	0	0	6	64	3	1	1	2	0	3	1	0	3	2	2	2	93	68.8%
g	2	0	13	1	6	2	57	1	2	3	2	1	3	2	2	1	3	2	103	55.3%
gw	2	1	0	1	1	0	2	17	0	0	0	0	0	0	0	0	1	3	28	60.7%
ts	2	7	2	0	1	0	3	1	51	25	17	2	2	0	1	0	0	0	114	44.7%
dz	1	3	0	0	1	1	1	0	33	56	18	1	2	1	1	1	2	1	123	45.5%
s	0	0	0	0	1	0	0	0	11	13	98	0	6	3	1	0	0	0	133	73.7%
f	0	0	1	0	0	0	1	0	2	1	4	40	2	1	0	0	1	0	53	75.5%
h	3	3	5	0	0	1	1	0	7	2	1	3	78	3	0	0	2	2	111	70.3%
l/n	1	3	2	0	1	0	1	0	4	2	2	3	4	87	7	11	6	1	135	64.4%
m	8	1	0	0	0	0	0	0	1	1	2	0	1	10	51	3	1	0	79	64.6%
0/ŋ	2	0	0	0	0	0	0	0	1	0	0	0	1	8	33	23	3	0	71	32.4%
j	0	2	1	0	1	0	0	0	8	5	1	0	1	2	3	0	65	0	89	73.0%
w	3	1	0	1	0	1	0	4	1	2	1	0	4	2	2	0	1	37	60	61.7%

Table 7.2 Confusion Matrix of *Initial* Classifier on Testing Data

S.S.	a	ɐ	ɛ/e	i	ɔ/o	oe	u	y	m/ŋ	Total	Accuracy
a	279	10	0	0	0	1	1	0	0	291	95.9%
ɐ	0	281	0	0	2	1	0	0	0	284	98.9%
ɛ/e	0	0	101	1	0	0	0	0	0	102	99.0%
i	0	0	0	258	0	0	0	0	0	258	100.0%
ɔ/o	0	0	0	0	221	0	1	0	0	222	99.5%
oe	0	0	0	0	0	101	3	0	0	104	97.1%
u	0	0	0	0	0	0	137	0	0	137	100.0%
y	0	0	0	0	0	0	0	68	0	68	100.0%
m/ŋ	0	0	0	0	0	0	0	0	4	4	100.0%

Table 7.3 Confusion Matrix of *Syllabic Segment* Classifier on Training Data

S.S.	a	ɐ	ɛ/e	i	ɔ/o	oe	u	y	m/ŋ	Total	Accuracy
a	250	15	1	3	10	2	10	0	0	291	85.9%
ɐ	3	192	11	21	8	12	35	1	1	284	67.6%
ɛ/e	0	7	85	7	0	1	2	0	0	102	83.3%
i	1	15	2	224	3	2	6	5	0	258	86.8%
ɔ/o	10	21	1	2	156	5	27	0	0	222	70.3%
oe	2	16	2	10	0	61	11	2	0	104	58.7%
u	2	7	0	14	9	6	98	1	0	137	71.5%
y	0	0	0	7	0	0	2	59	0	68	86.8%
m/ŋ	0	0	1	3	0	0	0	0	0	4	0.0%

Table 7.4 Confusion Matrix of *Syllabic Segment* Classifier on Testing Data

Ending	a/ɛ/ɔ/oe	i/y	u	m	n	ŋ	p	t	k	Total	Accuracy
a/ɛ/ɔ/oe	114	0	0	1	0	0	0	0	17	132	86.4%
i/y	0	259	0	0	3	0	0	1	0	263	98.5%
u	0	1	190	0	7	3	0	2	7	210	90.5%
m	0	0	0	77	19	5	0	1	0	102	75.5%
n	0	0	0	0	240	10	0	0	0	250	96.0%
ŋ	0	0	1	1	14	251	0	4	0	271	92.6%
p	0	0	0	0	0	0	26	0	13	39	66.7%
t	0	0	0	0	0	0	0	85	1	86	98.8%
k	0	0	0	0	0	0	0	0	117	117	100.0%

Table 7.5 Confusion Matrix of *Ending* Classifier on Training Data

Ending	a/ɛ/ɔ/oe	i/y	u	m	n	ŋ	p	t	k	Total	Accuracy
a/ɛ/ɔ/oe	83	1	6	0	0	4	1	2	35	132	62.9%
i/y	0	189	4	4	22	6	0	32	6	263	71.9%
u	0	0	153	1	23	4	0	4	25	210	72.9%
m	3	0	5	25	29	26	0	11	3	102	24.5%
n	0	0	5	0	211	31	1	2	0	250	84.4%
ŋ	0	3	25	5	48	158	0	21	11	271	58.3%
p	12	0	2	0	1	0	11	4	9	39	28.2%
t	0	6	2	0	6	0	0	64	8	86	74.4%
k	7	3	2	1	0	1	2	6	95	117	81.2%

Table 7.6 Confusion Matrix of *Ending* Classifier on Testing Data

Tone	1	2	3	4	5	6	7	8	9	Total	Accuracy
1	271	0	5	0	0	0	0	0	0	276	98.2%
2	0	217	0	0	4	0	0	0	0	221	98.2%
3	3	0	205	0	0	9	0	1	0	218	94.0%
4	0	0	0	209	0	0	0	0	1	210	99.5%
5	0	2	1	0	106	0	0	0	0	109	97.2%
6	0	0	6	0	0	186	0	0	2	194	95.9%
7	1	0	0	0	0	0	66	0	0	67	98.5%
8	0	0	2	0	0	0	0	80	4	86	93.0%
9	0	0	0	0	0	0	0	2	87	89	97.8%

Table 7.7 Confusion Matrix of Tone Classifier on Training Data

Tone	1	2	3	4	5	6	7	8	9	Total	Accuracy
1	270	0	5	0	0	0	1	0	0	276	97.8%
2	0	218	0	0	3	0	0	0	0	221	98.6%
3	1	4	193	0	0	20	0	0	0	218	88.5%
4	0	8	1	189	9	1	0	1	1	210	90.0%
5	1	2	0	1	103	2	0	0	0	109	94.5%
6	0	4	6	5	9	170	0	0	0	194	87.6%
7	2	0	0	0	0	0	58	5	2	67	86.6%
8	0	0	4	0	1	0	0	64	17	86	74.4%
9	0	0	0	0	0	4	0	9	76	89	85.4%

Table 7.8 Confusion Matrix of Tone Classifier on Testing Data

The overall recognition rates for the *initial* classifier, *syllabic segment* classifier, *ending* classifier and tone classifier are 58.6%, 76.3%, 67.3% and 91.2% respectively. Note that the descending order of recognition performance of the 4 primary level classifiers are: tone classifier, *syllabic segment* classifier, *ending* classifier, *initial* classifier. This order follows the ascending order of complication of the classification goals: pitches, vowels, consonants.

For the *initial* classifier, the major confusion occur between /k/ & /g/, and /ts/ & /dz/. They are pairs of similar consonants whereas the only difference between each pair is that one is aspirated and the other is non-aspirated. Also, major confusion occur between /ts/ & /s/, /dz/ & /s/ and /m/ & /0/ŋ/, as each pair of them belongs to the same sub-class of *initials*. The consonants with the lowest recognition rates are /kw/, /0/ŋ/ & /k/. The limited number of training candidates of /kw/ may degrade the capability to recognize /kw/.

For the *syllabic segment* classifier, the incapability in recognizing /m̩/ŋ̩/ may be due to the limited number of training candidates of them. It is observed that most confusion occur between /ɐ/ & /i/, /ɐ/ & /u/ and /ɔ/o/ & /u/. The vowel with the lowest recognition rate is /oe/.

For the *ending* classifier, the lowest recognition rates occur in /m/ and /p/. It can be seen that distinction among the nasals (/m/, /n/ and /ŋ/) is by no mean trivial. In deed, confusion also occur among the training data. Furthermore, many glides are wrongly classified as nasal stops. It may be due to the wrong speech segmentation, as the nasal stops are quite short in duration. However, most of the cases can be corrected in the syllable classifier, as the nasal stops only occur in entering tones.

For the tone classifier, it is noted that the major confusion occur between tone 3 & tone 6, tone 8 & tone 9. For tones 3 and 6, both of them are even tones (i.e. no rising or dropping in pitches) and the only difference is the pitch level. While tones 8 and 9 are both entering tones that most people get confused with them.

7.3 Overall Performance of the System

The experimental results of the whole system are summarized in Table 7.9.

Classifiers		*Ending* Correction	Recognition Rates	
			Training Data	Testing Data
Initial		N/A	98.8%	58.6%
Syllabic Segment		N/A	98.6%	76.5%
Ending		No	92.4%	67.3%
		Yes	93.9%	74.8%
Tone		N/A	97.1%	91.2%
Syllable	Concatenation without Correction	No	87.4%	29.5%
		Yes	89.4%	33.7%
	Concatenation with Correction	No	93.7%	44.5%
		Yes	93.4%	44.7%
	Fuzzy ART	No	98.4%	45.4%
		Yes	98.1%	45.3%

N/A - not applicable

Table 7.9 Summary of the Experimental Results

The performances of all individual classifiers are evaluated under both conditions with and without the incorporation of the *ending* correction. The performance of simple concatenation (i.e. concatenation without correction) is also evaluated for comparison.

With the help of the results from the tone classifier, the recognition rate on *ending* classification can achieve 74.8%, which is increased by 7.5% from the original

result. For simple concatenation without correction, the recognition rates on syllable are 29.5% without *ending* correction and 33.7% with *ending* correction. Note that these syllable recognition rates are greater than the average recognition rates calculated by multiplying the phonemes and tone recognition rates by 2.0% to 3.1%. The average recognition rates for cases without and with *ending* correction are 58.6% x 76.5% x 67.3% x 91.2% = 27.5% and 58.6% x 76.5% x 74.8% x 91.2% = 30.6% respectively. These results show that low recognition rates are not due to the concatenation processes, but due to the original low recognition rates of the phonemes recognition.

With the correction of the invalid syllables, the recognition rate on syllables with and without *ending* correction are increased by 15.0% and 11.0% respectively, i.e. for the approach of concatenation with correction, the recognition rate of the syllable classifier can become 44.5% to 44.7%. For the approach of Fuzzy ART, the recognition rate of the syllable classifier is 45.3% to 45.4%, which is similar to the previous approach. This similarity may be due to the fact that both approaches of the syllable classifier have made use of the same amount of information from the classifiers in the primary level.

In addition, both approaches of syllable classifier can achieve similar recognition rates without *ending* correction as the cases with *ending* correction. This may be because both approaches of syllable classifier can correct *endings* themselves and the effort paid by the *ending* corrector is duplicated and wasted.

Though the recognition rate on syllable is less than one-half, it is valuable to point out that top-**n** hit rates increase significantly as **n** increases. Top-**n** hit rate is defined as the rate that the output node corresponding to the target goal has one of

the **n** highest activations over all output nodes (1437 in our case). The top-**n** hit rates of the syllable classifier in both approaches (concatenation with correction and Fuzzy ART) with and without *ending* correction are listed in Table 7.10 for **n** = 1, 2, 3, 4, 5, 10, 15, 20, 30, 40. Also, a graph is plotted as top-**n** hit rates against **n** of syllable classifier in both approaches without *ending* correction for **n** = 1 up to 100 in Figure 7.1. The top-**n** hit rates for the cases with *ending* correction are not plotted in the graph, as to make the appearance of the graph clear enough.

n	Top-**n** Hit Rates			
	Concatenation with Correction		Fuzzy ART	
	without *Ending* Correction	with *Ending* Correction	without *Ending* Correction	with *Ending* Correction
1	44.5%	44.7%	45.4%	45.3%
2	55.6%	55.9%	55.4%	55.6%
3	61.2%	61.5%	59.7%	60.0%
4	64.4%	64.4%	63.9%	63.8%
5	66.8%	66.9%	66.7%	66.6%
10	73.5%	73.6%	76.5%	76.4%
15	78.5%	78.6%	82.4%	81.9%
20	82.0%	82.0%	85.0%	85.0%
30	86.1%	86.4%	87.9%	88.0%
40	88.2%	88.3%	90.3%	90.1%

Table 7.10 Top-n Hit Rates of the Syllable Classifiers

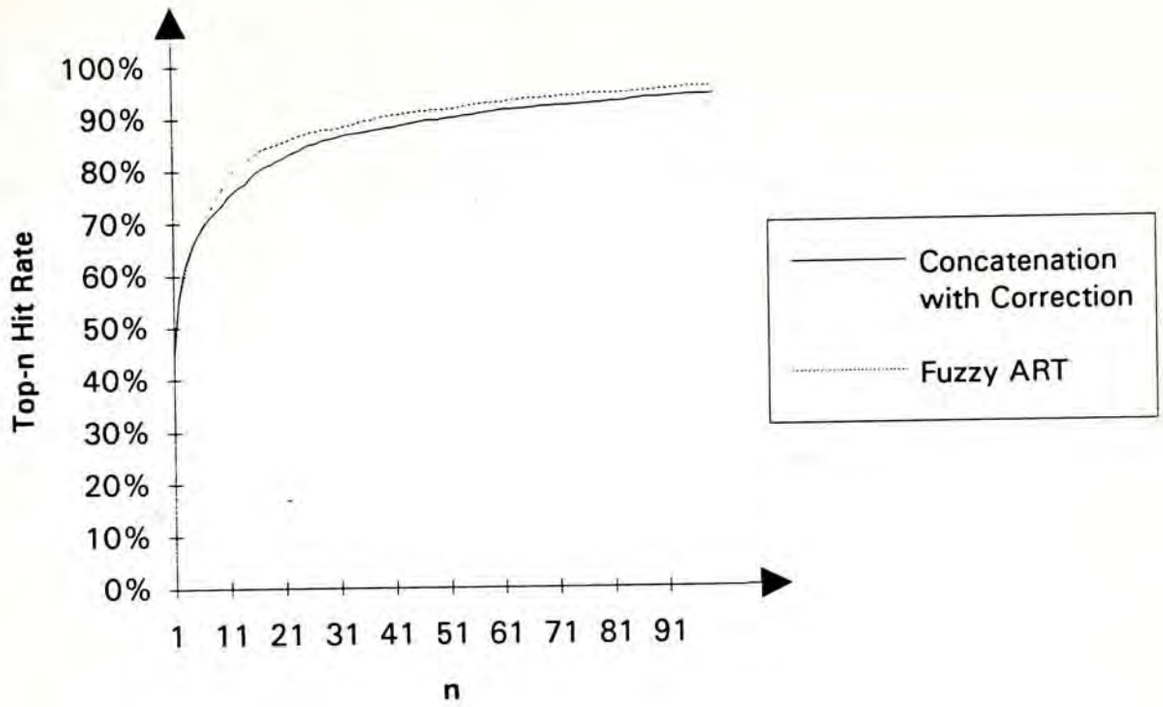

Figure 7.1 Graph of the Top-n Hit Rates of the Syllable Classifier

Note that the top-2 hit rates are greater than the top-1 hit rates (the original recognition rates) by more than 10.0% for both approaches of the syllable classifier, which are also greater than one-half in values. The top-5 hit rates are greater than two-third in values and the top-15 hit rate in Fuzzy ART syllable classifier is greater than four-fifth in value. The top-**n** hit rates for both approaches of the syllable classifier are very similar with small **n** and top-**n** hit rate for Fuzzy ART approach is greater than that for concatenation with correction approach for $\mathbf{n} \geq 7$. Note that there are only negligible differences ($\pm 0.5\%$) between the cases with and without *ending* correction in both approaches, so *ending* correction has no significance in both approaches of the syllable classifier.

Although the top-**n** hit rates with $\mathbf{n} > 1$ do not reflect the actual recognition rate of our Cantonese syllable recognition system, they can show the potential recognition power of the system. Since 15 is still quite small compared with

1437 (i.e. the vocabulary size), it is significant to have 82.4% in the top-15 hit rate out of 1437 categories. The sentence analyzer that proposed as further work in section 8.3 may select the correct goal among **n** possible choices to form valid and meaningful sentence with the help of other recognized syllables in the same sentence.

7.4 Discussions

Although the overall performance of the system is not satisfactory, the approach in the design of the system seems to be in the right direction. First, the use of 4 MLP classifiers in the primary level of the system to recognize the *initials, syllabic segments, endings* and tones make the number of recognition candidates greatly reduced from 1470 to 45. The efforts in training the system are greatly decreased. Second, the *ending* corrector in the intermediate level of the system is proved workable for having 7.5% increment in the recognition rate, though its effort seems to be duplicated after the usage of the syllable classifier. At last, the syllable classifier in the secondary level of the system gives an increase of about 15% on the syllable recognition. Both concatenation with correction approach and Fuzzy ART approach are workable, and Fuzzy ART approach seems to be a little bit better as shown in the top-**n** hit rate analysis. The syllable classifier is shown to have the capability to integrate the recognition results from the primary level classifiers. The unsatisfactory performance of the overall system is not due to the design of the system, but mainly due to the low recognition accuracy of the phoneme classifiers, for the recognition rates of all of them are below 80%, especially that the recognition accuracy of *initial* classifier is less than 60%.

8 Further Works

8.1 Enhancement on Speech Segmentation

Speech segmentation is an important and crucial speech preprocessing in our system. Since speech is dynamic in nature, speech segmentation algorithms have been proposed to pre-segment the speech signal before the classification is carried out. However, speech segmentation is an error-prone classification itself. So, when the speech segmentation is in error, the error will be propagated to the subsequent recognition procedures. Since the phoneme recognition uses the feature parameters from the segmented speech as the input for recognition, the recognition results may be improved with better speech segmentation algorithm.

Ng (1992) used the syllable composite models which were constructed from the trained HMMs to perform the proper segmentation on their corresponding syllables. This could give more accurate segmentation information for the MLP classifier.

8.2 Towards Speaker-Independent System

After the collection of speech data from different speakers, the recognition system should be enhanced to speaker-independent system. It is because speaker-independent recognition system is much more valuable than speaker-dependent system as the system can recognize speech utterances from essentially any speaker. Note that there may be an intermediate step: multi-speaker system before enhancement to speaker-independent system. It is because multi-speaker system may deal with only several speakers, whereas speaker-independent system should be trained and tested by speech data from a large number of speakers. Also, suitable speaker-independent feature parameters should be investigated and extracted from the speech data so as to make the recognition system independent of the speakers of the training data set.

8.3 Towards Speech-to-Text System

It is valuable to point out that there are many homophones in Cantonese, therefore the Chinese characters spoken cannot be obtained from the recognized Cantonese syllables immediately in most cases. For example, the syllable /dzi/ in tone 2 on its own can represent at least 7 commonly used Chinese characters: "son" (子), "elder sister" (姊), "purple" (紫), "stop" (止), "paper" (紙), "only" (只) and "finger" (指), etc. These homophones can only be distinguished by the context, i.e. the words, phrases and sentences they come from. So, several recognized syllables from a same sentence must be analyzed together with the help of the syntactic and semantic rules to give the meaning of the spoken sentence. The model for sentence analysis is shown in Figure 8.1. Also, connected speech recognition must be considered as the syllables in the sentences may not be isolated in the speech database.

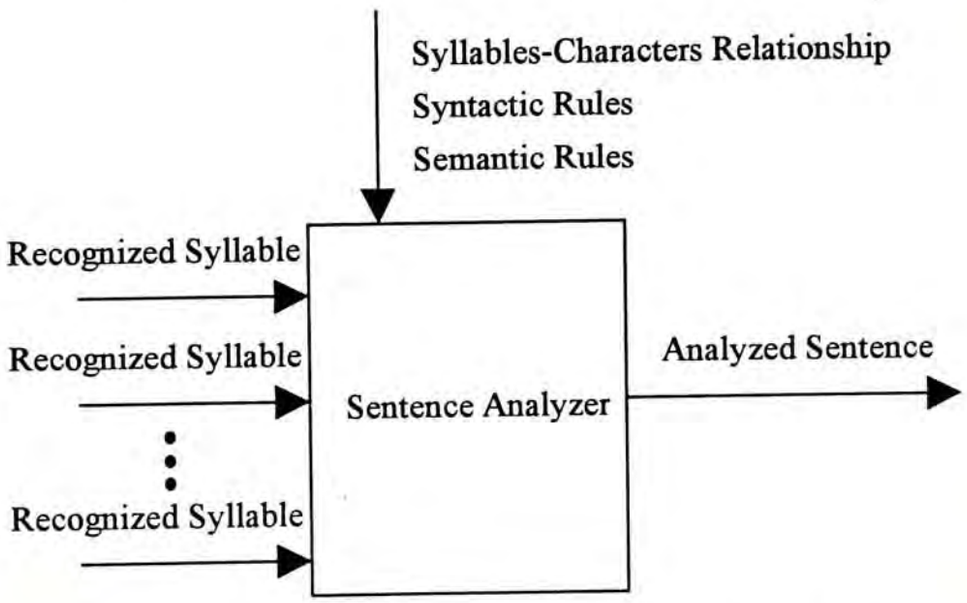

Figure 8.1 The Sentence Analysis Model

9 Conclusions

There are two main points in the development of an isolated Cantonese syllable recognition system: (1) Cantonese is a *monosyllabic tonal* language, and (2) Cantonese syllable full set contains a *large vocabulary* of more than a thousand in size. To deal with these two features, problem decomposition according to the phonology of Cantonese speech is employed for its efficient and effective recognition. A suitable design of classification model is crucial.

The phoneme-based hierarchical neural networks system designed recognizes Cantonese syllables by using syllable classifier to integrate the recognition results of their corresponding phonemes and tones. The system can reduce the number of recognition candidates substantially. The 1470 syllables can be classified by 45 candidates only (18 *initials*, 9 *syllabic segments*, 9 *endings* and 9 tones) with the use of syllable classifier in our system.

The performance of the system under speaker-dependent setting has been evaluated by computer simulation. The classification accuracy of the system on *initials, syllabic segments*, *endings*, tones and syllables are 58.6%, 76.5%, 74.8%, 91.2% and 45.4% respectively for the Cantonese database used. The unsatisfactory performance of the overall system is due to the low recognition accuracy of the phoneme classifiers. On the other hand, the syllable classifier is proven workable, as it can give an increase of 15.9% (from 29.5% to 45.4%) in the syllable recognition rate.

With the phoneme-based design, speech segmentation is very important to our system, because the feature parameters are supplied to the phoneme classifiers according to the phonemic parts segmented from the syllables. However, there is lack

of perfect speech segmentation algorithm. The segmentation of speech into phonemes is still a difficult task itself currently.

There are two main contributions of this research. Firstly, the isolated Cantonese syllable recognition system is designed in a modular approach as a phoneme-based system. This modular approach makes use of the knowledge from the phonology of Cantonese. The complicated task of recognizing Cantonese syllables with large vocabulary size can be broken down into easier sub-tasks of recognizing Cantonese phonemes and tones with very small vocabulary size. So, this modular approach is in a correct direction of large vocabulary Cantonese syllable recognition system design.

Secondly, both approaches of syllable classifier are proven workable and are success in integrating the recognition results of phonemes and tones. It plays a very important role in the whole system. Without the syllable classifier, low recognition accuracy will be given by simple concatenation even when the classifiers in the primary level give acceptable recognition accuracy.

Bibliography

Carpenter, G. A., S. Grossberg and D. B. Rosen (1991) "Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System", *Neural Networks*, vol. 4, pp. 759-771.

Chang, P. C., S. W. Sun and S. H. Chen (1990) "Mandarin Tone Recognition by Multi-Layer Perceptron", *Proceedings of 1990 International Conference on Acoustics, Speech, and Signal Processing*, volume 1 (Albuquerque, New Mexico, April 3-6, 1990), pp. 517-520.

Cheng, Y. H. (1991) *An Efficient Tone Classifier for Speech Recognition of Cantonese*, Master Thesis, Electronic Engineering Department, The Chinese University of Hong Kong, Hong Kong.

Chow, M. K. (周無忌) and B. C. Yiu (饒秉才), ed. (1988) *A Standard Cantonese Pronunciation Syllabary (廣州話標準音字彙)*, The Commercial Press, Hong Kong.

Hashimoto, O. Y. (1972) *Phonology of Cantonese*, Cambridge University Press.

Ho, W. H. (何文匯) (1987) *粵音平仄入門*, Publication (Holdings) Ltd., Hong Kong.

Ho, W. H. (何文匯) (1989) *粵語正音示例*, Publication (Holdings) Ltd., Hong Kong.

Huang, W. M. and R. P. Lippmann (1988) "Neural Net and Traditional Classifiers" in D. Anderson, ed., *Neural Information Processing Systems*, American Institute of Physics, New York.

Lee, K. F. (1989) *Automatic Speech Recognition: the Development of the SPHINX System*, Kluwer Academic Publishers, Norwell, Massachusetts.

Lee, T., P. C. Ching, L. W. Chan and B. Mak (1993) "An NN Based Tone Classifier for Cantonese" in *Proceeding of 1993 International Joint Conference on Neural Networks*, Nagoya, pp. 287-290.

Lippmann, R. P. (1987) "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, vol. 4, no. 2, pp. 4-22.

Lippmann, R. P. (1989) "Review of Neural Networks for Speech Recognition", *Neural Computation*, vol. 1, no. 1, pp. 1-38.

Liu, L. C. and W. J. Yang, H. C. Wang and Y. C. Chang (1989) "Tone Recognition of Polysyllabic words in Mandarin Speech, *Computer Speech and Language*, vol. 3, no. 3, pp. 253-264.

Markel, J. D. and A. H. Gray, Jr. (1976) *Linear Prediction of Speech*, Springer-Verlag, Berlin Heidelberg.

Ng, Y. P. (1992) *Automatic Speech Recognition Isolated Cantonese Words Using Hidden Markov Models and Segmental Neural Networks*, Master Thesis, Engineering Department, University of Cambridge.

Ng Tor Tai Chinese Language Research Centre (NTTCLRC), ed. (1980) *A Cantonese Syllabary with Model Pronunciations (粵音讀例)*, Chinese Language Research Centre, The Chinese University of Hong Kong.

Rabiner, L. R. (1989) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

Rabiner, L. R. (1992) "Speech Recognition Based on Pattern Recognition Approaches" in A. N. Ince, ed., *Digital Speech Processing: Speech Coding, Synthesis and Recognition*, Kluwer Academic Publishers, Boston.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1986) "Learning Internal Representations by Error Propagation" in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing (Explorations in the Microstructure of Cognition), Volume 1: Foundations*, Chapter 8, MIT Press, Cambridge, Massachusetts.

Sondhi, M. M. (1968) "New Methods of Pitch Extraction", *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 262-266.

Tohkura, Y. (1987) "A Weighted Cepstral Distance Measure for Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1414-1422.

van Hemert, J. P. (1991) "Automatic Segmentation of Speech", *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 1008-1012.

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang (1989) "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328-339.

Wang, J. F., C. H. Wu, S. H. Chang and J. Y. Lee (1991) "A Hierarchical Neural Network Model Based on a C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition", *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2141-2146.

Wilpon, J. G. and L. R. Rabiner (1985) "A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 3, pp. 587-594.

Wong, S. L. (黃錫凌) (1941) *A Chinese Syllabary Pronounced Accordingly to the Dialect of Canton (粵音韻彙)*, Chung Hwa Books, Hong Kong.

Yang, W. J., J. C. Lee, Y. C. Chang and H. C. Wang (1988) "Hidden Markov Model for Mandarin Lexical Tone Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 988-992.

Appendix A. Cantonese Syllable Full Set List

No.	Syllable*	C.	Word/Phrase
1	a1	鴉	烏鴉
2	a2	啞	啞巴
3	a3	亞	亞洲
4	ba1	巴	嘴巴
5	ba2	把	把持
6	ba3	壩	水壩
7	ba6	罷	罷工
8	da1	打	一打
9	da2	打	打扮
10	dza1	渣	渣打銀行
11	dza3	炸	爆炸
12	fa1	花	開花
13	fa3	化	變化
14	ga1	加	增加
15	ga2	假	眞假
16	ga3	嫁	出嫁
17	ga4	嘎	嘎仔(日本人)
18	gwa1	瓜	西瓜
19	gwa2	寡	寡婦
20	gwa3	掛	牽掛
21	ha1	蝦	鮮蝦
22	ha4	霞	晚霞
23	ha5	下	一下(不是'夏')
24	ha6	夏	夏天
25	ja5	也	也許
26	ja6	廿	廿八
27	ka1	卡	卡通片
28	kwa1	誇	誇張
29	la1	啦	啦啦隊
30	la3	喇	喇沙書院
31	ma1	媽	姑媽
32	ma4	麻	麻煩
33	ma5	馬	馬場
34	ma6	罵	責罵
35	na4	拿	拿手好戲
36	na5	那	那些
37	nga4	牙	牙齒
38	nga5	瓦	瓦片
39	nga6	迓	迎迓(不是'亞')
40	pa1	趴	趴低
41	pa3	怕	怕死
42	pa4	爬	爬山
43	sa1	沙	沙灘
44	sa2	灑	灑水
45	ta1	他	其他
46	tsa1	叉	刀叉
47	tsa3	詫	詫異
48	tsa4	茶	飲茶
49	wa1	娃	娃娃
50	wa2	畫	圖畫
51	wa4	華	中華
52	wa6	話	說話
53	ai1	唉	唉聲歎氣
54	ai3	隘	狹隘(同"嗌"交)
55	bai2	擺	搖擺
56	bai3	拜	拜訪
57	bai6	敗	敗家仔
58	dai1	獃	書獃子
59	dai2	歹	歹徒
60	dai3	帶	帶領
61	dai6	大	大陸
62	dzai1	齋	食齋
63	dzai3	債	還債
64	dzai6	寨	押寨夫人
65	fai3	快	快速
66	gai1	街	街邊
67	gai2	解	解決
68	gai3	界	界線
69	gwai1	乖	乖巧
70	gwai2	拐	拐杖
71	gwai3	怪	趣怪
72	hai1	揩	揩油
73	hai4	鞋	著鞋
74	hai5	蟹	大閘蟹
75	hai6	械	機械
76	kai2	楷	楷書
77	lai1	拉	拉車
78	lai3	癩	癩哈蟆
79	lai6	賴	依賴
80	mai4	埋	埋堆
81	mai5	買	購買
82	mai6	賣	賣身
83	nai5	奶	牛奶
84	ngai4	涯	天涯
85	ngai6	艾	少艾
86	pai1	派	派頭
87	pai2	牌	橋牌
88	pai3	派	氣派
89	pai4	排	排隊
90	sai2	徙	遷徙
91	sai3	曬	日曬雨淋
92	sai5	舐	舐犢情深(不是'徙')
93	tai1	呔	領呔
94	tai3	太	太陽

No.	Syllable*	C.	Word/Phrase
95	tsai1	猜	猜忌
96	tsai4	柴	柴台
97	wai1	歪	歪風
98	wai4	懷	胸懷
99	wai6	壞	壞人
100	au2	拗	拗手瓜
101	au3	坳	山坳
102	bau1	包	麵包
103	bau2	飽	食飽
104	bau3	爆	爆炸
105	bau6	鮑	管鮑之交
106	dzau1	嘲	嘲笑
107	dzau2	爪	鳳爪
108	dzau3	罩	燈罩
109	dzau6	棹	棹艇
110	gau1	交	交通
111	gau2	狡	奸狡
112	gau3	教	教師
113	hau1	敲	推敲
114	hau2	巧	技巧
115	hau3	孝	孝心
116	hau4	姣	發姣
117	hau6	效	效果
118	kau3	靠	可靠
119	mau1	貓	貓王
120	mau4	矛	矛盾
121	mau5	牡	牡丹
122	mau6	貌	禮貌
123	nau4	錨	拋錨
124	nau6	鬧	鬧市
125	ngau4	餚	佳餚
126	ngau5	咬	咬牙切齒
127	ngau6	樂	樂群館
128	pau1	拋	拋錨
129	pau2	跑	跑步
130	pau3	炮	炮丈
131	pau4	刨	刨冰
132	sau1	梢	樹梢
133	sau2	稍	稍作休息
134	sau3	哨	口哨
135	tsau1	抄	抄功課
136	tsau2	炒	炒魷魚
137	tsau4	巢	雀巢
138	dam1	擔	擔心
139	dam2	膽	蛇膽
140	dam3	擔	擔挑
141	dam6	淡	冷淡
142	dzam1	簪	髮簪
143	dzam2	斬	斬草除根
144	dzam3	湛	精湛

No.	Syllable*	C.	Word/Phrase
145	dzam6	站	車站
146	gam1	監	監督
147	gam2	減	減少
148	gam3	鑑	鑑定
149	ham3	喊	大喊
150	ham4	鹹	鹹水
151	ham6	陷	淪陷(不是'憾')
152	lam2	攬	攬住
153	lam4	藍	藍色
154	lam5	覽	遊覽
155	lam6	艦	軍艦
156	nam4	南	南方
157	nam5	腩	牛腩
158	ngam4	巖	巖石
159	sam1	衫	著衫
160	sam3	三	三思
161	tam1	貪	貪心
162	tam3	探	試探
163	tam4	談	談笑
164	tam5	淡	味好淡
165	tsam1	參	參加
166	tsam2	慘	慘況
167	tsam3	杉	大杉
168	tsam4	蠶	蠶絲
169	an3	晏	晏晝
170	ban1	班	超班
171	ban2	板	黑板
172	ban3	扮	裝扮(不是'辦')
173	ban6	辦	辦公
174	dan1	丹	王丹
175	dan2	蛋	雞蛋
176	dan3	誕	聖誕
177	dan6	但	但願
178	dzan2	棧	客棧
179	dzan3	讚	讚美
180	dzan6	賺	賺錢
181	fan1	翻	翻開
182	fan2	反	相反
183	fan3	氾	氾濫(不是'飯')
184	fan4	凡	平凡
185	fan6	飯	食飯
186	gan1	奸	奸商
187	gan2	簡	簡單
188	gan3	諫	進諫
189	gwan1	關	關心
190	gwan3	慣	習慣
191	han1	慳	慳錢
192	han4	閑	得閑
193	han6	限	限度
194	lan4	欄	欄杆

No.	Syllable*	C.	Word/Phrase
195	lan5	懶	懶惰
196	lan6	爛	爛仔
197	man4	蠻	野蠻
198	man5	晚	星島晚報
199	man6	萬	百萬富翁
200	nan4	難	難度
201	nan6	難	災難
202	ngan4	顏	顏色
203	ngan5	眼	眼睛
204	ngan6	雁	大雁
205	pan1	攀	攀山
206	pan3	盼	盼望
207	san1	山	高山
208	san2	散	驚風散(不是'傘')
209	san3	傘	雨傘
210	san4	潺	潺潺流水
211	tan1	灘	沙灘
212	tan2	坦	平坦
213	tan3	炭	火炭
214	tan4	檀	檀香山
215	tsan1	餐	早餐
216	tsan2	產	產品
217	tsan3	燦	亞燦
218	tsan4	殘	殘忍
219	wan1	彎	彎曲
220	wan4	還	償還
221	wan5	挽	挽救
222	wan6	幻	幻象
223	ang1	罌	錢罌
224	dzang1	爭	爭住舉手
225	gang1	耕	耕耘
226	hang1	坑	大坑道
227	hang4	行	行路
228	kwang1	框	鏡框
229	kwang3	逛	逛街
230	lang1	冷	打冷
231	lang5	冷	寒冷
232	mang4	盲	色盲
233	mang5	猛	威猛
234	mang6	孟	孟子
235	ngang6	硬	軟硬
236	pang1	烹	烹飪
237	pang4	膨	膨脹
238	pang5	棒	棒球
239	sang1	生	生仔(不是'笙')
240	sang2	省	節省
241	tsang1	撐	支撐
242	tsang2	橙	甜橙
243	wang4	橫	打橫
244	ap8	鴨	醜小鴨

No.	Syllable*	C.	Word/Phrase
245	dap7	嗒	嗒落好味
246	dap8	答	答案
247	dap9	踏	踏步
248	dzap9	雜	雜技
249	gap8	夾	夾心
250	hap8	呷	呷醋
251	hap9	峽	三峽
252	lap9	臘	臘味
253	nap9	納	採納
254	sap8	圾	垃圾
255	tap8	塔	燈塔
256	tsap8	插	插手
257	at8	壓	壓力
258	bat8	八	八卦
259	dat8	笪	大笪地
260	dat9	達	到達
261	dzat8	紮	包紮
262	fat8	法	辦法
263	gwat8	刮	搜刮
264	lat9	辣	辣椒
265	mat8	抹	抹檯(不是'抹'殺)
266	nat9	捺	撇捺
267	sat8	殺	殺人
268	tat8	撻	撻訂
269	tsat8	擦	擦鞋
270	wat8	挖	挖掘
271	wat9	滑	順滑
272	ak7	厄	厄運
273	bak7	迫	逼迫
274	bak8	伯	世伯
275	bak9	白	雪白
276	dzak8	窄	窄路相逢
277	dzak9	摘	摘荔枝
278	gak8	格	格式
279	gwak8	摑	摑了一巴
280	hak7	赫	顯赫
281	hak8	客	送客
282	jak8	喫	喫飯(不是'吃')
283	lak9	肋	肋骨
284	mak8	擘	擘開
285	ngak9	額	額頭
286	pak8	拍	節拍
287	sak8	索	索取(不是'朔')
288	tsak7	咋	叱咋
289	tsak8	拆	遷拆
290	tsak9	賊	賊匪
291	wak9	或	或者
292	Ai2	矮	小矮人
293	Ai3	縊	自縊
294	bAi1	跛	跛腳鴨

No.	Syllable*	C.	Word/Phrase
295	bAi3	閉	關閉
296	bAi6	幣	貨幣
297	dAi1	低	高低
298	dAi2	底	底線
299	dAi3	帝	皇帝
300	dAi6	弟	兄弟
301	dzAi1	擠	擠迫
302	dzAi2	仔	公仔麵
303	dzAi3	祭	祭祖
304	dzAi6	滯	停滯
305	fAi1	揮	指揮
306	fAi3	費	經費
307	fAi6	吠	狗吠
308	gAi1	雞	公雞
309	gAi3	計	計謀
310	gwAi1	歸	歸去
311	gwAi2	鬼	鬼神
312	gwAi3	貴	富貴
313	gwAi6	跪	跪地
314	hAi4	奚	奚落
315	hAi6	系	系統
316	jAi6	曳	搖曳
317	kAi1	溪	溪水
318	kAi2	啓	啓示
319	kAi3	契	契仔
320	kwAi1	規	規則
321	kwAi4	葵	葵青區
322	kwAi5	揆	英揆
323	lAi4	黎	黎明
324	lAi5	禮	送禮
325	lAi6	麗	麗晶酒店
326	mAi1	咪	無線咪
327	mAi4	迷	球迷
328	mAi5	米	米飯班主
329	nAi4	泥	黃泥
330	ngAi4	危	危險
331	ngAi5	蟻	蟻民
332	ngAi6	毅	堅毅
333	pAi1	批	批准
334	sAi1	西	西方
335	sAi2	洗	洗衫
336	sAi3	細	細心
337	sAi6	誓	發誓
338	tAi1	梯	樓梯
339	tAi2	體	身體
340	tAi3	替	代替
341	tAi4	堤	長堤
342	tsAi1	妻	夫妻
343	tsAi3	砌	堆砌
344	tsAi4	齊	整齊

No.	Syllable*	C.	Word/Phrase
345	wAi1	威	威風
346	wAi2	委	委員
347	wAi3	畏	畏懼
348	wAi4	圍	包圍
349	wAi5	偉	偉大
350	wAi6	胃	腸胃
351	Au1	歐	歐洲
352	Au2	嘔	作嘔
353	Au3	漚	"漚"病
354	dAu1	兜	肚兜
355	dAu2	斗	北斗
356	dAu3	鬥	戰鬥
357	dAu6	逗	逗號
358	dzAu1	周	周圍
359	dzAu2	酒	飲酒
360	dzAu3	奏	演奏
361	dzAu6	就	遷就
362	fAu2	否	否認
363	fAu4	浮	浮台
364	fAu6	埠	遊埠
365	gAu2	九	牌九
366	gAu3	救	急救
367	gAu6	舊	念舊
368	hAu2	口	開口
369	hAu4	猴	猴子
370	hAu5	厚	厚薄
371	hAu6	後	後悔
372	jAu1	優	優點
373	jAu2	黝	黝黑
374	jAu3	幼	幼兒
375	jAu4	油	豬油
376	jAu5	有	佔有
377	jAu6	右	左右
378	kAu1	溝	溝通
379	kAu3	扣	扣押
380	kAu4	求	祈求
381	kAu5	舅	舅仔
382	lAu1	褸	皮褸
383	lAu4	流	流水
384	lAu5	柳	楊柳
385	lAu6	漏	漏水
386	mAu1	痾	地痾
387	mAu4	謀	陰謀
388	mAu5	某	某年某月
389	mAu6	貿	貿易
390	nAu2	扭	扭計
391	ngAu1	勾	勾起心事
392	ngAu4	牛	牛郎
393	ngAu5	偶	偶然
394	sAu1	修	修理

No.	Syllable*	C.	Word/Phrase
395	sAu2	手	手指
396	sAu3	獸	野獸
397	sAu4	愁	憂愁
398	sAu6	受	接受
399	tAu1	偷	偷車
400	tAu3	透	透心涼
401	tAu4	頭	頭獎
402	tsAu1	秋	秋風秋雨
403	tsAu2	丑	小丑
404	tsAu3	湊	湊仔
405	tsAu4	綢	絲綢之路
406	Am1	鵪	鵪鶉
407	Am2	黯	黯然失色
408	Am3	暗	黑暗
409	bAm1	泵	水泵
410	dzAm1	針	針線
411	dzAm2	枕	枕頭
412	dzAm3	浸	浸水
413	dzAm6	朕	(皇帝自稱)
414	gAm1	金	黃金
415	gAm2	敢	勇敢
416	gAm3	禁	禁止
417	gAm6	撳	撳掣
418	hAm1	堪	不堪設想
419	hAm2	砍	砍伐
420	hAm3	勘	勘探
421	hAm4	含	含情
422	hAm6	憾	遺憾
423	jAm1	音	音樂
424	jAm2	飲	飲水
425	jAm3	蔭	林蔭
426	jAm4	淫	淫蟲
427	jAm6	任	任意
428	kAm1	襟	襟章
429	kAm4	琴	鋼琴
430	kAm5	妗	妗婆
431	lAm4	林	林青霞
432	lAm5	凜	威風凜凜
433	sAm1	心	心情
434	sAm2	審	審訊
435	sAm3	滲	滲透
436	sAm4	岑	岑建勳
437	sAm6	甚	欺人太甚
438	tsAm1	侵	入侵
439	tsAm2	寢	安寢
440	tsAm4	尋	尋找
441	bAn1	奔	奔馳
442	bAn2	品	品格
443	bAn3	殯	殯儀館
444	bAn6	笨	笨蛋
445	dAn2	躉	擁躉
446	dAn6	燉	燉冬菇
447	dzAn1	眞	眞假
448	dzAn3	振	振作
449	dzAn6	陣	陣地
450	fAn1	昏	黃昏
451	fAn2	粉	粉果
452	fAn3	訓	教訓
453	fAn4	焚	焚燒
454	fAn5	憤	氣憤
455	fAn6	份	份量
456	gAn1	巾	毛巾
457	gAn2	緊	緊張
458	gAn6	近	就近
459	gwAn1	君	欺君
460	gwAn2	滾	滾水
461	gwAn3	棍	木棍
462	gwAn6	郡	郡主
463	hAn2	很	很多
464	hAn4	痕	身痕
465	hAn6	恨	痛恨
466	jAn1	因	原因
467	jAn2	忍	忍受
468	jAn3	印	印刷
469	jAn4	人	好人
470	jAn5	引	引導
471	jAn6	刃	刀刃
472	kAn4	勤	勤力
473	kAn5	近	遠近
474	kwAn1	坤	乾坤
475	kwAn2	菌	細菌
476	kwAn3	困	圍困
477	kwAn4	群	群眾
478	mAn1	蚊	一隻蚊
479	mAn4	文	斯文
480	mAn5	敏	敏感
481	mAn6	問	問題
482	nAn2	撚	撚手小菜
483	ngAn4	銀	銀行
484	ngAn6	靭	靭力
485	pAn3	噴	噴氣
486	pAn4	貧	貧窮
487	sAn1	身	身體
488	sAn4	神	神仙
489	sAn5	腎	鴨腎(不是'愼')
490	sAn6	愼	愼重
491	tAn1	吞	吞嗞
492	tAn3	褪	褪色(不是'退')
493	tsAn1	親	親人
494	tsAn2	疹	麻疹

No.	Syllable*	C.	Word/Phrase
495	tsAn3	襯	陪襯
496	tsAn4	陳	陳大文
497	wAn1	溫	溫暖
498	wAn2	穩	穩健
499	wAn3	慍	慍色
500	wAn4	雲	風雲
501	wAn5	允	答允
502	wAn6	運	運作
503	Ang1	鶯	黃鶯
504	bAng1	崩	崩潰
505	dAng1	登	登高
506	dAng2	等	上等
507	dAng3	凳	檯凳
508	dAng6	鄧	鄧小平
509	dzAng1	憎	憎死你
510	dzAng6	贈	贈送
511	gAng1	羹	湯羹
512	gAng2	梗	贏梗
513	gAng3	更	更加
514	gwAng1	轟	轟炸
515	hAng1	亨	大亨
516	hAng2	肯	肯定
517	hAng4	恆	恆久
518	hAng5	悻	悻然
519	hAng6	幸	幸運
520	mAng4	盟	港同盟
521	nAng4	能	賢能
522	pAng4	朋	朋友
523	tAng4	藤	藤條
524	tsAng4	層	基層
525	wAng4	宏	宏觀
526	dzAp7	汁	果汁
527	gAp7	急	急促
528	gAp8	鴿	乳鴿
529	hAp7	洽	融洽
530	hAp9	合	合作
531	jAp7	泣	哭泣
532	jAp9	入	進入
533	kAp7	吸	吸收
534	kAp9	及	及時
535	lAp7	笠	笠落去
536	lAp9	立	立即(不是'臘')
537	nAp7	粒	粒粒橙
538	sAp7	濕	濕度
539	sAp9	十	十誡
540	tsAp7	輯	專輯
541	bAt7	筆	鉛筆
542	bAt9	拔	拔河
543	dAt9	凸	凸字
544	dzAt7	質	質詢

No.	Syllable*	C.	Word/Phrase
545	dzAt9	疾	眼疾
546	fAt7	忽	疏忽
547	fAt9	佛	活佛
548	gAt7	吉	吉利
549	gwAt7	骨	骨頭
550	gwAt9	掘	掘地
551	hAt7	乞	乞衣
552	hAt9	轄	管轄
553	jAt7	一	一二三
554	jAt9	日	日出
555	kAt7	咳	咳藥水
556	lAt7	甩	走甩
557	mAt7	乜	乜東東
558	mAt9	勿	請勿吸煙
559	ngAt9	屹	屹立
560	pAt7	匹	馬匹
561	sAt7	瑟	瑟縮
562	sAt9	實	事實
563	tsAt7	七	田七
564	wAt7	屈	委屈
565	wAt9	核	核突(不是'轄')
566	Ak7	握	握手(不是'厄')
567	bAk7	北	東北
568	dAk7	得	得獎
569	dAk9	特	特色
570	dzAk7	則	原則
571	hAk7	黑	黑色(不是'赫')
572	lAk9	勒	勒緊褲頭(不是'肋')
573	mAk9	陌	陌生
574	sAk7	塞	阻塞
575	tsAk7	測	預測(不是'𠊎宅')
576	bei1	悲	慈悲
577	bei2	比	比賽
578	bei3	秘	秘魯
579	bei6	備	設備
580	dei6	地	地點
581	fei1	非	非常
582	fei2	匪	土匪
583	fei4	肥	肥豬
584	gei1	基	基礎
585	gei2	己	自己
586	gei3	寄	郵寄
587	gei6	忌	禁忌
588	hei1	希	希望
589	hei2	喜	歡喜
590	hei3	汽	汽車
591	kei1	崎	崎嶇
592	kei2	棋	捉棋
593	kei3	冀	希冀
594	kei4	奇	奇怪

No.	Syllable*	C.	Word/Phrase
595	kei5	企	企業
596	lei4	離	離開
597	lei5	里	公里
598	lei6	利	利害
599	mei1	瞇	瞇埋眼
600	mei4	眉	眼眉
601	mei5	美	優美
602	mei6	味	好味
603	nei4	尼	尼姑
604	nei5	你	你我
605	nei6	膩	油膩
606	pei1	披	披星戴月
607	pei2	鄙	鄙視
608	pei3	屁	放屁
609	pei4	皮	皮膚
610	pei5	婢	奴婢
611	sei2	死	死亡
612	sei3	四	四面八方
613	be1	啤	啤酒
614	de1	爹	沙爹牛肉
615	dze1	遮	雨遮
616	dze2	者	長者
617	dze3	借	借錢
618	dze6	謝	多謝
619	fe1	啡	咖啡
620	ge3	嘅	我嘅書
621	je4	爺	老爺
622	je5	野	田野
623	je6	夜	夜晚
624	ke2	茄	番茄
625	ke4	騎	騎馬
626	me1	咩	羊咩
627	me2	歪	借歪
628	ne1	呢	這是什麼呢？
629	se1	些	些少
630	se2	寫	寫作
631	se3	瀉	肚瀉
632	se4	蛇	蛇膽
633	se5	社	社交
634	se6	射	發射
635	tse1	車	汽車
636	tse2	且	而且
637	tse3	斜	落斜
638	tse4	邪	邪教
639	beng2	餅	餅乾
640	beng3	柄	話柄
641	beng6	病	病假
642	deng1	釘	釘書機
643	deng2	頂	山頂
644	deng3	掟	掟煲
645	deng6	訂	落訂
646	dzeng1	精	精乖伶俐
647	dzeng2	井	打井
648	dzeng3	正	坐正
649	dzeng6	鄭	鄭成功
650	geng1	驚	心驚驚
651	geng2	頸	頸巾
652	geng3	鏡	照鏡
653	heng1	輕	輕飄飄
654	jeng4	贏	贏家
655	leng3	靚	靚女
656	leng4	靈	好靈
657	leng5	領	衫領
658	meng2	名	你叫乜名？
659	meng6	命	爛命一條
660	peng4	平	平價貨
661	seng1	腥	血腥
662	seng2	醒	嘈醒
663	seng4	成	做成生意
664	teng1	廳	坐花廳
665	teng5	艇	快艇
666	tseng1	青	青色(不是'清')
667	tseng2	請	請客
668	dek9	笛	笛子
669	dzek8	隻	一隻鞋
670	dzek9	蓆	草蓆
671	hek8	吃	吃喝
672	kek9	劇	喜劇
673	pek8	劈	劈開
674	sek8	錫	痛錫
675	sek9	石	大石
676	tek8	踢	踢波
677	tsek8	尺	咫尺
678	dzi1	諮	諮詢
679	dzi2	子	子女
680	dzi3	至	至尊
681	dzi6	自	自我
682	ji1	衣	衣服
683	ji2	倚	倚賴
684	ji3	意	意思
685	ji4	而	而且
686	ji5	以	以爲
687	ji6	二	第二
688	si1	詩	詩意
689	si2	史	歷史
690	si3	試	試驗
691	si4	時	時間
692	si5	市	鬧市
693	si6	事	事情
694	tsi1	雌	雌雄

No.	Syllable*	C.	Word/Phrase
695	tsi2	此	如此
696	tsi3	次	其次
697	tsi4	慈	仁慈
698	tsi5	似	類似
699	biu1	標	目標
700	biu2	表	儀表
701	diu1	凋	凋謝
702	diu3	吊	吊鐘花
703	diu6	掉	抹掉
704	dziu1	蕉	香蕉
705	dziu2	沼	沼澤
706	dziu3	照	照明
707	dziu6	趙	趙紫陽
708	giu1	嬌	詐嬌
709	giu2	矯	矯正
710	giu3	叫	呼叫
711	giu6	撬	撬開
712	hiu1	梟	梟雄
713	hiu2	曉	知曉
714	hiu3	竅	七竅生煙(不是'kiu3')
715	jiu1	腰	彎腰
716	jiu2	妖	妖怪
717	jiu3	要	要員
718	jiu4	搖	搖擺
719	jiu6	耀	耀安
720	kiu2	橋	一代橋王
721	kiu3	竅	竅門(不是'hiu3')
722	kiu4	橋	橋牌
723	liu1	撩	撩起(不是'僚')
724	liu2	料	有料
725	liu4	僚	官僚
726	liu5	了	了斷
727	liu6	料	料到
728	miu4	苗	豆苗
729	miu5	秒	分秒必爭
730	miu6	妙	奇妙
731	niu5	鳥	雀鳥
732	niu6	尿	驗尿
733	piu1	飄	輕飄飄
734	piu3	票	售票
735	piu4	嫖	嫖客
736	siu1	消	消費
737	siu2	小	小朋友
738	siu3	少	少年
739	siu4	韶	韶山
740	siu6	邵	邵逸夫
741	tiu1	挑	挑夫
742	tiu2	條	油條
743	tiu3	跳	跳舞
744	tiu4	條	條件
745	tiu5	窕	窈窕
746	tsiu1	超	超人
747	tsiu3	俏	嬌俏
748	tsiu4	潮	潮水
749	dim2	點	鐘點
750	dim3	店	商店
751	dzim1	尖	尖沙嘴
752	dzim3	佔	佔據
753	dzim6	漸	漸露頭角
754	gim1	兼	兼職
755	gim2	檢	檢查
756	gim3	劍	劍擊
757	gim6	儉	勤儉
758	him1	謙	謙虛
759	him2	險	危險
760	him3	欠	拖欠
761	jim1	閹	閹割
762	jim2	掩	掩護
763	jim3	厭	討厭
764	jim4	炎	發炎
765	jim5	染	傳染
766	jim6	驗	化驗
767	kim4	黔	黔驢技窮
768	lim4	廉	廉潔
769	lim5	斂	斂財(原爲'lim6')
770	lim6	臉	臉孔(原爲'lim5')
771	nim1	拈	信手拈來
772	nim4	粘	粘郵票
773	nim6	念	思念
774	sim2	閃	閃電
775	sim4	嬋	貂嬋
776	sim6	贍	贍養費(不是'善')
777	tim1	添	添丁
778	tim4	甜	甜品
779	tim5	恬	恬靜(原爲'tim2')
780	tsim1	簽	簽署
781	tsim2	諂	諂媚
782	tsim3	塹	天塹
783	tsim4	潛	潛水
784	bin1	邊	花邊
785	bin2	貶	貶值
786	bin3	變	變化
787	bin6	便	便利店
788	din1	癲	發癲
789	din2	典	字典
790	din3	墊	墊錢
791	din6	電	電視機
792	dzin1	煎	煎魚
793	dzin2	剪	剪紙
794	dzin3	箭	箭頭

No.	Syllable*	C.	Word/Phrase
795	dzin6	賤	犯賤
796	gin1	堅	堅持
797	gin3	見	意見
798	gin6	件	逐件
799	hin1	軒	軒尼詩道
800	hin2	顯	明顯
801	hin3	憲	憲法
802	jin1	煙	食煙
803	jin2	偃	偃旗息鼓
804	jin3	宴	宴會
805	jin4	言	言語
806	jin6	現	出現
807	kin4	乾	乾坤
808	lin2	鏈	金鏈
809	lin4	連	連接
810	lin6	練	練習
811	min4	棉	棉花糖
812	min5	免	免費
813	min6	麵	公仔麵
814	nin4	年	年月
815	pin1	編	編寫
816	pin3	騙	騙子
817	sin1	先	優先
818	sin2	癬	腳癬
819	sin3	線	針線
820	sin5	鱔	黃鱔
821	sin6	善	善哉
822	tin1	天	天空
823	tin4	田	田園
824	tsin1	千	老千
825	tsin2	淺	淺水灣
826	tsin4	前	前後
827	tsin5	踐	實踐(不是'淺')
828	bing1	冰	溜冰
829	bing2	丙	甲乙丙
830	bing3	併	合併(不是'拼')
831	bing6	並	並且
832	ding1	仃	孤苦伶仃
833	ding2	鼎	鼎足而立
834	ding3	訂	訂定
835	ding6	定	決定
836	dzing1	晶	晶瑩
837	dzing2	整	整齊
838	dzing3	政	政策
839	dzing6	靜	靜止
840	ging1	京	北京
841	ging2	警	警告
842	ging3	敬	敬佩
843	ging6	勁	勁抽
844	gwing2	炯	炯炯有神
845	hing1	兄	兄弟
846	hing3	慶	慶祝
847	jing1	英	英雄
848	jing2	影	電影
849	jing3	應	反應
850	jing4	形	形像
851	jing6	認	認識
852	king1	傾	傾心
853	king2	頃	公頃(不是'king5')
854	king4	鯨	鯨魚
855	ling1	拎	拎起
856	ling4	零	零的突破
857	ling5	領	領袖
858	ling6	另	另外
859	ming4	名	姓名
860	ming5	冥	冥王星
861	ming6	命	生命
862	ning4	寧	寧靜
863	ning6	佞	巧言佞色
864	ping1	娉	娉婷
865	ping3	聘	聘請
866	ping4	平	平均
867	sing1	星	星星月亮
868	sing2	醒	甦醒
869	sing3	性	人性
870	sing4	成	成功
871	sing6	盛	盛事
872	ting1	聽	聽見(不是'廳/ting3')
873	ting2	亭	涼亭(不是'停')
874	ting3	聽	聽覺
875	ting4	停	停止
876	ting5	挺	挺胸(原爲'ting3')
877	tsing1	清	清澈
878	tsing2	請	邀請
879	tsing3	稱	稱職
880	tsing4	情	感情
881	wing1	扔	扔掉
882	wing4	榮	光榮
883	wing5	永	永遠
884	wing6	泳	暢泳
885	dip9	蝶	蝶式
886	dzip8	接	間接
887	gip8	劫	打劫
888	hip8	怯	怯場
889	hip9	協	協助(不是'怯')
890	jip9	頁	黃頁
891	lip9	獵	打獵
892	nip9	聶	聶榮臻
893	sip8	涉	干涉
894	tip8	貼	張貼

No.	Syllable*	C.	Word/Phrase
895	tsip8	妾	妾氏
896	bit7	必	必須
897	bit8	鱉	甕中之鱉(不是'別')
898	bit9	別	別去
899	dit8	跌	跌低
900	dit9	秩	秩序
901	dzit8	節	節日
902	dzit9	截	截止
903	git8	結	結果
904	git9	傑	傑出
905	hit8	歇	歇腳(不是'揭')
906	jit9	熱	熱狗
907	kit8	揭	揭曉
908	lit9	列	排列
909	mit7	搣	搣下你
910	mit9	滅	滅口
911	pit8	鼈	驚鴻一鼈
912	sit8	洩	洩漏
913	sit9	蝕	"蝕"本
914	tit8	鐵	鋼鐵
915	tsit8	設	假設
916	bik7	迫	強迫
917	dik7	的	的確
918	dik9	滴	點滴
919	dzik7	即	隨即
920	dzik9	夕	朝夕
921	gik7	激	感激
922	gik9	極	太極
923	gwik7	隙	罅隙(原爲'kwik7')
924	jik7	益	得益
925	jik9	翼	雞翼
926	lik7	礫	瓦礫
927	lik9	力	力量
928	mik9	覓	尋覓
929	nik7	匿	匿藏
930	pik7	辟	開辟
931	sik7	色	色彩
932	sik9	食	飲食
933	tik7	惕	警惕
934	tsik7	戚	親戚
935	wik9	域	地域
936	ou3	澳	澳洲
937	bou1	煲	煲水
938	bou2	保	保持
939	bou3	報	報紙
940	bou6	步	進步
941	dou1	刀	飛刀
942	dou2	賭	賭錢
943	dou3	到	遲到
944	dou6	道	道路
945	dzou1	租	交租
946	dzou2	早	早餐
947	dzou3	灶	爐灶
948	dzou6	做	做工
949	gou1	高	高尚
950	gou2	稿	起稿
951	gou3	告	控告
952	hou1	蒿	茼蒿菜
953	hou2	好	好壞
954	hou3	耗	消耗
955	hou4	豪	豪門
956	hou6	號	編號
957	lou1	撈	撈偏門
958	lou2	佬	大佬
959	lou4	勞	勞工
960	lou5	老	老師
961	lou6	路	出路
962	mou1	髦	時髦(不是'毛')
963	mou4	巫	巫婆
964	mou5	母	母親
965	mou6	務	任務
966	nou4	奴	奴才
967	nou5	努	努力
968	nou6	怒	憤怒
969	ngou4	熬	煎熬
970	ngou6	傲	驕傲
971	pou1	鋪	床鋪
972	pou2	普	普通
973	pou3	鋪	地鋪
974	pou4	袍	禮袍
975	pou5	抱	擁抱
976	sou1	蘇	蘇東坡
977	sou2	嫂	大嫂
978	sou3	數	數學
979	tou1	滔	浪滔滔
980	tou2	土	泥土
981	tou3	吐	吐痰
982	tou4	桃	桃花
983	tou5	肚	大肚
984	tsou1	操	體操
985	tsou2	草	草原
986	tsou3	燥	乾燥
987	tsou4	曹	曹操
988	o1	柯	荊柯
989	bo1	波	電波
990	bo3	播	廣播
991	do1	多	多餘
992	do2	躲	躲藏(不是'朵')
993	do6	墮	墮落
994	dzo2	左	左翼

No.	Syllable*	C.	Word/Phrase
995	dzo3	佐	員佐級
996	dzo6	助	助聽器
997	fo1	科	科目
998	fo2	火	救火
999	fo3	貨	送貨
1000	go1	哥	表哥
1001	go3	個	個性
1002	gwo1	戈	戈壁灘(不是' 哥')
1003	gwo2	果	果斷(不是'go2')
1004	gwo3	過	經過(不是' 個')
1005	ho1	苛	苛刻
1006	ho2	可	認可
1007	ho4	何	何必
1008	ho6	賀	賀禮
1009	lo1	囉	囉嗦
1010	lo2	裸	裸體
1011	lo4	羅	開羅
1012	mo1	麼	甚麼
1013	mo2	摸	摸索
1014	mo4	磨	磨擦
1015	no2	娜	婀娜(原爲'no5')
1016	no4	挪	挪威
1017	no6	糯	糯米飯
1018	ngo4	俄	俄羅斯
1019	ngo5	我	自我
1020	ngo6	餓	肚餓
1021	po2	頗	頗有成績
1022	po3	破	破爛
1023	po4	婆	老婆
1024	so1	梳	梳頭
1025	so2	所	所以
1026	so4	傻	傻笑
1027	to1	拖	拍拖
1028	to4	鴕	鴕鳥
1029	to5	妥	妥當(原爲'to2')
1030	tso1	初	當初
1031	tso2	楚	痛楚
1032	tso3	錯	認錯
1033	tso4	鋤	鋤頭
1034	tso5	坐	坐低(不是' 助')
1035	wo1	窩	狗窩
1036	wo4	和	和氣
1037	wo5	禍	禍𠱁
1038	wo6	禍	闖禍
1039	oi1	哀	悲哀
1040	oi2	藹	和藹
1041	oi3	愛	可愛
1042	doi6	代	代替
1043	dzoi1	災	天災
1044	dzoi2	宰	主宰

No.	Syllable*	C.	Word/Phrase
1045	dzoi3	再	再次
1046	dzoi6	在	存在
1047	goi1	該	應該
1048	goi2	改	改變
1049	goi3	蓋	杯蓋
1050	hoi1	開	開門
1051	hoi2	海	海洋
1052	hoi6	害	害人害己
1053	koi3	鈣	鈣質
1054	loi4	來	來賓
1055	noi6	內	內部
1056	ngoi4	呆	癡呆
1057	ngoi6	外	外國
1058	soi1	鰓	魚鰓
1059	toi1	胎	胎兒
1060	toi4	台	台灣
1061	toi5	怠	怠慢
1062	tsoi2	彩	彩色
1063	tsoi3	菜	青菜
1064	tsoi4	才	才幹
1065	on1	安	平安
1066	on3	案	方案
1067	gon1	肝	心肝
1068	gon2	趕	追趕
1069	gon3	幹	幹線
1070	hon1	看	看管
1071	hon2	罕	罕有
1072	hon3	漢	漢奸
1073	hon4	韓	韓國
1074	hon5	旱	旱災
1075	hon6	汗	汗水
1076	ngon6	岸	岸邊
1077	ong3	盎	春意盎然
1078	bong1	幫	黑幫
1079	bong2	榜	標榜
1080	bong6	磅	磅重
1081	dong1	當	當初
1082	dong2	黨	共產黨
1083	dong3	檔	開檔
1084	dong6	蕩	遊蕩
1085	dzong1	妝	化妝
1086	dzong3	葬	殮葬
1087	dzong6	狀	作狀
1088	fong1	方	方向
1089	fong2	訪	採訪
1090	fong3	放	開放
1091	fong4	防	邊防
1092	gong1	剛	剛才
1093	gong2	港	港口
1094	gong3	降	降低

No.	Syllable*	C.	Word/Phrase
1095	gwong1	光	光明
1096	gwong2	廣	廣闊
1097	hong1	康	健康
1098	hong2	慷	慷慨
1099	hong3	炕	炕熱
1100	hong4	杭	杭州
1101	hong6	巷	巷仔
1102	kong3	抗	抵抗
1103	kwong3	礦	礦場
1104	kwong4	狂	瘋狂
1105	long4	狼	豺狼
1106	long5	朗	開朗
1107	long6	浪	浪子
1108	mong4	忙	繁忙
1109	mong5	網	網球
1110	mong6	望	希望
1111	nong4	囊	背囊
1112	ngong4	昂	昂首挺胸
1113	ngong6	戇	戇居
1114	pong3	謗	誹謗(不是'蚌')
1115	pong4	旁	旁邊
1116	pong5	蚌	蚌的啓示
1117	song1	桑	滄海桑田
1118	song2	爽	秋高氣爽
1119	song3	喪	頹喪
1120	tong1	湯	煲湯
1121	tong2	倘	倘若
1122	tong3	燙	燙傷
1123	tong4	堂	飯堂
1124	tsong1	倉	貨倉
1125	tsong2	廠	工廠
1126	tsong3	創	創傷
1127	tsong4	床	床單
1128	wong1	汪	水汪汪
1129	wong2	枉	冤枉
1130	wong4	皇	皇帝
1131	wong5	往	往返
1132	wong6	旺	興旺
1133	got8	割	分割
1134	hot8	渴	口渴
1135	ok8	惡	兇惡
1136	bok8	博	博愛
1137	bok9	薄	刻薄
1138	dok9	度	量度
1139	dzok8	作	寫作
1140	dzok9	鑿	言之鑿鑿
1141	fok8	霍	霍亂
1142	gok8	角	主角
1143	gwok8	國	中國(不是'角')
1144	hok8	殼	脫殼
1145	hok9	學	學習
1146	kok8	確	的確
1147	kwok8	擴	擴大(不是'礦')
1148	lok8	烙	烙鐵(不是'落')
1149	lok9	落	落雨
1150	mok7	剝	剝皮
1151	mok9	寞	寂寞
1152	nok9	諾	承諾
1153	ngok9	岳	岳父
1154	pok8	撲	撲朔迷離
1155	sok8	朔	撲朔迷離
1156	tok8	托	襯托
1157	wok9	穫	收穫
1158	doe2	朵	一朵花(不是'躲')
1159	hoe1	靴	皮靴
1160	toe3	唾	唾液
1161	doey1	堆	堆積
1162	doey2	隊	排隊
1163	doey3	對	反對
1164	doey6	隊	隊友
1165	dzoey1	追	追趕
1166	dzoey2	嘴	嘴巴
1167	dzoey3	醉	醉酒
1168	dzoey6	罪	罪行
1169	goey1	居	居民
1170	goey2	舉	舉起
1171	goey3	句	句子
1172	goey6	巨	巨大
1173	hoey1	虛	虛無
1174	hoey2	許	許多
1175	hoey3	去	去向
1176	joey5	蕊	花蕊(不是'銳')
1177	joey6	銳	銳利
1178	koey1	區	區分
1179	koey4	渠	渠道
1180	koey5	距	距離
1181	loey4	雷	雷聲
1182	loey5	呂	呂方
1183	loey6	類	分類
1184	noey5	女	女人
1185	soey1	雖	雖然
1186	soey2	水	汽水
1187	soey3	稅	交稅
1188	soey4	誰	誰是兇手
1189	soey5	緒	情緒
1190	soey6	瑞	瑞典
1191	toey1	推	推理
1192	toey2	腿	大腿
1193	toey3	退	退後
1194	toey4	頹	頹廢

No.	Syllable*	C.	Word/Phrase
1195	tsoey1	吹	吹牛
1196	tsoey2	取	取巧
1197	tsoey3	趣	趣味
1198	tsoey4	除	除夕
1199	doen1	敦	敦煌
1200	doen6	鈍	遲鈍
1201	dzoen1	津	津貼
1202	dzoen2	準	準時
1203	dzoen3	進	進入
1204	dzoen6	盡	盡快
1205	joen6	潤	滋潤
1206	loen2	卵	卵子(原爲'loen5')
1207	loen4	輪	輪盤
1208	loen6	論	辯論
1209	soen1	殉	殉職
1210	soen2	筍	雨後春筍
1211	soen3	信	寫信
1212	soen4	純	純潔
1213	soen6	順	順利
1214	toen5	盾	矛盾
1215	tsoen1	春	春天
1216	tsoen2	蠢	愚蠢
1217	tsoen4	秦	秦國
1218	dzoeng1	張	張揚
1219	dzoeng2	掌	掌握
1220	dzoeng3	帳	蚊帳
1221	dzoeng6	象	大笨象
1222	goeng1	疆	新疆
1223	hoeng1	香	香港
1224	hoeng2	享	享受
1225	hoeng3	向	方向
1226	joeng1	央	中央
1227	joeng4	羊	羔羊
1228	joeng5	養	養生
1229	joeng6	樣	樣子
1230	koeng4	強	強大
1231	koeng5	漾	漾水
1232	loeng2	兩	半斤八兩
1233	loeng4	良	良心
1234	loeng5	兩	兩個人
1235	loeng6	亮	閃亮
1236	noeng4	娘	娘親
1237	soeng1	雙	成雙成對
1238	soeng2	想	思想
1239	soeng3	相	食相
1240	soeng4	常	經常
1241	soeng5	上	上去
1242	soeng6	尚	高尚
1243	tsoeng1	鎗	手鎗
1244	tsoeng2	搶	搶劫
1245	tsoeng3	唱	唱戲
1246	tsoeng4	長	長短
1247	dzoet7	卒	卒仔
1248	loet9	律	律師
1249	soet7	率	率先
1250	soet9	術	出術
1251	tsoet7	出	出術
1252	doek8	啄	啄木鳥
1253	dzoek8	爵	爵士
1254	dzoek9	著	著意
1255	goek8	腳	洗腳
1256	joek8	約	約會
1257	joek9	虐	虐待
1258	koek8	卻	冷卻
1259	loek9	略	戰略
1260	soek8	削	削弱
1261	tsoek8	卓	卓越
1262	fu1	呼	歡呼
1263	fu2	虎	老虎
1264	fu3	富	豐富
1265	fu4	扶	扶持
1266	fu5	婦	婦女
1267	fu6	父	父母
1268	gu1	姑	姑媽
1269	gu2	古	古董
1270	gu3	故	故宮
1271	ku1	箍	箍頸
1272	wu1	烏	烏黑
1273	wu2	糊	芝麻糊
1274	wu3	惡	厭惡
1275	wu4	湖	西湖
1276	wu6	戶	戶口
1277	bui1	杯	茶杯
1278	bui3	貝	寶貝
1279	bui6	焙	焙乾(同'背'書)
1280	fui1	灰	灰色
1281	fui3	悔	後悔
1282	kui2	潰	崩潰
1283	mui4	梅	梅花
1284	mui5	每	每天
1285	mui6	昧	愚昧
1286	pui3	配	配合
1287	pui4	培	培養
1288	pui5	倍	加倍
1289	wui1	偎	依偎
1290	wui2	會	開會
1291	wui4	回	回家
1292	wui5	會	會不會
1293	wui6	匯	匯款
1294	bun1	般	一般

No.	Syllable*	C.	Word/Phrase
1295	bun2	本	根本
1296	bun3	半	一半
1297	bun6	伴	伴侶
1298	fun1	寬	寬鬆
1299	fun2	款	款式
1300	gun1	官	官方
1301	gun2	管	管制
1302	gun3	冠	冠軍
1303	mun4	門	開門
1304	mun5	滿	滿足
1305	mun6	悶	悶局
1306	pun1	潘	潘金蓮
1307	pun3	判	判決
1308	pun4	盤	地盤
1309	pun5	伴	有伴
1310	wun2	碗	洗碗
1311	wun4	援	支援
1312	wun5	浣	浣紗女(不是'碗')
1313	wun6	換	交換
1314	ung2	擁	擁跌人
1315	ung3	甕	引君入甕
1316	bung2	捧	捧書(不是'pung2')
1317	dung1	冬	冬瓜
1318	dung2	董	古董
1319	dung3	凍	凍冰冰
1320	dung6	動	運動
1321	dzung1	宗	宗教
1322	dzung2	總	總管
1323	dzung3	眾	群眾
1324	dzung6	仲	仲裁
1325	fung1	風	風車
1326	fung2	俸	長俸
1327	fung3	諷	諷刺
1328	fung4	馮	馮先生
1329	fung6	奉	奉旨
1330	gung1	公	公平
1331	gung2	鞏	鞏固
1332	gung3	貢	西貢
1333	gung6	共	共通
1334	hung1	空	空間
1335	hung2	孔	孔子
1336	hung3	控	控告
1337	hung4	紅	紅色
1338	jung1	翁	老翁
1339	jung2	擁	擁護
1340	jung4	容	容納
1341	jung5	勇	勇敢
1342	jung6	用	利用
1343	kung4	窮	窮人
1344	lung1	隴	山隴
1345	lung4	龍	九龍
1346	lung5	壟	壟斷
1347	lung6	弄	玩弄
1348	mung2	懵	懵然不知(原爲'mung5')
1349	mung4	蒙	蒙古
1350	mung6	夢	發夢
1351	nung4	農	農業
1352	pung2	捧	捧場(不是'bung2')
1353	pung3	碰	碰碰車
1354	pung4	蓬	蓬鬆(不是'逢')
1355	sung1	鬆	蓬鬆
1356	sung2	聳	危言聳聽
1357	sung3	送	送客
1358	sung4	崇	崇高
1359	tung1	通	普通
1360	tung2	統	統一
1361	tung3	痛	痛苦
1362	tung4	同	相同
1363	tsung1	匆	匆忙
1364	tsung2	寵	寵物
1365	tsung4	蟲	毛蟲
1366	tsung5	重	輕重
1367	but8	砵	衣砵(不是'勃')
1368	but9	勃	興致勃勃
1369	fut8	闊	廣闊
1370	kut8	括	包括
1371	mut8	抹	抹殺
1372	mut9	沒	淹沒
1373	put8	潑	活潑
1374	wut9	活	活潑
1375	uk7	屋	房屋
1376	buk7	卜	占卜
1377	buk9	僕	僕人
1378	duk7	督	港督
1379	duk9	讀	讀書
1380	dzuk7	足	足跡
1381	dzuk9	族	族譜
1382	fuk7	福	有福
1383	fuk9	伏	埋伏
1384	guk7	谷	谷氣
1385	guk9	局	局長
1386	huk7	哭	痛哭
1387	huk9	酷	酷熱(不是'浩')
1388	juk7	旭	旭日
1389	juk9	肉	豬肉
1390	kuk7	曲	歌曲
1391	luk7	碌	忙碌
1392	luk9	六	六月
1393	muk9	木	木板
1394	puk7	仆	仆低

No.	Syllable*	C.	Word/Phrase
1395	suk7	叔	叔父
1396	suk9	屬	屬於
1397	tuk7	禿	光禿禿
1398	tsuk7	促	急促
1399	dzy1	豬	肥豬
1400	dzy2	煮	煮飯
1401	dzy3	注	注意
1402	dzy6	住	住宅
1403	jy1	於	由於
1404	jy2	瘀	好瘀
1405	jy4	如	如果
1406	jy5	雨	落雨
1407	jy6	預	干預
1408	sy1	書	書包
1409	sy2	鼠	老鼠
1410	sy3	恕	饒恕
1411	sy4	殊	特殊
1412	sy6	樹	種樹
1413	tsy2	處	處理(不是'柱')
1414	tsy3	處	辦事處
1415	tsy4	廚	廚房
1416	tsy5	柱	燈柱
1417	dyn1	端	開端
1418	dyn2	短	長短
1419	dyn3	鍛	鍛練
1420	dyn6	段	分段
1421	dzyn1	專	專門
1422	dzyn2	轉	轉變
1423	dzyn3	鑽	鑽探
1424	dzyn6	傳	傳記
1425	gyn1	捐	捐錢
1426	gyn2	卷	春卷
1427	gyn3	眷	家眷
1428	gyn6	倦	倦意
1429	hyn1	圈	圓圈
1430	hyn2	犬	犬隻
1431	hyn3	勸	勸告
1432	jyn1	淵	深淵
1433	jyn2	苑	俊民苑
1434	jyn3	怨	怨言
1435	jyn4	原	原因
1436	jyn5	軟	柔軟
1437	jyn6	願	願望
1438	kyn4	拳	拳擊
1439	lyn2	戀	戀愛(原爲'lyn5')
1440	lyn4	聯	聯合國
1441	lyn6	亂	混亂
1442	nyn5	暖	溫暖
1443	nyn6	嫩	幼嫩
1444	syn1	孫	兒孫
1445	syn2	選	選舉
1446	syn3	算	計算
1447	syn4	旋	旋風
1448	syn5	吮	吮手指
1449	syn6	篆	篆書(不是'算')
1450	tyn4	團	團結
1451	tyn5	斷	斷氣(不是'段')
1452	tsyn1	村	農村
1453	tsyn2	喘	喘氣
1454	tsyn3	寸	尺寸
1455	tsyn4	全	全部
1456	dyt9	奪	奪取
1457	dzyt8	拙	笨拙(不是'絕')
1458	dzyt9	絕	決絕
1459	hyt8	血	鮮血
1460	jyt8	乙	甲乙丙(不是'月')
1461	jyt9	月	月亮
1462	kyt8	決	決定
1463	lyt8	劣	惡劣(不是'lyt9')
1464	syt8	說	說話
1465	tyt8	脫	灑脫
1466	tsyt8	撮	撮合
1467	m̩4	唔	唔該
1468	ng4	蜈	蜈蚣
1469	ng5	五	五福
1470	ng6	誤	誤會

C. - Character

* The tone number of the syllable is written just behind the syllabic structure. Also, the following phonemes are labeled in other forms in this list for simplicity.

original phonemes	ɐ	ɛ	ɔ	ŋ
new labels	A	e	o	ng