A SURVEY OF MANDARIN CHINESE SPEECH RECOGNITION TECHNIQUES

By

YING-CHIEH CHIANG

Bachelor of Education
National Changhua University of Education
Taiwan, Republic of China
1992

Master of Education
University of Central Oklahoma
Edmond, Oklahoma
1998

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2003

A SURVEY OF MANDARIN CHINESE SPEECH

RECOGNITION TECHNIQUES

Thesis Approved:

Thesis Advisor

J Chandler

G. E. Hedrick

Timothy J. Pettibone

Dean of the Graduate College

ACKNOWLEDGMENTS

My experience of academic pursuit and career development is marked by severe challenges and tremendous efforts to win the challenges. My vision started to degenerate when I was 12 years old. At the age of 17, I became totally blind. During this period of time I learned independent life skills, orientation and mobility, listening skills, and Braille. This was the beginning of a life-long battle against my disability. Living with this disability, I started to compete with myself and with people who have normal vision.

Later on, I successfully finished college education, following which I was employed to do research work in the area of visual impairments. From my own learning and working experience, I found that huge barriers existed on the way to success for blind people, but the barriers could be lessened through education and technology. Appropriate education, availability of equipment, and access to technology give blind people opportunities to become productive members of the society and to achieve self-fulfillment. I myself benefited from computers with access devices both in school and at work. Since the adaptive technology and related education were not well developed yet in my country, I decided to come to the United States to pursue learning in these areas.

Resources can effectively deal with the barriers that blind people face from two fields-special education and computer technology. That is why I chose to study in these two fields in the U. S. I received a master's degree in special education in 1998. Now I am going to graduate from a second master's degree in computer science.

At this time I'd like to express my sincerely gratitude to my department and Student Disability Services. They provide necessarily assistants and supports so that I could finish this degree. Thank my advisor Dr. Dai for guiding me doing research. Without my friends I couldn't independently live in the U. S.; thank all of you.

Also I'd like to express my greatest thankfulness to Education Department of Taiwan for 4-year full scholarship, thus, I have this opportunity to pursue my career goal.

PREFACE

It is a dream so far but is believed that might come true soon in the near future. The dream is to develop a computer, which can hear (perceive) and speak (respond) human language. If it happens, many persons with disabilities will benefit from it.

A magnificent building is built from its base. This work will, the same, go from the basis: to investigate the methodologies have been applied in Mandarin Chinese speech recognition. The work will review the papers and summarize the techniques in the different phases of processing recognition, such as speech segmentation, feature extraction, recognizing syllable, and language processing. The most importance is to give more substantial significant concepts on handling Mandarin Chinese speech recognition. Hopefully, that can form a clear configuration of Chinese speech recognition system framework in readers' mind.

Although there exist many theoretic foundations based on complicated statistics and mathematics for recognition modeling algorithms, this paper will not discuss too many statistical and mathematical parts. If interesting, readers can get them from References listed in the last section of the paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Chinese language is a complicated language. Each Chinese character has its own shape (form), sound (syllable), and sense (meaning). These three components of characters play important roles, for providing us with significant semantic sources of interrelated implication, and of understanding much easier upon both spoken and written language. Lacking one of them, especially for character-form, will lead to difficulty of comprehension of the language.

Without form of character, Chinese braille represents each Chinese character by syllables on a format of consonant/vowel/tone (three cells of braille symbols). More extra contextual meanings to be referred are therefore required, otherwise the meaning of words is extremely ambiguous. There are more than 10,000 characters sharing 1,345 syllables, that is, 10000/1345, approximately 7 characters per syllable on average. In other words, blind persons use braille based on pronunciation as their written language of Chinese while sighted persons are using graph-based characters. There is no equivalence between Chinese characters and braille symbols. The problem leads to obstacles in learning, working, and communication between sighted persons and blind persons. These obstacles have not yet been satisfactorily overcome even through computer technologies have been widely applied today. However, it is believed that computer technologies will provide the most feasible solutions for these problems, since speech synthesis and speech recognition have been introduced and the products are gradually becoming mature.

Being a user, it is almost coming true that a blind person can directly interact with a computer only by speech, for example, while using ViaVoice implemented with a screen reader such as JAWS (Job Access with Speech). But most of these well-developed products are based on western alphabetic languages. The first consideration therefore raised is to see what problem issues are in Chinese language for developing up such products. This paper will focus on the issues that are related to the voice dictation of Mandarin Chinese with very large vocabulary, with the following motivations:

(1) Mandarin is the official Chinese language and is widely used in most Chinese communities, although hundreds of dialects exist. Moreover, most available research papers and experimental prototype systems are based on Mandarin Chinese. That means it is likely to collect more information.

(2) In order to be widely used in daily tasks with a computer, the voice dictation should cover daily frequently used words. That would be a very large vocabulary, more than 100,000 words.

Let us take a look on the traditional input methods for Chinese characters. That might be helpful to identify the problems and to see what barriers there are. Keyboard should be originally a good alternative. But limited by Chinese language characteristics, it is not really convenient for a blind person. Basically, characters input by keyboard can be categorized into two modes: the radical symbol input mode and the phonetic symbol input mode. In the case of radicals, there is no regularity although they can be classified into 214 fundamental radicals in a Chinese dictionary. In the aspect of phonetics, Chinese language is a monosyllabic and tonal language, that is, each character has a single syllable with a tone. There exist 37 phonetic symbols and 5 tonal symbols. Regardless of the radical scheme or the phonetic scheme, we cannot obtain one-to-one mapping between 26 English characters and the Chinese on the general keyboard originally designed for alphabetic languages. In the past decade, dozens of products have appeared after many efforts have been made to design different rules for decomposing characters to fewer radicals or to change the mapping layout of Chinese phonetic symbols on keyboards. However, those of radical input still cannot keep away from special training or memorizing cumbersome decomposing rules, and those of phonetic symbol input always require more keystrokes and selecting desired character from a homonym list. Remember, as mentioned above, that a blind person does not have character-form in mind; therefore these two modes are impractical for him/her.

Then, we might ask the following questions as criteria for developing speech recognition system for inputting Chinese characters:

(1) Is the performance better than other Chinese input techniques?

(2) Does it adapt to a new speaker with less training?

(3) Can it correctly recognize isolated syllables or complete sentences without user intervention?

(4) Can it tolerate environmental noise?

(5) Can it give correct and accurate results in the presence of emotional stress and tone changes in a user's utterance?

If all the answers are yes, Chinese speech input method will be a good alternative and solution for blind persons to input Chinese characters.

2. Motivations and Purposes

In the previous chapter, the problems have been identified and their effects have been described. Here we will address the motivation and the purpose.

It is a dream so far but is believed that might come true soon in the near future. The dream is to develop an intelligent computer, which can hear (perceive) and speak (respond) human language. In other words, the computer can be accessed by speech and provide responses by speech whenever messages pop up or any act performed. If it happens, many persons with disabilities will benefit from it.

A magnificent building is built from its base. This work will go from the basis: to investigate the methodologies that have been applied in Mandarin Chinese speech recognition. The work will review the papers and summarize the techniques in the different phases of processing recognition, such as speech segmentation, feature extraction, recognizing syllable, and language processing. The most importance is to give more substantial significant concepts on handling Mandarin Chinese speech recognition. Hopefully, this will form a clear configuration of Chinese speech recognition system framework in readers' mind.

Although there exist many theoretical foundations based on complicated statistics and mathematics for recognition modeling algorithms, this paper will not discuss too many statistical and mathematical parts. Interested readers can get them from References listed in the last section of the paper.

3. Characteristics and Syllable Structure of Mandarin Chinese

Besides some characteristics of Chinese have been mentioned in Chapter 1, we highlight the syllable structure and some grammatical characteristics of the language in this chapter.

3.1 Simple Syllable Structure

Chinese is known as a monosyllabic and tonal language. The syllable structure is very simple that can be denoted as C/M/V/T, where C = 21 consonants, M = 3 medials (glides), V = 13 vowels, and T = 4 lexical tones + 1 neutral tone. The parameters C, M, and V can be null. But, in order to reduce recognition difficulty caused by co-articulation of medials and vowels, it should not use such smaller unit for syllable recognition. Hence, simplify C/M/V/T to C/V/T (tonal-syllable), where V includes M, V, and their combination. Some recognition methodologies disregard tones (Lee, 1997), so it is further simplified as C/V (base-syllable). That can reduce the number of syllables from 1,345 tonal syllables to 408 base syllables.

Some papers use initial-final instead of C/V (consonant-vowel) as a representation of Mandarin Chinese syllable structure. Regardless of the terminologies, initial-final structure or consonant-vowel structure, they are merely roughly to represent two parts of a syllable, not exactly to represent each part of a syllable. So it is important to distinguish when talking about the real consonants and vowels. In addition, it is also important to notice that some vowels in Chinese are with nasal sound ending such as /an/, /ang/, /en/, /eng/. These vowels are quite confusing while recognizing.

According to vocal tract features, these consonant and vowel symbols can be organized as shown in Table 3-1 (the phonetic symbols adopt International Phonetic Alphabet, IPA).

	Consonants	Vowels
Category 1	Null	Null
Category 2	/b/, /p/, /m/, /f/	/a/, /ai/, /au/, /an/, /ang/
Category 3	/d/, /t/, /n/, /l/	/o/, /ou/
Category 4	/g/, /k/, /h/	/e/, /eh/, /ei/, /en/, /eng/, /er/
Category 5	/ji/, /chi/, /shi/	/u/, /ua/, /uo/, /uai/, /uei/, /uan/, /uen/, /uang/, /ueng/
Category 6	/j/, /ch/, /sh/, /r/	/iue/, /iuan/, /iun/, /iung/
Category 7	/tz/, /ts/, /s/	/i/, /iu/, /ia/, /ie/, /iai/, /iau/, /iou/, /ian/, /in/, /iang/, /ing/

Table 3-1 Category of consonant and vowel symbols

3.2 Too Many Homonyms

As mentioned in Chapter 1, more than 10,000 Chinese characters share 1,345 syllables. When someone speaks "two", it is difficult to determine that it is "two" but neither "to" nor "too" without continuing content. Certainly you cannot. That is a linguistic restriction. The situation is much more serious in Chinese spoken language. No wonder it is difficult for a computer to handle a language with 7 characters per syllable on the average. However, when someone speaks "two pens" or "two students go to school," you definitely know whether to choose "two" or "to". It is obvious that with more words given it is easier to determine the correct word by meaning and word-order (phoneme-order). This is an important concept for building a linguistic model.

3.3 No Word Boundary

In English, spacing is an apparent identifier of word boundary. But there is no natural spacing between words in printed and written Chinese sentences. Each Chinese character can be a word and two or more characters also can be from a compound-character word. For example, /du-2 shu-1/ (the ending number stands for tone), you can view it as one word standing for "reading" or two words standing for "read" and "book" respectively. A sentence, in fact, can be looked upon as a sequence of characters as well as a sequence of words. It is quite different from English.

3.4 Open Character and Word Generating

English consists of 26 characters (letters) in uppercase and lowercase. We surely know that no new English character will be generated. But we do not know exactly how many characters there will be in Chinese. New characters are generated to satisfy needs at any time. Furthermore, any two or more characters can be combined as a new word. Collecting characters and words as much as possible is important in this case. Periodic update lexicon and new word add-in function in user interface should be considered.

4. Chinese Speech Recognition System Framework

The design of a system framework will affect system performance and its usage, especially when Chinese spoken language characteristics are used in the application of recognition technique. For example, adopting different feature extraction methods will obtain different feature patterns that represent different characteristics on the input acoustic signal. There are two main approaches used in methodology design for Chinese speech recognition: the template based approach and the consonant-vowel (initial-final) structural based approach (Fu et al., 1995).

Usually, a template based approach is suitable for some sound command systems with fast response time like speech command driven dialing, speech controlling system, and so on. Only limited vocabulary speech recognition system can be achieved by this method. Its implementation is to use one or several templates to represent each candidate syllables. During recognition, match each candidate syllable with a set of templates by various feature extraction techniques and template matching algorithms. Large vocabulary (1,345 syllables) with this approach, large-scale comparison and template searching make such system difficult to give fast response since the system requires at least 1,345 templates.

In the consonant-vowel structural based approach, syllable recognition is based on the characteristics of Chinese phoneme and syllabic structure. Recognition of Chinese syllable can be divided into seven phases: end-point detection, speech segmentation, feature extraction, vowel (final) recognition, consonant (initial) recognition, tone recognition, and Chinese language syntactic and semantic processing (Fu et al., 1995).

The flow of a typical system is shown as follows:

(1) Record input speech data by a microphone.

(2) Implement end-point detection on recorded digital signal to obtain isolated syllables.

(3) Segment the syllable into equal-segments or consonant and vowel part using various algorithms.

(4) Extract the speech features from the equal-segments or the consonant and the vowel.

(5) Perform recognition of base-syllable by the feature sequence.

(6) Perform recognition on tone.

(7) Use recognized consonant, vowel and tone, perform phoneme to character transcriptions by Chinese language processing model.

The flow of a typical system is also showed in Figure 4.1(see page10).

It is important to notice that the consonant and vowel may not be segmented from the syllable explicitly in some approaches such as using time frame. Moreover, consonant-vowel is applied in the modeling of recognition algorithm and training. For example, in the application of Hidden Markov model (HMM) in Golden Mandarin I (Lee et al., 1993), states for consonants and vowels are separately trained and combined. During recognition, input syllable do not need segmentation but obtaining consonant and vowel by searching through the HMM network. However, in some systems with neural network method (Wang et al., 1991), consonant and vowel are explicitly segmented. Recognition

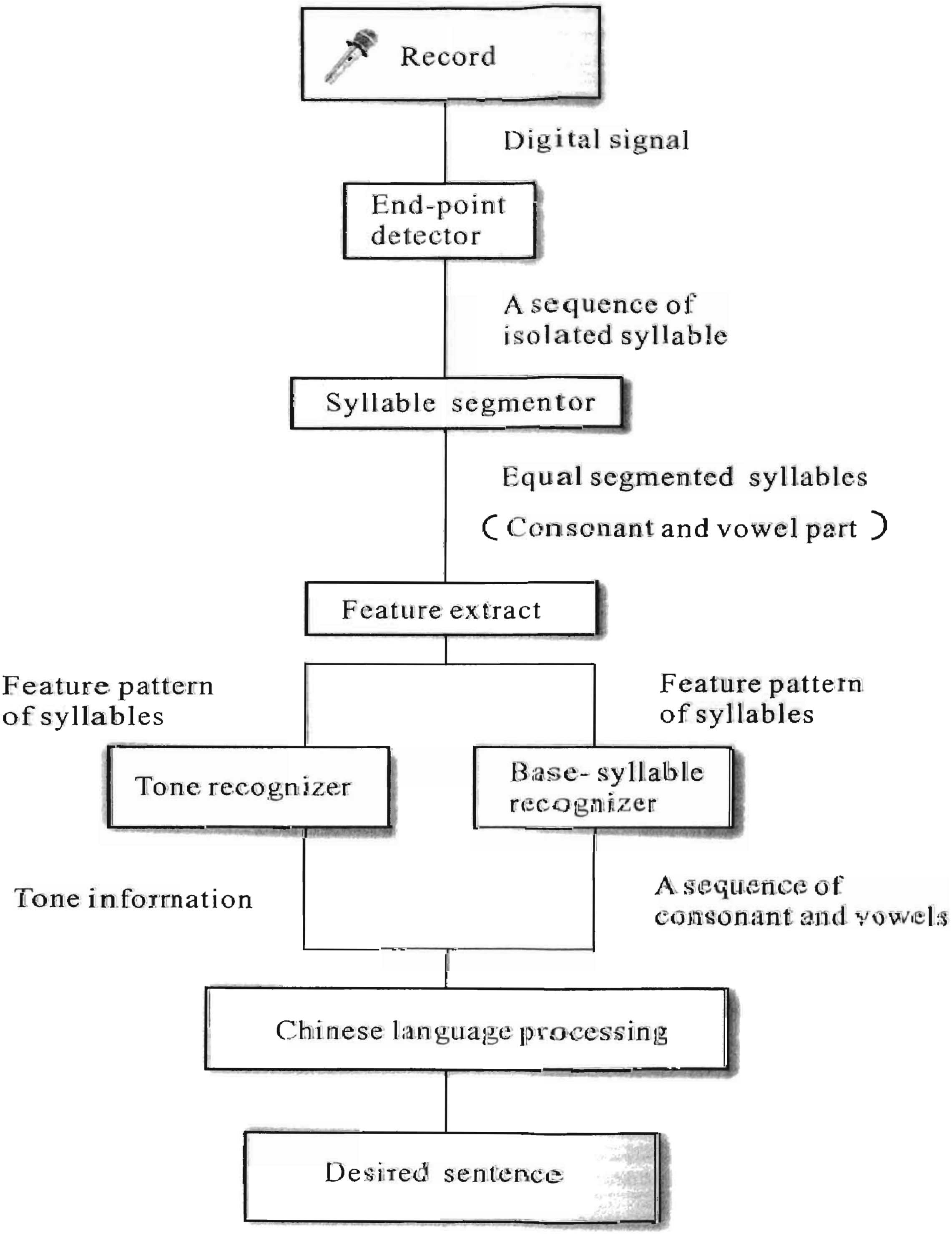

Record
Digital signal
End-point detector
A sequence of isolated syllable
Syllable segmentor
Equal segmented syllables
(Consonant and vowel part)
Feature extract
Feature pattern of syllables
Feature pattern of syllables
Tone recognizer
Base- syllable recognizer
Tone information
A sequence of consonant and vowels
Chinese language processing
Desired sentence

Figure 4.1 The flow of a typical recognition system

of vowels and consonants are organized in a hierarchical manner. In this case, segmentation of consonant and vowel becomes important since a false segmentation point will cause degradation of system performance.

For phase 6 above, it is usually performed concurrently with phase 5. Vowel is usually used in tone recognition since the vowel is the voiced part of a syllable, which contains tone information. The language processing uses some specific properties of linguistics for the most desired character sequence output followed by the previous phases. The language processing is important since there are many homonym characters in Chinese language and anonymous situations in Chinese words and sentences. In complete sentence recognition, the recognized tone will not be very useful since the tone of a character will change in the sentence due to stress and emotional expression. Thus, in some systems, tone recognition is omitted and language processing is employed to deal a sequence of base-syllables with N-gram model to generate correct sentence. Where N-gram can be a number of characters or a number of words, which will be described in Section 5.5.

The following is a real example for Chinese speech recognition system framework. Automatic Speech Recognition and Voice Response (ASRVR) system proposed by (Li and Liu, 1999) has three phases: segmentation, preprocessing and recognition. Speech segmentation process includes Convex Hull method for estimating the minimum number of sub-words S_{min}, *Spectral Variation Function (SVF)* for estimating the maximum number of sub-words S_{max}, and Normal Decomposition for finding the optimal number of sub-words within the preset range S_{opt}. The boundaries of sub-words are located by Level Building Dynamic Programming (LBDP)-based method. The preprocessing is to deal with the variability in speech signals that affect performance of the system. Three feature extraction methods proposed in this phase are based on Mel-frequency cepstral coefficients (MFCCs), relative spectral processing (RASTA), and forward-backward dynamic cepstral coefficients (FBDCCs). Speech recognition is the final phase of the system; the approaches include Navie Bayesian classification, HMM with Viterbi algorithm, and Back-propagation with two-hidden-layered structure. ASVSR system is an experimental prototype for multilingual (Cantonese, English, and Mandarin) performed in different technique combinations.

For the Mandarin Chinese dictation with large vocabulary, in the past decade, Lee's research group has developed Golden Mandarin system I (Lee et al., 1993), II (Lee et al., 1994), and III (Lyu et al., 1995). From these three different versions of the system, we can follow up their improvement and find the system framework changed and different approaches applied in different version.

In their early implementation, Golden Mandarin I, discrete hidden Markov model is used to model the consonant and vowel structure. Later investigation shows that dynamic features of Chinese syllables are not significant in recognition due to the short time duration and relatively stable structure of Chinese vowel. Therefore, the main problem is on the classification of consonant. In Golden Mandarin II, segmental probability model is used. This approach disregards the dynamic property of speech signal and equally segments the syllable. Although the recognition accuracy is not as good as using approaches having dynamic nature, recognition speed is enhanced. Time and training can be saved with this approach in speaker adaptation. In their latest system, Golden Mandarin III, the modeling of consonant-vowel structure is further enhanced using rightcontext-dependent phoneme-like-units (RCD-PLU). Continuous hidden Markov Model with 2 states and 3 Gaussian mixtures per state is used in the modeling. It should be noticed that RCD-PLU emphasize on intra-syllabic co-articulation between consonant and vowel. It is matched with a technique using consonant-vowel transient features that is used in template-based approach.

In real-life application, a successful system like Golden Mandarin III is approaching the stage of useful product. With such a system, speech input can be used as a solution to input Chinese characters to computer system.

5. Techniques for Recognition of Mandarin Chinese

In Chinese speech recognition, syllable recognition is important because it will affect the recognition accuracy rate later on processing the conversion from syllables to characters. In other words, the recognized sequence of syllables will become the input data of language processing. The recognition process can generally be divided into end-point detection, feature extraction, pattern recognition and language processing. End-point detection is used to locate the starting and ending point of the input speech data. This process locates the voiced speech part and skips the non-speech part, which can minimize the processing time. Recognition accuracy is also increased since the test pattern is not confused with the non-speech part. Feature extraction is to find a good representation form for given speech data. Speech data are transformed into a more usable and compressed form for further processing.

Linear prediction analysis followed by vector quantization (VQ) is widely used in various systems. The speech data are segmented into small time frames and are pre-emphasized by windowing. After linear prediction analysis, the result of linear prediction coefficients will be transformed into cepstrual coefficients that are the logarithm of inverse fast Fourier transform. The cepstrual coefficients form a feature vector that is used as a representation for a small time frame. In this way, input speech signals are transformed into a sequence of feature vectors. VQ is the process to search for a similar feature vector to represent the input feature vector. A vector codebook is constructed in advance, during training, by collecting and processing sufficient number of feature vectors. The codeword index of that similar vector is used as a pattern for pattern recognition. After VQ, the sequence of feature vectors is transformed into a sequence of indices. Various algorithms and methodologies are used to recognize the generated sequence to obtain a sequence of syllables. The recognized syllables will be dealt with language processing to solve the ambiguity and homonym words. The desired sentences will finally be produced.

As mentioned in Chapter 4, two approaches, template-based and consonant-vowel-based, are used in the recognition of Chinese syllables. They are discussed in Section 5.1 and Section 5.2, respectively.

Mandarin Chinese is a tonal language. Tone recognition is one of the issues in Chinese speech recognition. The general approach is to consider the pitch contour in the vowel part of the syllable. However, tone recognition is only useful in isolated syllable dictation. In complete sentence recognition, tone information becomes less informative and tone recognition is mainly performed by Chinese language processing model which transcript phoneme sequence to complete sentence. More discussion is in Section 5.3.

Speech variation is a factor that has an effect on recognition accuracy. Some approaches can solve such problem discussed in Section 5.4.

Chinese language processing is used in sentence hypothesis and phoneme to character transcription. Various methods, Markov language model and statistical approaches, are applied. However, different methodologies give preferentiality to different corpus and environments. For this issue, some standards in construction of corpus are worth investigating. Those are discussed in Section 5.5.

5.1 Template Based Approaches

Template-based systems are often used at the beginning stage of Chinese speech recognition research as a pioneering experiment on the usability of speech recognition of Chinese syllables. It recognizes Chinese syllable by template matching. Different techniques and methodologies are developed with different kinds of speech signal representations in reference templates. Some approaches adopt the dynamic properties of speech signals. Others employ the transient part between consonant and vowel. Different distortion measurements are used to match the input syllable and templates. The template that is closest to the input syllable is considered as the same. A limited (small) vocabulary system using this approach can achieve a high performance with fast response time. Template-based approaches such as dynamic programming, feature matrix, neural network and probabilistic approaches are introduced and summarized in Fu et al.'s (1995) paper. Interested readers can find more detailed information from their paper.

5.2 Syllable Recognition

Consonant-vowel structural based system uses the consonant and vowel characteristic of Mandarin Chinese language. A system using this method models input syllable as a concatenation of consonant and vowel. Training and recognition will be based on the structures discussed in Section 3.1. However, using this approach does not imply that the input syllable will be segmented into two parts explicitly. It depends on how the system is designed and how the recognition techniques are applied. Using consonant-vowel structure modeling, the whole set of syllables can be recognized by identifying the parts of consonants and vowels. Together with tone recognition (some systems without tone recognition but constructing a lexical lattice based on recognized base-syllables) and

Chinese language processing, continuous speech recognition with very large vocabulary can be achieved.

VQ can be applied further to reduce the number of bits needed to represent a speech unit. It is a technique for the selecting from a subset of possible combinations of parameter values. It requires less bits per frame than independent coding of the parameters, but it can be computationally expensive to obtain the maximum benefit that is, in principle, possible with this technique.

The multi-dimensional feature space for any practical method of speech signal analysis is not uniformly occupied. The types of spectrum cross-section that occur in speech signals cause certain regions of the feature space, for example, corresponding to the spectra of commonly occurring vowels and fricatives, to be highly used, and other regions to be at most sparsely occupied. It is possible to make a useful approximation to the feature vectors that actually occur by choosing only a small subset of feature vectors, and replacing each measured vector by the one in the subset that is nearest to it according to a suitable distance metric. This process is VQ.

For example, 50 bits might be used in linear predictive coding (LPC) vocoders to send 10 coefficients in scalar quantization, but that same information could be coded with perhaps 10 bits in a VQ scheme. A properly chosen set of 1,024 spectra (2^{10} for a 10-bit VQ) should be able to adequately represent all possible speech sound. Obviously, 9 bits is sufficient for representing all Mandarin Chinese base-syllables (408).

For a system that uses the characteristics of consonant and vowel, recognition of consonants and vowels is the very important part. In the early publications, several authors proposed methodologies that consonants and vowels are recognized individually, that is, a syllable is first segmented in two parts and then recognized separately.

Consonants are composed of voiced and unvoiced consonants and nasals, which make consonants more confusing when recognizing alone. Using vowel state VQ on classification of consonants, the consonant part is first segmented from the syllable and classified into voiced and unvoiced by extracting features like averaged energy, syllable duration and zero-crossing rate (Suppose we have sampled a sine wave 2,000 times per second, so we have samples at 0.5 ms, 1.0 ms, 1.5, ms, etc. To keep it simple, we may assume, without loss of generality that the samples are taken at the times where the sine wave is zero called zero-crossing). Then, using these feature vectors sets up a feature codebook. Recognition is done by finite-state VQ. In such system, the vowel part is known in advance. Consonant classification is completed when the recognized vowel group is reached.

Another consonant classification systems that use statistical models are constructed based on the acoustic features of the 16 channels filter-bank output. For each filter bank channel, consonant duration, average intensity, the first moment and the second central moment of intensity in frequency domain, and the relative intensities in the specified frequency bands are recorded as acoustic features. For an unknown consonant, recognition is finished when found the probability that the unknown consonant fit all candidate consonant models. The model with the highest probability is chosen as the correct consonant.

Statistical Markov model is adopted by another consonant classification system. Three different HMM are experimented to recognize the consonants. The first one is the

HMM using discrete observation with a VQ preprocessor. The other two are mixture autoregressive HMM and multivariate Gaussian HMM. These show that HMM is efficient for recognition of Mandarin consonants (Fu et al., 1995).

For the recognition of vowel parts, there are three approaches using for recognizing Mandarin Chinese vowels--they are the Segmental Model Approach, Three-pass Approach and Multi-section VQ. The Segmental Model Approach is based on different segmental model for different type of vowels. The vowel part is segmented into several parts according to the transitivity of speech signals. Each segmented part is relatively stable in nature and limited by the transient part at each end. The vowel is recognized from recognition of the pattern and characteristic of each segment. The three-pass approach has three stages in the recognition procedure. The first stage is to screen for pure vowel in the vowel part. The second stage is to classify vowels in small confusing groups. The third stage is to match for each vowel by VQ. Multi-section VQ is used with branch-and-bound classification to recognize the vowels. Multi-section VQ is the most capable one in terms of relatively high recognition rate and low computation utilization.

Many recognition systems use Hidden Markov Model or modified HMM. HMM is a 5-tuple (Q, V, π, A, B), where:

Q is the set of states, $\{q_1, q_2, ..., q_n\}$,

V is the output alphabet set, $\{v_1, v_2, ..., v_m\}$,

$\pi(i)$ is the probability of being in state q_i at time t = 0,

A is an nxn matrix $\{a_{ij}\}_{nxn}$ transition probabilities, where a_{ij} = Pr[entering state q_j at time t+1 | in state q_i at time t] (independent of t), and

B is an nxm matrix $\{b_j(k)\}_{nxm}$ output probabilities, where $b_j(k)$ = Pr[producing v_k at time t | in state q_j at time t] (independent of t).

An HMM is a two-stage probabilistic process that can be used as a powerful representation for speech. It is a well-behaved mathematical construct and a number of detailed algorithms exist for solving problems associated with HMMs. Introductions to these can be found in (Rabiner, 1989). The following will give some examples of HMMs and describe how they can be applied to speech recognition.

An HMM consists of a number of internal states, and the model passes from an initial state to the final state as a step-by-step process generating an observable output at each step (state transition). For example, the states may correspond to phonemes contained in a word with the observable output corresponding to the presence of absence of a number of acoustic features. At each step, the model can either move to a new state or stay in the current one. The model is "hidden" in that we cannot observe the state directly but only its output. From the sequence of observable outputs, we can attempt to guess when the model was in each state. On the other hand, we can say whether some sequence of outputs was likely to have been generated by a particular HMM.

An HMM can be used to represent a syllable with internal states representing characteristic acoustic segments, possibly phonemes or allophones. The output of a state is a frame or vector of acoustic parameters or features. This output is probabilistic to allow for variability in pronunciation and hence differences in the acoustic representation. The duration of an acoustic segment is a function of the number of steps in which the model is in the state corresponding to the segment. Staying in the same state, that is, lengthening a phone, depends on the probability associated with the transition from that state to itself such as P_{11} in Figure 5.1 (see page 24). Arcs such as P_{13} may be included to indicate that an intermediate state S_2 is optional during pronunciation, such as the second phoneme in /ji-i-ang/.

HMMs are a powerful representation of speech and consequently are used in many speech recognition systems currently under development. Although the description of following examples is for isolated syllable recognition, HMMs can also be used with continuous speech to represent not only the phones of each syllable but also the probabilities of transitioning from one syllable to another. For continuous speech recognition, the observed sequence would be generated by passing through a sequence of state corresponds to an entire syllable in that way encoding syntactic and semantic information (refer to the description in Section 5.5).

Consider two boxes full of cards. On each card one consonant or vowel phonetic alphabet is printed from among /b/, /f/, /h/, /ing/, /e/, /i/, etc., and there are equal numbers of each. If you pick one card from each box you have a certain probability of obtaining a Mandarin Chinese syllable, say /bing/ or /he/. Of course, unacceptable syllables, such as /bf/, are also possible. All combinations are equally likely.

We can train these boxes to prefer a certain syllable and certain pronunciations of that syllable. Call the boxes A1 and A2. Whenever we say /bing/ we discard one card of every alphabet but /b/ from A1 and one card of every alphabet but /ing/ from A2. Other syllables have no effect. We repeat this process many times.

After the training stage of discarding cards, if we randomly select a card for observation (not discarding) from each box, the probability of other alphabet combinations. We can also say that these two boxes, when used as indicated, are a model of the syllable /bing/.

Suppose that we pronounce /bing/ in two ways, /bing/ and /biing/, where the doubled phonetic /i/ indicates greater (lengthening) vowel duration. Recall that between two slashes are phonetic alphabet spellings from Table 3.1 (Chapter 3). Further suppose that we pronounce each variant half the time when training the boxes. To model the two pronunciations best, we must make an extra notation on box A1 or A2 indicating that when we are selecting cards for observation, half the time we should examine two cards from A1 (/b/ and /i/) or A2 (/i/ and /ing/) instead of one. Furthermore, suppose that about one time in ten we stutter on /b/ of /bing/, pronouncing it /b-bing/. This tells us to adjust the model so that we observe two cards from A1 every tenth time, on average. We illustrate the model in Figure 5.2 (see page 24).

In the terminology of speech recognition, the boxes are called states; the arrows connecting them are called transitions; and the cards that are selected for observation are called outputs. The transition arrows indicate the probability of going to whatever state they're pointing at. When in state A1, after observing a card, you proceed to state A2 nine times out of ten because the probability of that state transition is 0.9. One time in ten you remain in state A1 and observe another card in state A2. After observing a card from A2 you must stop.

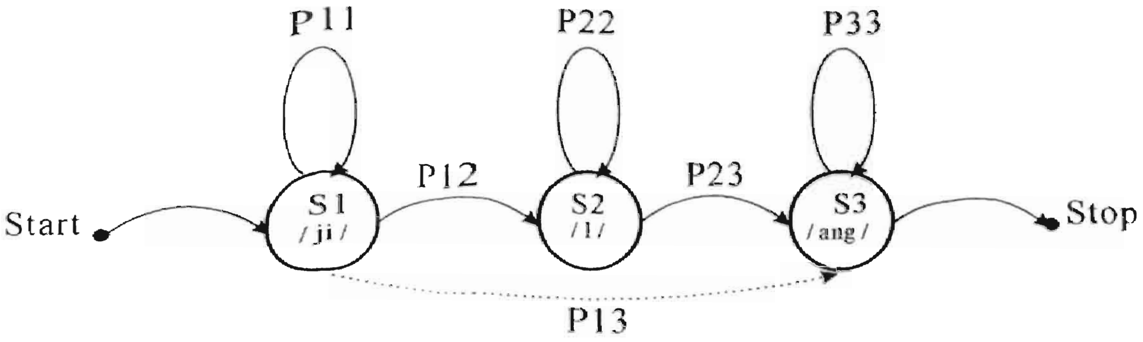
Although the boxes prefer /bing/, it is a probabilistic matter and other combinations are possible. This is a two-state probabilistic process in that there are probabilities associated with card observation and probabilities associated with box selection.

We can repeat this process with more boxes. For example, we can train the syllable /he/ just as we trained /bing/, associating /he/ with the boxes B1 and B2. To account for variant pronunciations of /he/, /hee/, /heee/, etc., we can adjust the state transition probabilities leaving B2 to make remaining in B2 and possibly adding an /e/ more likely.

The A-boxes now model /bing/, and the B-boxes model /he/. Let us combine the models into an elementary syllable recognizer, using a pattern matching, statistical technique. Suppose that we have an unknown syllable, X. We randomly select and observe cards from the A-boxes as described above, repeating the process a very large number of times. This corresponds to the statistical, probabilistic nature of the process. We record the percentage of time that X is matched by the observed cards. We repeat the process using the B-boxes, again noting the percent of matches.

If the matching percentage from the A-boxes is sufficiently greater than that from the B-boxes and exceeds some threshold above zero, then we may conclude that X is /bing/. Conversely, a high percentage of matches from the B-boxes indicate that X is /he/. If both percentages are below the threshold, we may conclude that X is neither /bing/ nor /he/. We also can train a bi-syllable model showed in Figure 5.2, C-boxes, for /bing-1 he-2/ (glacier). Box C1 contains /bing-1/, /bing-3/, and /bing-4/. Box C2 contains /he-1/, /he-2/, and /he-4/. Follow the same training processing like the above training for A-boxes and B-boxes. In C-boxes, only /bing-1 he-2/ can be highly acceptable in Chinese lexicon. Apparently, this is different level of HMM in Mandarin Chinese recognition and in this way that continuous lexical processing can be long-drawn-out.

This system is, in essence, a simple hidden Markov model. It is "Markov" because that is the name given to state transition models in which the next state is determined



(1)start ► s1 ► s2 ► s3 ►stop ► / jiiang /
(2)start ► s1 ► s2 ► stop ► / jiang /

Figure 5.1 A Markov Model for / jiiang / and /jiang / (second phoneme omitted)

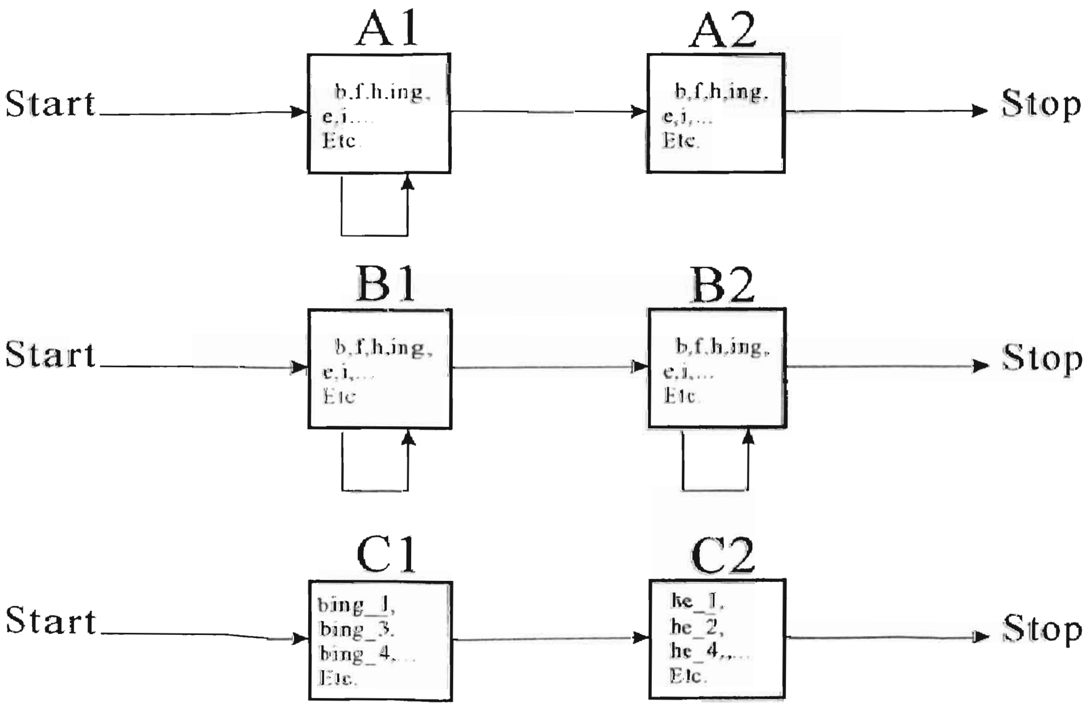


Figure 5.2 Three models for training / bing /, / he /,
and /bing_1 he_2 /, respectively

solely from the current state. It is "hidden" because the actual state sequences are concealed from us. For example in observing cards, it is possible that /b-i-ing/ comes from the state sequence A1-A1-A2, though it is more likely to come from A1-A2-A2. All we know is that the output is /bing/ and we can only infer probabilistically which set of boxes corresponds to the input.

In using HMMs for actual speech recognition, each "card" corresponds to a vector of acoustic parameters. These may be derived from LPC or cepstral coefficients; they may be codebook entries from a VQ process; or they may be something else. Each vector has probability associated with it, analogous to the different distributions of phonetic alphabets in boxes. The probabilistic nature of these parameters helps account for the variability in pronunciation in a way analogous to how we did it with spelling (/bing/ vs. /biing/ vs. /bbiing/). The state transition probabilities accomplish the effect of time normalization by allowing the system to remain in the same state, corresponding to a longer-duration pronunciation of some speech unit. It is also possible for states to be skipped, thus accounting for pronunciations in which sounds are omitted, such as the second /i/ in /jiiang/.

HMMs are the most common type of reference model in use today. Their two-pronged probabilistic nature makes them extremely effective for representing speech, since allophonic distribution is statistical in nature. From a mathematical point of view, HMMs are well-behaved and thoroughly understood. Computationally, many algorithms have been devised for carrying out both training and recognition as efficiently as the mathematics allows. Because the acoustic outputs are probabilistic, retraining an HMM to the changing habits of one speaker, or to entirely new speakers, is feasible.

To a certain extent, the whole syllable is segmented in time frames and the recognition is to search through HMM to find the best path. Since the HMM topology is constructed using consonant-vowel structure, the best path gives the resulting combination of consonant and vowel pair.

Golden Mandarin I is the first system that uses discrete HMM. The system makes use of the consonant-vowel structure of Mandarin Chinese in the modeling of HMM topology where the consonant and vowel are modeled by states separately, as shown in the above examples. The recognition rate is not satisfied in that system. Therefore, they developed several training approach in HMM training. As Mandarin Chinese syllable can be divided into consonant and vowel, all syllables are segmented into two parts. In their two-pass training approach, the 38 vowels are trained by 38 HMMs. And 408 HMMs are used to train the 408 consonant parts of the 408 syllables which are composed from all possible concatenations of consonants and vowels. The 408 consonant HMMs are concatenated with the 38 vowel HMMs to form 408 syllable HMMs where the last state of consonant HMM is the first state of vowel HMM. This process is to model the transient region between consonant and vowel. Thus, the difference in the consonants and the similarity in the vowels can be emphasized. In order to minimize the errors in the HMMs due to insufficient training data, an improved two-pass training is used. Since some vowels using the same consonants shared the same consonant HMMs, the total number of consonant HMMs is reduced from 408 to 99. These 99 consonant HMMs are trained with more data. Then, the 408 syllable HMMs formed from the 99 consonant HMMs and 38 vowel HMMs are more robust but less accurate. Later, three-pass training is applied. The trained 99 consonants are reassigned to 408 consonant HMMs. These 408 consonant

HMMs are re-trained by the consonant data. The resulting consonant HMMs are concatenated with the 38 vowel HMMs to form the 408 syllable HMMs. This training approach improves the recognition rate (Lee, 1997).

Segmental probability model (SPM) is proposed in Golden Mandarin II, which is found to be very suitable for recognition of isolated Mandarin syllables especially considering the simple monosyllabic structure of Chinese language (Lee et al., 1993; Lee et al., 1994; Lyu et al., 1994; Shen et al., 1994; Shen and Lee, 1996). It is very similar to continuous density HMM (CHMM) but the state-transition probabilities are deleted and the syllable utterance is divided into *N* segments (states) with equal length. The stochastic state transition behavior in CHMM is replaced by a deterministic process, while the stochastic observation behavior remains unchanged, represented by Gaussian mixtures. The proposal of segmental probability model is to ease the intensive state path searching in HMM. A concatenated syllable matching algorithm is used instead of the conventional Viterbi search algorithm (Aravind, 1999; Shen, 1998). The primary problem of recognizing Mandarin is to distinguish the confusing consonants, the state transition probabilities of HMM is considered not important.

In training and recognition stage, syllable is divided into N segments corresponding to the N state in SPM. During training, the N feature vectors are used to train the N covariance matrix for the N states. In recognition, observation probability of an unknown syllable is produced by each SPM. The SPM producing the highest probability is the desired output.

Two-dimensional cepstrum (TDC) analysis is used as a feature representation of syllables for isolated syllable recognition. An extended system using this approach is proposed to solve the problem of complete Mandarin vocabulary recognition, to the system, TDC is still used as a feature representation scheme. However, TDC is calculated for consonant and vowel separately. A syllable is first segmented into a consonant and vowel. A small portion of TDC is used as a feature vector for consonant and vowel. In the training process, a model of Gaussian mixture distribution is generated for each consonant and vowel. During recognition, the vowel is recognized first and the top three vowel models giving the highest probabilities are selected. The consonant is recognized by the consonant model, which can be connected to the candidate vowels. The highest probability from the consonant-vowel pair is chosen as the recognized syllable.

5.3 Tone Recognition

Mandarin Chinese Language is a tonal language. There are basically four lexical tones, that is, the high-level tone (Tone 1), the mid-rising tone (Tone 2), the mid-falling-rising tone (Tone 3), the high-falling tone (Tone 4), and one neutral tone (Tone 5). The primary difference among the tones is in the pitch contours, there exist standard patterns for the pitch contours for the four lexical tones but not for the neutral tone, and the pitch contours are essentially independent of the vocal tract shape or parameters of the syllables. Neural tone is different from lexical tone with lower energy and short time duration. Tone recognition plays an important role in Chinese syllable recognition. Since different characters can share the same consonant and vowel but with different tones, tone recognition is necessary to give a correct result in isolated-syllable recognition.

Before early 1990s, a popular method for recognizing tone languages is the two-step method--first, to recognize the base syllable by its consonants and vowels; second, to recognize the tone of the syllable by classifying the pitch contour of that syllable using discriminative rules. The recognition of tonal syllables is a combination of the recognition of base syllables and the recognition of tones. The above method works well in isolated-syllable speech recognition. But it shows difficulties in handling continuous speech. Later, the one-step method recognizes vowels and consonants and tones in a single step are proposed for the recognition of continuous speech tonal languages. Basic acoustic units with different tones are treated as different phonemes since tone is a property of syllable.

Basically recognition of tone is first to segment a syllable into a sequence of small time frames. Pitch information is then extracted from each time frame. It is assumed that pitch information is stable within each time frame. Tone recognition is implemented by recognizing pitch information sequence patterns. Different methodologies have been applied on tone recognition.

Demi-Syllable Approach

This approach is to decompose each Chinese syllable to a consonant and a vowel. For example, in Mandarin, /tan-3/ can be decomposed to /t/ and /an-3/, /tian-3/ to /t/ and /ian-3/, /dan-3/ to /d/ and /an-3/, /dian-3/ to /d/ and /ian-3/, etc. Because the vowel parts contain tone information but consonants do not, the decomposition let only treat tonal-vowels for recognizing tones. The number of recognized units is reduced.

By grouping the medial (glide) with the consonant to become a preme. For example, separate /bian-3/ to /bi/ and /an-3/ and /dian-3/ to /di/ and /an-3/. The total number of phonemes using can be further reduced (Chen et al., 2001).

Main Vowel Approach

According to the assumption that the pitch information on the main vowel is sufficient to determine the tone of the whole syllable, the examples in the above section of Demi-syllable approach, the main vowel only remains /a/ with tone 3. In Mandarin Chinese the vowel without nasal sound ending is only nine. The total number of phonemes is greatly reduced (Chen et al., 2001).

Center-Clipping Autocorrelation Method

Center-clipping autocorrelation method is proposed for Chinese four-tone recognition. The input syllable is center-clipped and the resulting clipped signal shows the pitch period. Then, analyze the pitch period to form a two-dimensional vector by using data selection, error correction, data smoothing and curve fitting. Statistics for this two-dimensional vector is collected and the four tone decision is performed by examining the values of the two-dimensional vector (Guan and Chen, 1993).

5.4 Speech Signal Variability Handling

Noisy environment, meaningless sound, signal distortion, and so on usually degrade seriously on performance of speech recognition. They are illustrated as following:

- Background and channel noise.
- Electrical noise from different microphones or other recording devices.
- Meaningless sounds (a sneeze) or filler words ("uh" or "um") between words.
- Different speaking rate, mood and styles of speakers.

Taking the speech input and converting it into some types of parametric representation for increasing recognition accuracy rate can deal these problems. The Relative Spectral Processing (RASTA) (Li and Liu, 1999) is a noise reduction method that can extract noisy-robust features from speaker-independent, continuous speech recognition systems. In order for spectral smoothing and enhancement of spectral peaks, it uses some moderate forms of automatic gain control.

A segment-based C_0 (the zero-th order of cepstral coefficient) adaptation (SCA) scheme (Hong and Chen, 1999), for parallel model combination (PMC)-based Mandarin Chinese speech recognition, incorporates a new C_0 model of speech signal into the PMC method to improve the gain matching between the clean-speech HMM model and current noise model. The SCA-PMC method composes two phases—training and testing. During the training phase, the task is to construct the C_0 model by mutually modeling the normalized C_0 with other MFCC recognition features to form C_0 normalized HMM model. During the testing phase, the input utterance is pre-segmented into segmented-syllable. The task is to perform C_0-denormalization operations to expand the C_0-normalized HMM model, and uses them in the PMC method. Because the use of more precise gain matching in the PMC, the consequence of noise compensation can be achieved. The recognition accuracy rate of base-syllable is significantly increased for continuous noisy Mandarin Chinese speech recognition.

The advantages of SCA-PMC method are shown as below:

(1) It can track both local phonemic loudness variation and the global intonation loudness variation. This makes it has better noise compensation effect. This also makes insensitive to the volume adjustment of the recording device.

(2) It can take C_0 as an additional recognition feature to assist in the recognition.

(3) The gain matching between the speech models and the noise models, required in the PMC model combination, is implicitly achieved by the proposed C_0 adaptation scheme.

(4) The de-normalization factor maximum C_0 is always estimated from a frame with high signal-to-noise ratio (SNR). This makes it be a reliable estimated.

5.5 Chinese Language Processing

In speech recognition, an acoustic processing is used to recognize a sequence of input speech signals and gives a sequence of consonants, vowels, and tones. The sequence of consonants, vowels, and tones then forms a word lattice that is the candidates of words sharing the same syllables. If the acoustic recognition is based on base-syllable and disregards tones, it has to produce a lattice of tonal-syllable candidates and this tonal-syllable lattice will be used to construct a word lattice. The construction of word lattice is implemented by a matching process between all possible paths in the tonal-syllable lattice with word selecting from a lexicon that stores a large number of words. In order to make such a matching process efficient for a large vocabulary dictation task, the words in the lexicon are usually stored in a tree structure. Such a lexicon tree can also be re-organized into a backward tree structure so that a forward-backward searching algorithm can be used for word matching in the lexicon. In such case, the backward lexicon tree is helpful for linguistic decoding. A language-processing model is used to find out a most promising sentence hypothesis for the given lattice. In Chinese language processing, sentence hypothesis includes fast phoneme to syllable conversion, syllable searching and homonym characters solving.

The Markov Chinese language model is used to solve phoneme character decoding problem. The probability of generating a Chinese sentence is the product of the successive state transition probabilities where the state can be one or more character or words (a word is compose of several character). The state transition probabilities are trained by a large amount of text. In usage, with the input of syllable sequence, assuming a sentence consists of a sequence of words $W = w_1 w_2 w_3 \ldots w_n$, its probability can be decomposed into a conditional form as following:

$$P(w_1\ w_2 \ldots w_n) = P(w_1)\ P(w_2 \mid w_1) \ldots P(w_n \mid w_{n-N}\ w_{n-N-1} \ldots w_{n-1})$$

The $P(w_n \mid w_{n-N}\ w_{n-N-1} \ldots w_{n-1})$ is called an N-gram probability.

Suppose that an N-1-order Markov process producing a sequence of words, the probability of the i-th word w_i given in a conditional form depending on the last N-1 words as:

$$P(w_i \mid w_1\ w_2 \ldots w_{i-1}) = P(w_i \mid w_{i-(N-1)} \ldots w_{i-2}\ w_{i-1})$$

The smoothing method of bi-gram is to use the combination of bi-gram probabilities and uni-gram probabilities. The smoothing method of tri-gram is based on the weighted sum of tri-gram, bi-gram, uni-gram, and zero-gram probabilities. The optimal weights are obtained by the Baum-Welch citation.

In the case of transformation from syllables into Chinese characters, every syllable has several candidate characters, consisting of all nodes of multi-level graph. There are adjacent probabilities between two words, which are weighted between two nodes. Thus, a multi-level graph of transforming syllables to Chinese characters is built. The aim is to search a best path of maximum probability in the graph. Viterbi algorithm is used in the transformation, which suitable to search fast for a best path from a weighted multi-level graph.

By considering and observing the characteristics of Chinese word structure, bi-character words occupy more than 70% of the most frequently used top 50,000 words in Mandarin Chinese. Most word classifications are based on bi-character word because a word formed with the more character compound will decrease the occurrence of homonym word but will increase the collection of words from larger text corpus. Since a sentence can be regarded as a sequence of characters or a sequence of words and every character can be a morpheme in a word with its own meaning and linguistic feature, this leads to the concept of Chinese language modeling based on characters rather than words. Actually, the character-based models provide some information existing in the word-based models and some additional information. The total number of commonly used characters in the character-based models is much smaller than that of commonly used words, so the number of bi-gram (character), tri-gram, or similar parameters will be much smaller for character-based models. Therefore, these parameters can be estimated with better accuracy as compared to those for word-based models if the training corpus is limited in reality. The smaller number of parameters makes it possible to obtain model parameters with higher order, such as N-grams, and makes storage, retrieval, and implementation of such language models easier.

Since the segmentation of a sentence into words is not only one of its kinds, training word-based language models requires a large enough training text corpus that is the same segmented into words every time. Because the word-based approach is certainly difficult, the character-based language models are straightforward for the problem of consistently segmenting the training corpus into words due to that can be directly bypassed. However, a word bi-gram (two words composed) is more effective than a character bi-gram (two characters composed), a word tri-gram (three words composed) is more effective than a character tri-gram (three characters composed) and so forth. So that successful interpolation of the character-based and word-based language models is certainly better. The reason is that in Chinese every word is composed of one to several characters that also have their own meaning. For example, the word bi-gram probability (assume each of these two words has two characters) for a word appearing after another is very close to the probability for these four characters appearing in a sequence (a character 4-gram). Each of these four characters has its own meaning and this word bi-gram provides stronger linguistic constraints and more specific meaning.

The word class is a more useful basis for Chinese language modeling, which is to group different words together with similar linguistic properties as a class. Thus, a very large number of words can be categorized into a quite small number of word classes and training of higher-order language models becomes possible.

An example of simple word classification technique, how to group the words into appropriate word classes on which the language models can be constructed, is according to the word's starting and ending characters. For example, the words starting with the character /tian-1/ (day) such as /tian-1 kueng-1/ (sky), /tian-1 chii-4/ (weather), etc., can be in the same class; the words ending with the character /tian-1/ (day) such as /ming-2 tian-1/ (tomorrow), /tzuo-2 tian-1/ (yesterday), etc., can be in the same class. In this way, every word generally belongs to two word classes, so the total number of word classes will be the total number of characters times two. The categorization of this technique is very simple so that any new word added to the lexicon can be automatically categorized into its corresponding class without any problem. However, not all words having the same starting or ending characters always have identical linguistic properties. Therefore, some words (a small portion) may be inappropriately assigned in the same class with the same starting character.

For the different classifications and making them much more useful, large text corpus must be constructed by collecting large amount of text from different sources such as books, magazines, newspapers, etc. For different usages, the terminologies from different fields such as computer science, medical science, psychology, etc. also must be collected and categorized in some reasonable and efficient ways for further application.

6. Conclusion and Suggestion

From the above chapters, it can be seen that many efforts have been made for developing Mandarin Chinese speech recognition by different research groups in Chinese communities. Various approaches have been proposed for improving recognizing accuracy rate and system performance. Some papers furthermore show that combining different approaches in a system can enhance the accuracy rate of recognition and speed up computation. It is encouraging that Golden Mandarin (III) has been used in personal computer with Windows system. Automatic Speech Recognition and Voice Response (ASRVR) is an ideal system for multilingual users if its performance could be greatly improved.

Some suggestions are given as following:

(1) Create some specific speech grammar for training and correcting interface. For example, we can say "/bing-1 he-2 de-5 bing-1/" (in English, it is something like we say "b of boy") that associated word must be no confusing. The system can be trained by our pronunciation of /bing-1 he-2/ and the system will select the exact character /bing-1/ (stands for ice) from the word /bing-1 he-2/ (glacier) in the lexicon, since the higher-order word becomes less ambiguous. In this manner, it is capable of solving the ambiguous characters such as characters in a person's name.

(2) Although there are many homonyms in Mandarin Chinese, there are approximately 170 syllables sharing only one character. These characters such

as /bai-1/ (white), /pau-3/ (run), /bei-3/ (north), etc. are quite distinguishable from those homonyms. Using these characters for word classification basis and selecting them as the highest priority for language processing might greatly reduce the ambiguity.

(3) While the Mandarin Chinese speech recognition has gradually matured, hope that some developers can integrate it with Mandarin Chinese speech synthesis and create a friendly interface to benefit those persons with disabilities in their education and vocation.

REFERENCES

1. Aravind, Ganapathiraju (1999). "Implementation of Viterbi Search Algorithm." Proceedings of the IEEE Southeast Conference, Lexington, Kentucky, USA, pp. 32-35.

2. Byrne, W., Venkataramani, V., Kamm, T., Zheng, T.F., Song, Z., Fung, P., Liu, Y. and Ruhi, U. (2001). "Automatic generation of pronunciation lexicons for Mandarin spontaneous speech." Acoustics, Speech, and Signal Processing, IEEE International Conference on, vol.1, pp. 569 –572.

3. Cao Yang, Deng Yonggang, Zhang Hong, Huang Taiyi and Xu Bo (2000). "Decision tree based Mandarin tone model and its application to speech recognition." Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on, vol.3, pp. 1759 –1762.

4. Chen C., Li Julian, Shen Haiping, Fu Liqin, Guo Kang (2001). "Recognize tone languages using pitch information on the main vowel of each syllable." Acoustics, Speech, and Signal Processing, 2001. Proceedings. IEEE International Conference on, vol.1, pp. 61 –64.

5. Chen Yeou-Jiunn, Wu Chung-Hsien and Yan Gwo-Lang (1999). "Utterance verification using prosodic information for Mandarin telephone speech keyword spotting." Acoustics, Speech, and Signal Processing, 1999. IEEE International Conference on, vol.2, pp. 697 –700.

6. Deb, Roy and Malamud Carl (1997). "Speaker identification based text to Audio Alignment for an Audio Retrieval System." Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Munich, Vol. 2, pp. 1099-1103.

7. Fu, Stephen W.K., Lee, C.H. and Clubb, Orville L. (1995). "A Survey on Chinese Speech Recognition." Communication Langagiere et Interaction Personne Systeme.

8. Gu Liang and Rose, Kenneth (2001). " Perceptual harmonic cepstral coefficients for speech recognition in noisy environment." Acoustics, Speech, and Signal Processing, 2001. Proceedings. IEEE International Conference on, vol.1, pp. 125 –128.

9. Guan Cuntai and Chen Yongbin(1993), "Speaker-independent tone recognition for Chinese speech." Acta Acustica, Vol. 18, No. 5, pp. 380-385.

10. He Qiang, Mao Shiyi and Zhang Youwei (2000). "Smoothed unit HMM in Mandarin speech recognition." Signal Processing Proceedings. WCCC-ICSP 2000. 5th International Conference on, vol.2, pp. 792 –795.

11. Holmes, J.N. (1993). *Speech Synthesis and Recognition*. NY: Chapman & Hall, Second Edition.

12. Hong Wei-Tyng and Chen Sin-Horng (1999). "A segment-based C_0 adaptation scheme for PMC-based noisy Mandarin speech recognition." Acoustics, Speech, and Signal Processing Proceedings. IEEE International Conference on, vol.1, pp. 433 –436.

13. Hong Wei-Tyng and Chen Sin-Horng (1999). "Robust SBR method for adverse Mandarin speech recognition." Electronics Letters, Vol. 35, pp. 875 –876.

14. Huang Hank, Chang Han and Seide Frank (2000). "Pitch tracking and tone features for Mandarin speech recognition." Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on, vol.3, pp. 1523 –1526.

15. Hung Wei-Wen and Wang Hsiao-Chuan (2000). "A fuzzy approach for the equalization of cepstral variances." Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on, Vol.3, pp. 1611 –1614.

16. Jiang Minghu, Zhu Xiaoyan, Xia Ying, Tan Gang, Yuan Baozong and Tang Xiaofang (2000). "Segmentation of Mandarin Braille word and Braille translation based on multi-knowledge." Signal Processing Proceedings. WCCC-ICSP 2000. 5th International Conference on, vol.3, pp. 2070-2073.

17. Lee Jin-shan (1990). "Continuous Mandarin Speech Recognition for Chinese Language with Large Vocabulary Based on Segmental Probability Model." IEEE ASSP Magazine, pp. 26-41.

18. Lee Lin-Shan (1997). "Voice Dictation of Mandarin Chinese: Computer Data Entry Without a Keyboard via Speech Recognition." IEEE Signal Processing Magazine, pp. 63-101.

19. Lee Lin-Shan, Chen Keh-Jiann, Tseng Chin-Yu, Lyu Renyuan, Chien Lee-Feng, Wang Hsin-Min, Shen Jin-Lin, Shen Jia-Lin, Lin Sung-Chiien, Yuang Yen-Ju, Bai Bo-Ren, Nee Chi-Ping, Liao Chun-Yii, Lin Shueh-Sheng, Yang Chung-Shu, Hung I-Jung, Lee Ming-Yu, Wang Rei-Chang, Lin Bo-Shen, Chang Yuan-Cheng,

Yang Rung-chiung, Huang Yung-Chi, Lou Chun-Yuan and Lin Tung-Sheng (1994). "Golden Mandarin (II) - An Intelligent Mandarin Dictation Machine for Chinese Character Input with Adaptation/Learning Function." International Symposium on Speech, Image Processing and Neural Networks, 13-16, pp. 155-159.

20. Lee Lin-shan, Tseng Chiu-yu, Chen Keh-Jiann, Hung I-Jung, Lee Ming-Yu, Chien Lee-Feng, Lee Yumin, Lyu Renyuan, Wang Hsin-min, Wu Yung-Chuan, Lin Tung-Sheng, Gu Hung-yan, Nee Chi-ping, Liao Chun-Yi, Yang Yeng-Ju, Chang Yuan-Cheng and Yang Rung-chiung(1993), "Golden Mandarin (II) - An Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with very large vocabulary." International Conference on Acoustics, Speech, and Signal Processing, April 21-24, 1997, vol. 2, pp. 503-506.

21. Lee Lin-Shan, Tseng Chiu-Yu, Gu Hung-Yan, Liu Fu-Hua, Chang Chen-Hao, Lin Yueh-Hong, Lee Yumin, Tu Shih-Lung, Hsieh Shew-Heng, and Chen Chian-Hung(1993), "Gold Mandarin (I)—Real-Time Mandarin Speech Dictation Machine for Chinese language with vary large vocabulary." IEEE Transactions on Speech and Audio Processing, vol. 1, No. 2.

22. Li Bavy N.L. and Liu James N.K. (1999). "A comparative study of speech segmentation and feature extraction on the recognition of different dialects." Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. IEEE International Conference on, vol.1, pp. 538 –542.

23. Li Jingjiao, Xia Xiaodong and Gu Shusheng (1999). "Mandarin four-tone recognition with the fuzzy C-means algorithm." Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE'99. IEEE International, vol.2, pp. 1059 –1062.

24. Liu Mingkuan, Xu Bo, Huang Taiyi, Deng Yonggang and Li Chengrong (2000). "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling." Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on, vol.2, pp. 1025 - 1028.

25. Lyu Ren-Yuan, Chien Lee-Feng, Hwang Shiao-Hong, Hsieh Hung-Yun, Yang Rung-Chiuan, Bai Bo-Ren, Weng Jia-Chi, Yang Yen-Ju, Lin Shi-Wei, Chen Keh-Jiaann, Tseng Chiu-Yu and Lee Lin-Shan (1995). "Golden Mandarin (III) – A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary." IEEE, pp.57-60.

26. Lyu Ren-Yuan, Hong I-Chung, Shen Jin-Lin, Lee Ming-Yu and Lee Lin-Shan (1998). "Isolated Mandarin Base-Syllable Recognition Based upon the Segmental Probability Model." IEEE Transactions on Speech and Audio Processing, vol. 6, No. 3, pp. 293-299.

27. Lyu Ren-yuan, Shen Jia-lin, Hong I-Chung, Lee Ming-Yu, Lee Lin-shan(1994), "A new approach for Mandarin base-syllable Recognition based upon Segmental Probability Model (SPM)." Proceedings of the 1994 International Conference on Computer Processing of Oriental Languages May 10-13 Taeyin, Korea, pp. 201-206.

28. Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE, Vol. 77, pp. 257-286.

29. Halstead Robert Jr., Benjamin Serridge Jean-Manuel, Van Thong and William Goldenthal (1996). "Viterbi Search Visualization using Vista: A Generic Performance Visualization Tool." Fourth International Conference on Spoken Language Processing, vol.3, pp.1910-1913.

30. Shen Jia-lin and Lee Lin-shan (1996). "Fast and accurate recognition of very-large-vocabulary continuous Mandarin speech for Chinese language with improved Segmental Probability Modeling." Proceedings of IEEE conference on Acoustics, Speech, Signal Processing, pp. 125-128.

31. Shen Jia-lin, Wang Hsin-min, Bai Bo-ren and Lee Lin-shan L (1994). "An Initial Study on A Segmental Probability Model Approach to Large-Vocabulary Continuous Mandarin Speech Recognition." International Conference on Acoustics, Speech, and Signal Processing, pp. 133-136.

32. Wang Hsin-Min, Ho Tai-Hsuan, Yang Rung-Chiung, Shen Jia-Lin, Bai Bo-Ren, Hong Jenn-Chau, Chen Wei-Peng, Yu Tong-Lo and Lee Lin-Shan (1997). "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data." Speech and Audio Processing, IEEE Transactions on, vol. 5, Issue 2, pp. 195-200.

33. Wang Hsin-min, Lyu Renyuen, Shen Jia-lin and Lee Lin-shan(1994), "An initial study on large-vocabulary continuous Mandarin speech recognition with limited training data based on sub-syllabic models." Proceedings of International Computer Symposium, pp. 1140-1145.

34. Wang Jhing-Fa, Wu Chung-Hsien, Chang Shih-Hung and Lee Jau-Yien(1991), "A Hierarchical Neural Network Model based on a C/V segmentation algorithm for isolated Mandarin speech recognition." IEEE Transaction on Signal Processing, vol. 39, No. 9, pp. 2141-2146.

35. Xuan Zhu, Li Husheng, Liu Jia and Liu Runsheng (2001). "Efficient decoding algorithms for Mandarin connected digit speech recognition Intelligent Multimedia." Video and Speech Processing. Proceedings of 2001 International Symposium on, pp. 555-558.

36. Zheng Fang, Chai Haixin, She Zhijie, Wu Wenhu and Fang Ditang(1998). "A Real-World Speech Recognition System Based on CDCPMs." Journal of Computer Processing of Oriental Languages, vol. 11, No. 3, pp. 121-232.

VITA

Ying-Chieh Chiang

Candidate for the Degree of

Master of Science

Thesis: A SURVEY OF MANDARIN CHINESE SPEECH RECOGNITION TECHNIQUES

Major Field: Computer Science

Biographical:

Education: Received Bachelor of Education in Guidance from National Changhua University of Education, Taiwan, Republic of China in June, 1992; received Master of Education degree in Special Education from the University of Central Oklahoma, Edmond, Oklahoma in August, 1998; Doctoral Study in Special Education: Visual Impairment Assistive Technology, May, 1999 (Incomplete); Illinois State University, Normal, Illinois; completed the requirements for the Master of Science degree with a major in Computer Science at Oklahoma State University, Stillwater, Oklahoma, December, 2002.

Working Experience: Research Assistant, Department of Special Education, National Changhua University of Education, Taiwan, R.O.C. 1992-1995. Planned and implemented renovation of talking books and the Talking Books Library for the Blind of this university. Simplified the transcription of printing materials into Chinese braille. Developed non-obstacle environment. Developed Chinese braille computer system.

Other Professional Activities: Guest speaker in the doctoral level course SP6563 Program Development in Special Education at Oklahoma State University. February 23, 2000. Presented on how the screen reader, braille display, and Optical Character Recognition (OCR), can benefit people with visual impairment. Guest speaker in the course Communication Skills for Students with Visual Disabilities in the Department of Special Education at

Illinois State University. March 23 and 30, 1999. Taught the use of Braille n' Speak (a braille note taker). Presented "The Future Development of Blind People".

Memberships: Member of the Committee of University and College Entrance Exams, Taiwan, R.O.C. 1996; Member of the Committee of University and College Entrance Exams for the Visually and/or Hearing Impaired, National Education Department, Taiwan, R.O.C. 1995.