

A Large-Vocabulary Speech Recognition System for Taiwanese (Min-nan)

Ren-yuan Lyu¹, Yuang-jin Chiang², Wen-ping Hsieh², Ren-zhou Fang²

¹Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan

²Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan

Email: rylyu@mail.cgu.edu.tw, rylyu@ms1.hinet.net

Tel: 886-3-3283016#5677

Abstract

In this paper an initial study and some preliminary work about Taiwanese (Min-nan, Southern Hokkien) speech recognition has been described, including a set of phonetic transcription symbols, a Taiwanese pronunciation lexicon more than 50 thousand words, several sets of phonetically balanced words, and a set of speech data. The inter-syllabic right context dependent Initial/Finals or phonemes are shown to be very useful in the acoustic modeling. Furthermore, we adopted not only model clustering based on acoustic decision tree to improve the data sharing, but also a hybrid duration model to improve the accuracy of state duration modeling in CHMM. Promising recognition rate and satisfactory recognition speed can be achieved. A prototype real-time speech recognition system has been constructed and some preliminary experimental results have also been reported. The recognition task defined here is a large-vocabulary, isolated-word (multi-syllabic), and speaker-dependent system.

1. Introduction

Taiwanese is the mother tongue of more than 75% of the population in Taiwan. It belongs to a larger Chinese dialectical family called Min-nan (or Southern-Min, Southern-Hokkien), which is also used by many overseas Chinese living in Singapore, Malaysia, Philippine, and other areas of Southern-East Asia. It was estimated that this language has more than 49 millions speakers and is ranked in the 21th place in the world [Ethnologue96]. In the past few decades, scientists and engineers in Taiwan conduct speech recognition research on Mandarin, the major Chinese dialect spoken by most Chinese, including those who living in Mainland China and Taiwan. Some achievements have been achieved in recent years.[L.Lee93][Ly95][Lyu98][H.Wang97] Since Taiwanese is another major language spoken in this land, and Taiwan is basically a multilingual society, a similar large-vocabulary speech recognition system for Taiwanese speech is worth of studying.

In this paper, some preliminary work has been done, including the study of Taiwanese phonetics and linguistics, setting up a Taiwanese pronunciation lexicon and a set of phonetic alphabet to symbolize Taiwanese speech, selecting several sets of phonetically balanced words to be used in speech data collection, and recording a Taiwanese speech database. Finally, a real-time system has been constructed to demonstrate the validity of the approaches proposed in this paper.

The basic technology adopted here is the continuous Hidden Markov Model (CHMM) because of its success in speech recognition in the past decades [Rabiner93]. We adopt CHMM to model the Taiwanese sub-syllabic phonetic units, including the Initial/Finals and phonemes, considering both the inside- and inter-syllabic coarticulation. Since the training data here is relatively few, sharing of data among the models is also considered.

Decision tree clustering using acoustic/phonological knowledge provides an efficient approach of data sharing. [Odell95] We choose clustering in the sub-phonetic level, i.e., states in the whole phonetic models, in which each specific state in each model has its own decision tree. A set of 25 yes/no phonological questions was used to construct each decision tree. After the tree was constructed, models within each leaf of the tree were tied together, and the states of these models shared the same training data.

It has been shown that traditional left-to-right HMM suffers from its poor state duration modeling, we also propose a hybrid-duration model to alleviate such a difficulty.

A promising final result, with the error rate being 6.94%, for the speaker dependent case was obtained, and a real-time prototype system in a Pentium-II personal computer running MS-Windows95/98/NT was implemented for further study and to validate the approaches we proposed here.

2. Taiwanese Phonetics/Linguistics

Taiwanese, like Mandarin as a member of Sino-Tibetan language family, is a tonal, syllabic language. Each Taiwanese sentence can be looked upon as a string of words. Most Taiwanese words can be written in the form of Chinese characters (Hanzi), but there are still a large portion of daily used words without commonly accepted written forms.

Nowadays, many writers use their own favorite spelling systems (usually in form of English characters) to spell out the pronunciation of those words without commonly accepted Chinese written forms. Some writers, however, insist on using Chinese characters to represent all Taiwanese words. Take the following sentence for an example. “咱 edang 去佗位 cit-tor ? ” (translated in Mandarin Chinese: “我們可以去哪裡玩 ? ”; in English: “Where can we go to play ? ”). One of another forms of the sentence by using Chinese characters completely is “咱會凍去佗位 迤迤 ? ”, where “會凍” is meaningless itself and just sound the same as “edang”, while “迤迤 “ does not even exist in the big-5 code set for traditional Chinese and may be strange to many literate people.

For those which have the Chinese written forms, each word, like in the case of Mandarin Chinese, can be composed of one to several Chinese characters and each Chinese character is usually pronounced as a mono-syllable with a lexical tone. For those which do not have Chinese written forms, each word is usually represented as a string of English characters to form one to several syllables, with or without a hyphen ‘-’ between two syllables. Phonetically speaking, in both cases, each Taiwanese word can be composed of one to several syllables. Each syllable, carrying one particular lexical tone, can be further decomposed into an optional Initial and a Final. An Initial is just a consonant phoneme, while a Final may be a vowel, a vowel plus a nasal consonant, or a vowel plus stop consonant. The linguistically/phonetically hierarchical structure of a Taiwanese sentence can be shown as in <table.1>.

<table.1> The linguistic/phonetic hierarchy of a Taiwanese sentence.

Table 1 The linguistic/phonetic hierarchy of a Taiwanese sentence.																	
Sentence (Form 1)	咱 edang 去佗位 cit-tor																
Sentence (Form 2)	咱會凍去佗位迤迤																
word	咱 (we)	edang (can)		去 (go)	佗位 (where)		cit-tor (to play)										
morpheme / syllable	咱	e	dang	去	佗	位	cit			tor							
tonal-syllabl e	lan4	e1	dang3	ki3	dor4	ui2	cit7			tor5							
tonal-syllabl e (with sandhi)	lan4	e2	dang4	ki4	dor1	ui2	cit6			tor5							
base-syllable	lan	e	dang	ki	dor	ui	cit			tor							
Initial/Final	l	a	n	e	d	a	ng	k	i	d	or	u	i	c	it	t	or
phoneme	l	a	n	e	d	a	ng	k	i	d	or	u	i	c	i	t	or

There are 18 Initials (including one null Initial), about 90 Finals, and 7 lexical tones in Taiwanese. [Wang57]. These basic phonetic units can be further combined to form a set of about 2000 tonal syllables and about 800 base-syllables, which were counted from 58270 words of the electronic lexicon available to us. The set of 18-Initials/90-Finals can be transformed to a set of 35 phonemes, if we divide each Final into one to several phonemes.

All the phonemes are listed in <table.2>, each of which has a corresponding Chinese character with that phonetic units as part of its pronunciation. The phonemes are represented by 3 alternative symbolic systems, including the International Phonetic Alphabet (IPA), the Church Roman (a phonetic alphabet system traditionally used in Church to read the Bible in Taiwanese) [Cheng97] and a set of newly proposed phonetic alphabet called Tongyong Pinyin, where Tongyong Pinyin is designed to encode all native languages used in Taiwan, including Mandarin, Taiwanese, and Hakka in a uniform form, and thus is adopted in our research widely.[Yu99] In <table.3> and <table.4>, all 18 Initials and about 90 Finals are also listed for reference in Tongyong Pinyin.

<table.2> A List of Phonemes in Taiwanese

consonant				consonant				Vowel			
Chinese Character	IPA	Church Roman	Tong-yong	Chinese Character	IPA	Church Roman	Tong-yong	Chinese Character	IPA	Church Roman	Tong-yong
保	p	p	b	資	ts	ch	z	阿	a	a	a
坡	p'	ph	p	此	ts'	chh	c	伊	i	i	i
冒	m	m	m	思	s	s	s	有	u	u	u
帽	b	b	v	如	z	j	r	鞋	ɛ	e	e
刀	t	t	d	好	x	h	h	烏	ɔ	o'	o
討	t'	th	t	英	ø			蚵	ɔ	o	or
怒	n	n	n					餡	ã	a ⁿ	ann
路	l	l	l					嬰	ĩ	i ⁿ	inn
糕	k	k	g					樣	ũ	u ⁿ	unn
科	k'	kh	k					嬰	ẽ	e ⁿ	enn
雅	ŋ	ng	ng					惡	õ	o ⁿ	onn
鵝	g	g	q								

Chinese Character: a character which has the corresponding phoneme as its partial pronunciation

IPA: the International Phonetic Alphabet

Church Roman: A traditional phonetic symbol set for reading the Bible in church in Taiwanese

Tongyong: A specially designed Taiwanese Phonetic Alphabet used throughout this paper

<table.3> 18 Initials of Taiwanese in Tongyong Pinyin.

b	p	m	v	d	t	n	l	g	k	ng	g	z	c	s	r	h	Null
---	---	---	---	---	---	---	---	---	---	----	---	---	---	---	---	---	------



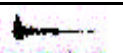







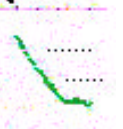
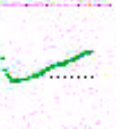


<table.4> 94 finals of Taiwanese in Tongyong Pinyin

a	ah	ia	iah	un	ut	iaunn	iaunnh
e	eh	ior	iorh	uan	uat	uann	uannh
i	ih	iu	iu	uang	uak	uenn	uennh
o	oh	iau	iauh	m	mh	uinn	uinnh
or	orh	iam	iap	ng	ngh	uainn	uainnh
u	uh	iang	iak	ann	annh		
ai	aih	iong	iok	enn	ennh		
au	auh	im	ip	inn	innh		
am	ap	in	it	onn	onnh		
an	at	ing	ik	ainn	ainnh		
ang	ak	ua	uah	aunn	aunnh		
en	et	ue	ueh	iann	iannh		
om	op	ui	uih	ionn	ionnh		
ong	ok	uai	uaih	iunn	iunnh		

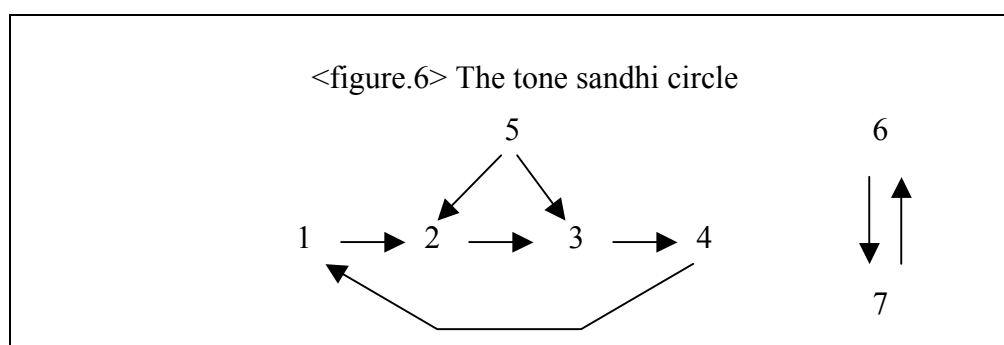
Furthermore, Taiwanese is also a tonal language with more complex tonal structures than that of Mandarin. It has 7 lexical tones, two of which are carried in syllables ending with final /-p, -t, -k, -h/ (called entering-tone traditionally) and the other five are carried in those ending without final /-p, -t, -k, -h/ (called non-entering tone traditionally). An example of these 7 tones with one corresponding Chinese character for each tone is shown in <table.2>. Some acoustic characteristics, including the waveform, the contour of fundamental frequency, the description of relative frequency level, and a proposed

digit-to-tone mapping are also shown in <table.5>.

<table.5> The 7 lexical tones of Taiwanese

Chinese Character	東	洞	棟	黨	同	獨	督
Waveform							
Fundamental Frequency							
Relative Frequency	High Level	Mid Level	Low Falling	High Falling	Rising	High Stop	Low stop
Digit to Tone Mapping	1	2	3	4	5	6	7

The tone sandhi issue is relatively complex in Taiwanese. Almost each syllable (may corresponding to one Chinese character if it exists) has 2 kinds of tones called base-tone and sandhi-tone depending on the position it appears in words or sentences. One of the most frequently used sandhi rules is the so-called “tone sandhi circle”. [Chiang98] It says that if a syllable appears at the end of a sentence, or at the end of a word, then it is pronounced as its base-tone. In most the other cases, however, it is pronounced as its sandhi tone. To describe the sandhi rule conveniently, let’s defined the number 1 to 7 to encode the 7 Taiwanese tones as follows: (1) High-Level (like 東), (2) Mid-Level (like 洞), (3) Low-Falling (like 棟), (4) High-Falling (like 黨), (5) Rising (like 同), (6) High-Stop (like 獨), (7) Mid-Stop (like 督). Under such a mapping from digit to tone pattern, the “tone sandhi circle” says that tone 1 will change to tone 2, tone 2 will change to tone 3, ..., and so on. There are 2 different sandhi situations for tone 5 depending on 2 major different sub-dialects. The “tone sandhi circle” can be shown in <figure.6>



More tone sandhi rules, including triple-adjective tone sandhi, neutral tone sandhi exist and were described in detail in the literatures [Cheng97][Wang57].

Another linguistic/phonetic issue about Taiwanese is the phonetic variations of people whose ancestors came from different sub-dialectic areas of Fujian Province in China. There are two main sub-dialects, i.e., Qyuan-zhou, and Zhang-zhou. These two main sub-dialects, however, are mutually intelligible and most people in Taiwan even don’t realize which sub-dialect they are using. Take the following characters as examples, “字、皮、鞋、旅、嬰、羊”. Their pronunciation in both sub-dialects are shown in <table.7>.

<

<table.7> Two main sub-dialects in Taiwanese for several example Chinese characters.

Chinese Character	字	皮	鞋	旅	嬰	羊
Qyuan-zhou	li	pe	ue	li	inn	iunn
Zhang-zhou	ri	pue	e	lu	enn	ionn

One more systematic phonetic variation occurs in the different situations of the language being used. When Taiwanese is used to read the classical literature like poetry, it uses the classical pronunciation system, otherwise it use the oral pronunciation system in daily lives. Take the 4-word phrase “落花流水” as an example. Its two distinct pronunciations in both usage situations are shown in <table.8>

<table.8> Two distinct pronunciations in both usage situations for Taiwanese

Chinese Character	落	花	流	水
Classic usage	lok	hua	liu	sui
Oral usage	lor	hue	lau	zui

3. The Task Definition

To deal with all the linguistic/phonetic issues well at once in this initial study is impractical. Instead, we define the task as a more feasible one by considering the most important issues but ignoring the other issues.

Because of the non-uniformity of the written form and the shortage of the text material for Taiwanese, there are more difficulties to construct a general-purpose dictation machine for Taiwanese than for Mandarin. To create a real-time system as soon as possible for further research, we adopt word as the basic speech recognition unit because it's not so difficult to set up a Taiwanese pronunciation lexicon. We decided not to deal with the issues about grammar which govern the context information of words in a sentence. All we want to do is to recognize an utterance and then output the corresponding written form already exists in the lexicon.

As described previously, a Taiwanese word may be composed of from one to several syllables. Since the one-syllable word has many homonyms, which should be discriminated from one another by context information, and thus is beyond this initial study, we consider to recognize only the multi-syllabic words. We further found that there are relative few homonyms for a multi-syllabic word found in the lexicon even when the tones are disregarded. We thus decided not to deal with the issues of tones and then reduce all phonologically allowed tonal syllables to about 800 base syllables. That is, each word in the lexicon is represented as a concatenation of base syllables. The word recognition task becomes the recognition of base syllable strings. After the base-syllable strings are recognized, it's easy to extract the text of the word, and output the result to the users.

Furthermore, to avoid to consider the phonetic variations, the system is constrained to speaker-dependent mode. That is, before using, users should provide their own voice to train the system.

4. The Database

To begin with the study of the large-vocabulary speech recognition of a new language,

like Taiwanese we are studying now, one of the most important preliminary work is to construct a pronunciation lexicon. We have set up a Taiwanese pronunciation lexicon of more than 50 thousand (50K) words [Chiang94], each of them has a corresponding string of phonetic symbols encoded in Tongyong phonetic alphabet. From the 50K-word lexicon, a subset of about 20K frequently used words have been selected for the task. In this 20K lexicon, there are in fact 19152 ordinarily used Taiwanese words, composed of 48318 syllables, i.e., each word contains 2.52 syllables in average.

Another important preliminary task is to select a training script which contains as few words but as many phonetic varieties as possible. To achieve this, a word selective procedure is set to choose appropriate words as follows:

- 1) *Determine the phonetic unit to be used in the recognition system;*
- 2) *Each new word selected contains the maximal number of possible new phonetic units;*
- 3) *Include all distinct speech units which appear in the lexicon.*

As a result, a minimal set of 472 words containing all the 1024 distinct phonemes and phoneme pairs found in the lexicon were selected. In addition, 4 more sets of words, which contain as many distinct phoneme pairs as possible, were also selected to enhance the phonetic varieties. Furthermore, a set of single-syllabic words, containing all tonal phonologically possible syllables, was picked out, too. The statistics of all the sets of words used in the training session is listed in <table.9>.

For evaluation of the recognition system, we select several sets of words with different features:

- 1) R1000: 1000 randomly selected words, each of which contains 2.55 syllables;
- 2) H500: 500 highest frequently used words, each of which contains 2.12 syllables;
- 3) N407: 407 place names, each of which contains 2.08 syllables;
- 4) P396: 396 phonetically rich words, each of which contains 3.24 syllables.

The statistics of the evaluation set is listed in <table.10>.

The speech database currently used for training and evaluation were recorded by two adult speakers, including one male and one female, over a period of one month. A close-talk head-set microphone plugging in a SoundBlaster card in a Pentium-II personal computer was used. The speech waveform was sampled at 16 KHz. The statistics of the speech database is also listed in <table.9> and <table.10>.

<table.9> Statistics of the Lexicon and the Training Word Sets

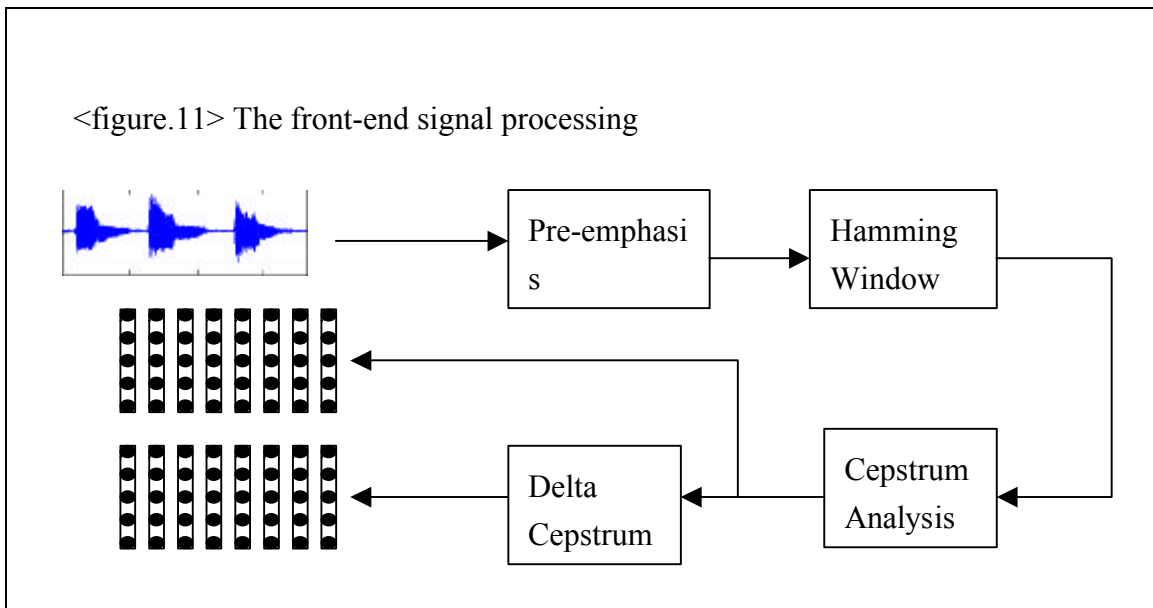
		Number of words	Number of distinct phonemes and phoneme pairs	Speech Length in seconds	
				Male	Female
Training Word Sets	Single_syllble	2,874	213	1417	1486
	Min_word	472	1,029	459	445
	Ext_word	1,045	1,029	965	981
	The whole	4391	1029	2841	2912

<table.10> Statistics of the Lexicon and the Testing Word Sets

		Number of words	Number of syllables per word	Speech Length in seconds	
				Male	Female
Testing Word Sets	R1000	1000	2.55	826	656
	H500	500	2.12	361	397
	N407	407	2.08	304	311
	P396	396	3.24	385	256
	The whole	2303	2.49	1876	1620

5. Front-end Signal Processing

The speech waveform was multiplied by a 16-ms Hamming window first. A set of 12-dimensional mel-cepstral coefficients and 1-dimensional log energy was extracted to form a 13-dimensional feature vector for each frame which shifts forward every 8 ms. A time window of 5 frames of feature vectors were used to compute the corresponding 13-dimensional delta coefficients. These 2 sequences of feature vectors and delta feature vectors were treated as statistically independent and modeled by separate Gaussian mixture densities in CHMM. The overall block diagram for front-end signal processing was shown as in <figure.11>:



6. Selection of Speech Units

In this paper, we adopted Initial-Finals and phonemes, considering the context dependency both inside a syllable and inter syllables, as the basic speech units to be modeled as CHMM. It is believed that the coarticulation effect inside a syllable is more severe than that between 2 syllables for the monosyllabic language, such as Mandarin or Taiwanese. So, it is natural for researchers to consider the inside-syllable coarticulation in the previous literatures. [Lyu95] In such a case, only Initials can be right context dependent

(RCD) and all Finals are context independent (CI). There are thus 147 RCD Initial models and 77 CI Final models. However, when the speed of utterance increases the coarticulation across 2 syllables becomes severe. In addition, for the vowel-vowel concatenation between 2 neighboring syllables, the coarticulation effect may be very severe even when the speed of utterance is slow. To alleviate such a difficulty, the inter-syllabic modeling was considered.

However, the number of general RCD Finals is so large that we chose not to use it directly. Instead, we added the inter-syllabic RCD (ISRCD) bounded phones explicitly to model the coarticulation effect. For examples, the bi-syllabic word “pue-e” (皮鞋), will be looked upon as the concatenation of /p+u/, /ue/, /e+e/, and /e/, where the unit /e+e/ is what we called the inter-syllabic RCD bounded phone. By this approach, 105 additional units were obtained. As we will see in the experiments, such an explicit consideration about the inter-syllabic coarticulation does decrease the word error rate at a little cost of additional computation.

In addition to Initial/Finals, we also adopted phonemes, considering the right context dependency both inside a syllable and inter syllables, as the basic speech units to be modeled as CHMM. There are 208 inside-syllable RCD (iRCD) phones. To deal with the problem of co-articulation across syllables, we consider using the outside-syllable RCD phones (oRCD), in which the right most phone in a syllable is dependent on the left most phone of the next syllable. From the same 20K lexicon, 1029 oRCDs are extracted. Since each new model not found in the set of iRCDs comes from certain iRCD with syllable boundary as its right context, it is reasonable to use such a kind of iRCD models as seed models to generate all its corresponding oRCD models.

Similar to the case in Initial/Finals, the number of oRCDs is relatively large in view of the limited amount of speech data available, we have to consider some data sharing approaches to make full use of the speech data. The acoustic decision tree method is an efficient approach to achieve this purpose and will be described in detail in next section.

7. State Clustering Based on Acoustic Decision Tree

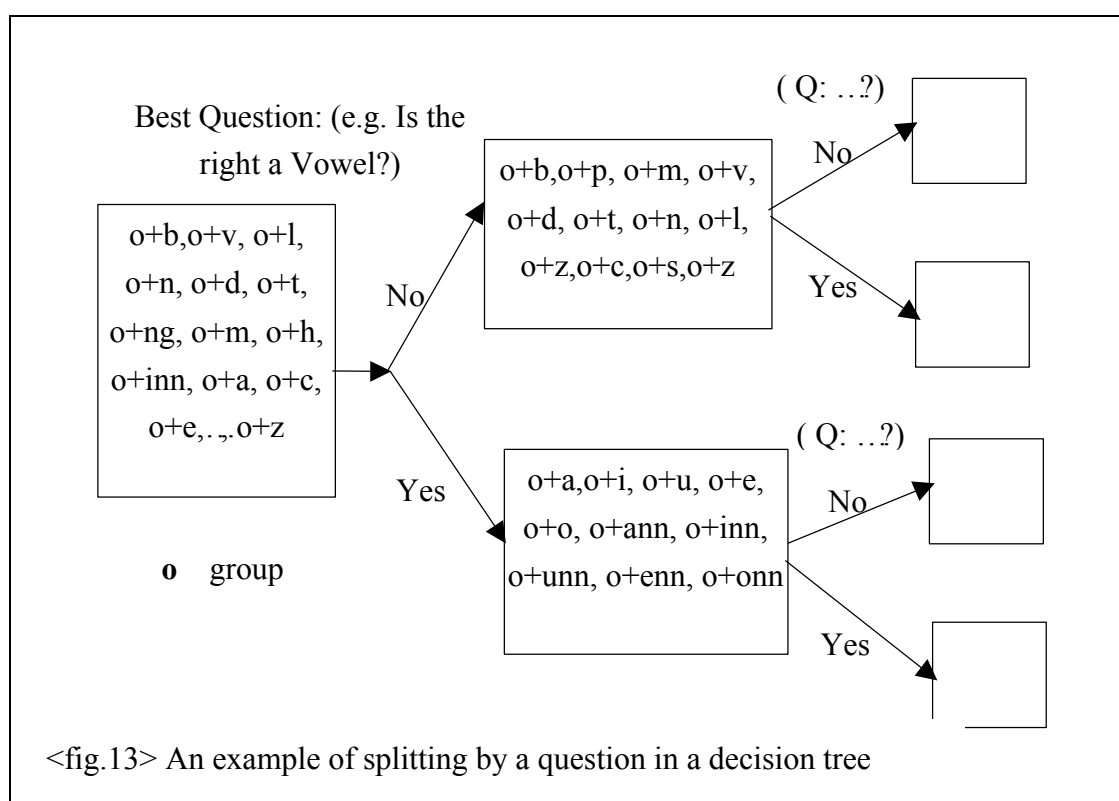
It is known that increasing the specificity of models may decrease the trainability because of a limited amount of training data. To get the optimal trade-off between specificity and trainability, data should be shared among similar models. The method of model clustering based on acoustic decision tree is known to deal with this problem quite well. [Odell95]

An acoustic decision tree is a binary tree in which each node contains a set of acoustic models with similarities. A yes/no question is attached to each node, which is then split into 2 child node, each containing a subset of models with higher degree of similarity. Those questions are designed from knowledge of phonetics. A set of 25 questions specially designed for Taiwanese speech was proposed here. All questions are listed in <table.12>.

<table.12> 25 phonetic questions for Acoustic Decision Tree

	Question	Phoneme Members for Yes		Question	Phoneme Members for Yes
Q1	vowel?	a, ann, e, enn, i, inn, o, onn, or, u, unn	Q14	bilabial?	b, p, m, v
Q2	nasal_vowel?	ann,enn,inn,onnn,unn	Q15	alveolar?	d, t, n, l
Q3	a_vowel?	a, ann	Q16	velar?	k, g, q, ng
Q4	e_vowel?	e, enn	Q17	pre-alveolar?	z, c, s, r, t
Q5	i_vowel?	i, inn	Q18	affricative?	c, z
Q6	o_vowel?	o, onn	Q19	liquid?	l
Q7	u_vowel?	u, unn	Q20	fricative?	s, r
Q8	glottal stop or unreleased p, t, k?	-p,-t,-k,-h	Q21	's'?	s
Q9	front vowel?	i, e, inn,enn	Q22	'r'?	r
Q10	mid vowel?	a, ann, or	Q23	'z'?	z
Q11	back vowel?	u, unn, o, onn	Q24	'c'?	c
Q12	voiced consonant?	m, n, -m, -n, -ng, b, d, g, z, l, q, v, r	Q25	't'?	t

For each node, the “best” question is chosen to be asked and the node is split according to the answer to the best question. As shown in <fig.13>, where the parent node contains phoneme ‘o’ as the kernel, then they are split into 2 subsets by a “best” question.



To set a criterion to decide the best question for splitting a tree node, we define a measure for average distance (dissimilarity) L_S of a set of models as follows,

$$L_S = \sum_{m \in S} N_{mi} |\bar{\mathbf{m}}_{mi} - \bar{\mathbf{c}}|^2$$

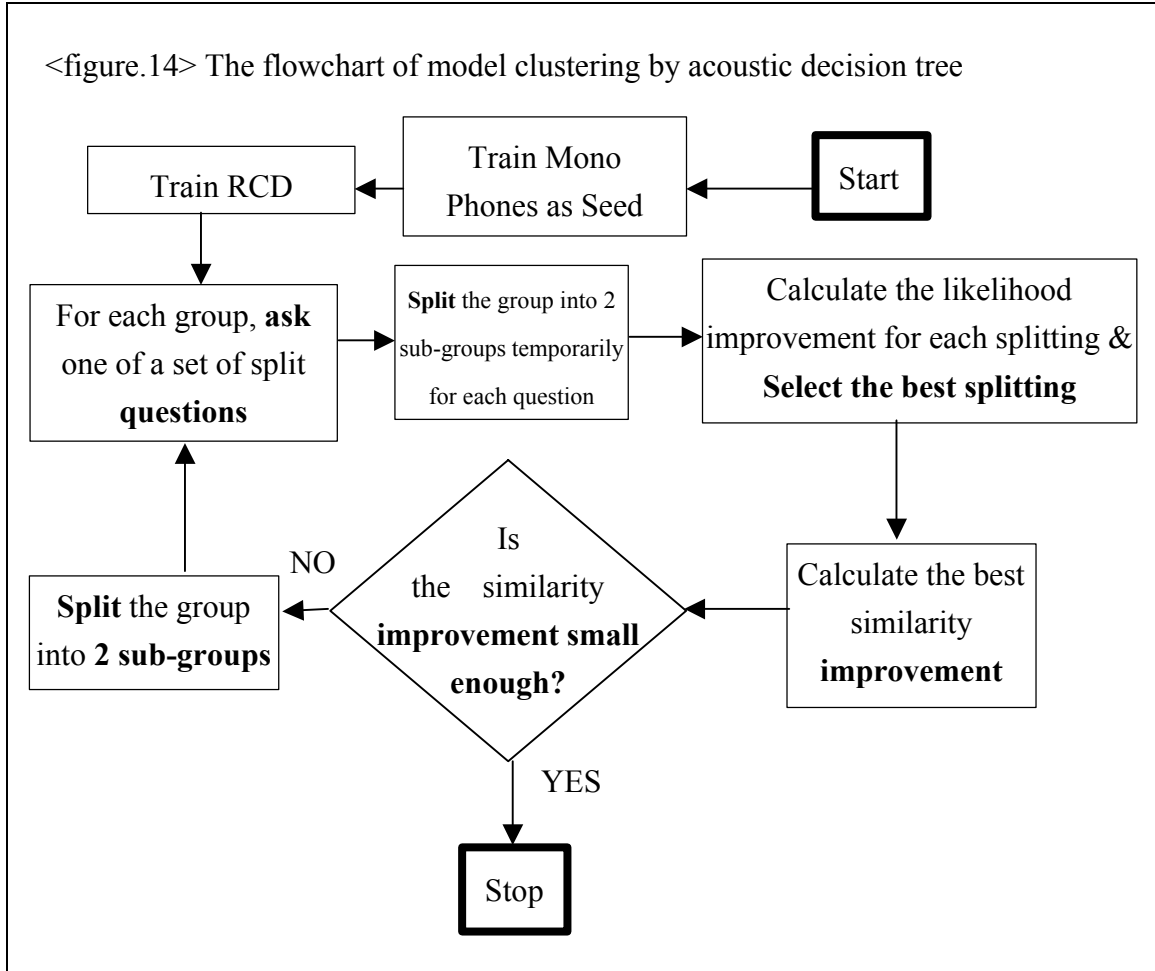
where N_{mi} is the occupancy of state i of model m , $\bar{\mathbf{m}}_{mi}$ is the mean vector of state i of a model m in the tree node S , and $\bar{\mathbf{c}}$ is the center vector of all $\bar{\mathbf{m}}_{mi}$ in node S . Each splitting should maximize the decrease of L_S , i.e., the increase of similarity, when the parent node S splits into 2 children nodes S_1 and S_2 .

In other words, we have to maximize

$$\Delta L = L_S - (L_{S_1} + L_{S_2}).$$

A threshold of ΔL is set by experiments to stop the process of splitting the tree. After the tree is constructed, states within each leaf node are tied together and the key states of these models share the same training data.

We construct decision trees for each state of the kernel phone. All the 25 questions were asked to make judgments about right contexts. The procedure to cluster acoustic models by decision tree is shown in <figure.14>.



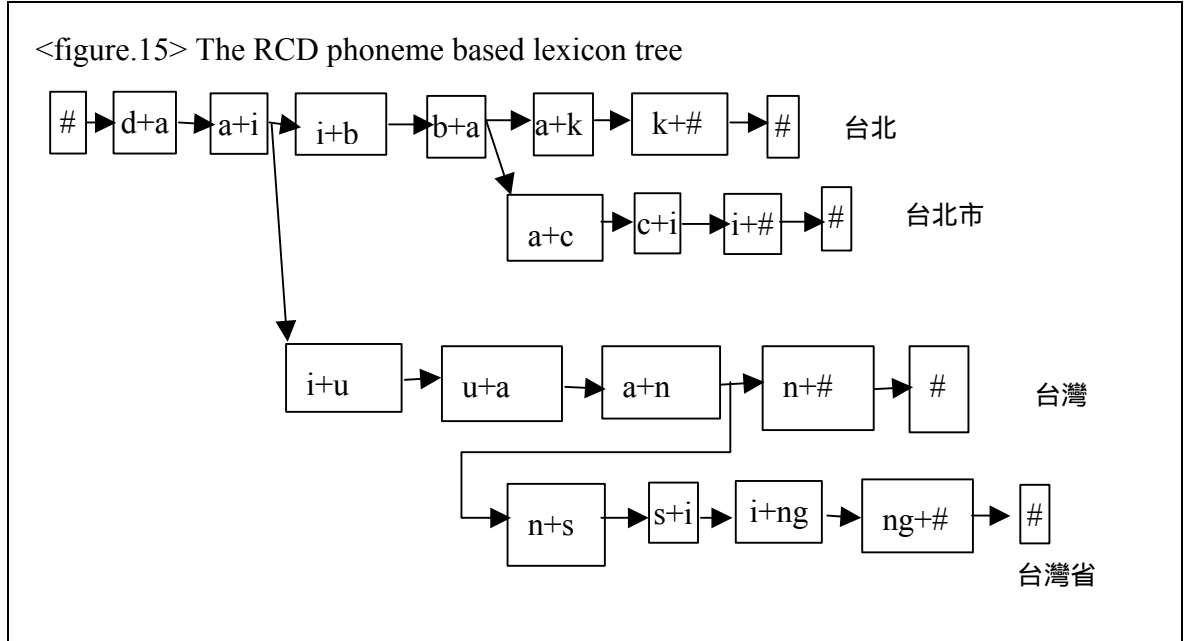
The advantage of tree classification is that it can generate a vocabulary-independent system. For each new RCD unit, we can ask the same 25 questions in a hierarchical manner and classify it into a leaf node. Whether or not this unit has been shown in the training corpus, it shares the model of that leaf node and thus can be recognized.

The acoustic decision tree can be applied to any level of speech unit. However, since

each state has its own effect on the model, we choose state as the units to be clustered. In order to get better trade-off between number of states and recognition accuracy, experiments of clustering different states in a model have been tried and all the results are reported in the experimental section.

8. Lexicon Tree Search

The 20K-word lexicon is organized in terms of the chosen speech units as a tree data structure to be used as the search space. There are about 58K nodes in the lexicon tree, with each node containing one chosen speech unit. Compared with a plain linear lexicon, which contains about 124K nodes, the tree lexicon saves more than a half storage space. In addition, the searching speed is much faster in the tree lexicon. A sub-tree is shown in <figure.15>. A widely used Viterbi beam search is then used to find N best paths and then the N candidates of the recognized words. [C.H.Lee89]



9. Hybrid Duratin Model

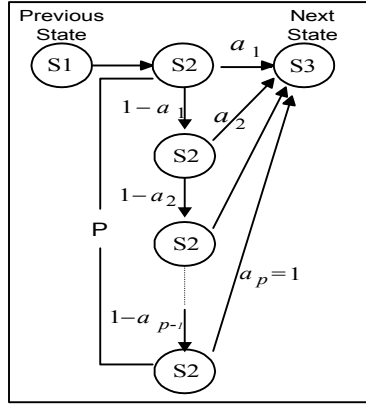
It has been shown that the traditional left-to-right topology of HMM suffered from its inaccurate state duration modeling. The state transition probabilities are constant and this implies, if unconstrained, the number of consecutive frames generated from a state would have a geometric distribution. The duration distribution of a typical state is thus:

$$f(d) = (1 - a_{self}) \times (a_{self})^{d-1}, \text{ where } d \in N,$$

where a_{self} is the transition probability of a state transiting to itself.

We see that the actual observed duration is distributed more like a gamma function. The discrepancy between modeled and actual distribution suggests more refined modeling for the duration distribution.

In order to model the duration more accurately, a finite duration model topology for HMM has been proposed [Picone90]. This is done by replacing the self-looping state with a string of P replicated states, as illustrated in <fig.16>.



<fig.16> The topology of a particular state in the finite duration model

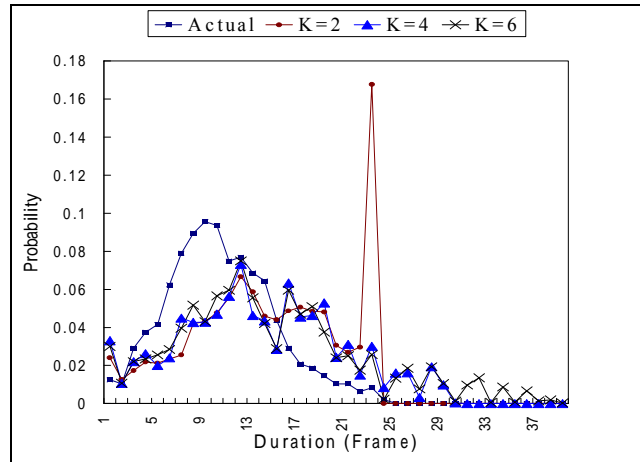
The number of replicated states, P , is determined by

$$P = E(S) + K \times stddev(S),$$

where S is the number of frames that have been mapped to the state obtained from the infinite duration experiment, $E(S)$ is the expected value of S , $stddev(S)$ is the standard deviation of S , and K is a small fixed integer to be determined experimentally. The duration distribution of a particular state is then

$$f(d) = a_d \prod_{i=1}^{P-1} (1 - a_i), \text{ where } a_i = 0 \text{ if } i \geq P$$

The resulting estimated duration distributions and the actual observed duration from the infinite duration case for a particular state are sketched in <fig.17>. We see that as K grows, the distribution is closer to the actual one.



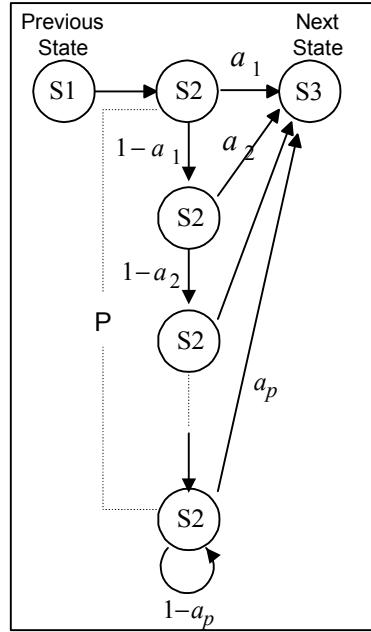
<fig.17> The duration distribution for finite-duration modelin with different K values

Unfortunately, this topology comes with a price in computation time. Since the number of replicated states is large due to the large variance of duration, this leads to a relatively large parameter space and search space, and makes it impractical to be used in a real time system. To alleviate such a deficiency, we try to combine the finite duration modeling with traditional infinite duration modeling in hope to improve the WER with little computations. This leads to a hybrid duration model to be discussed in next

paragraph.

By making the last replicated state have self loop in the finite duration model, we have the first hybrid duration model, as illustrated in <fig.18>.

<fig.18> The topologies of a particular state in First Hybrid Duration Modeling



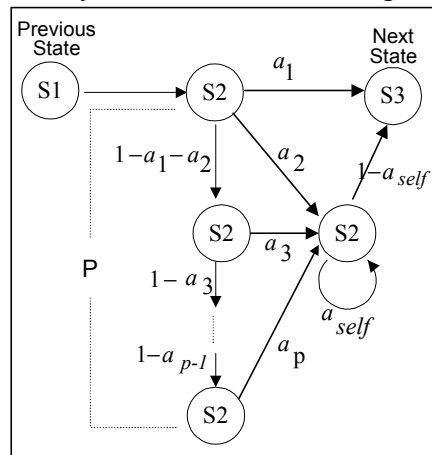
The distribution of the first hybrid duration model is as follows:

$$f(d) = a_d \prod_{i=1}^{P-1} (1-a_i), \text{ where } a_i = 0 \text{ if } i > P$$

Under this model, the recognition result is comparable to the finite duration case, while the computational efficiency is improved.

After several trials, we finally come up with a somewhat more complicated model as illustrated in <fig.19>.

<fig.19> The topologies of a particular state in Second Hybrid Duration modeling

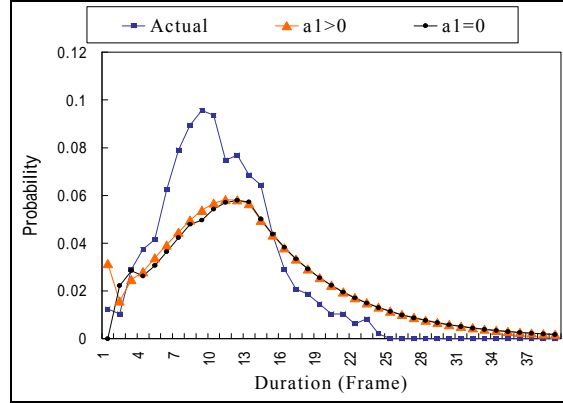


The distribution for the second hybrid duration model can be shown as

$$f(d) = \begin{cases} a_1 & \text{if } d=1 \\ a_2 \times (1 - a_{self}) & \text{if } d=2 \\ f(d-1) \times a_{self} + (1 - a_1 - a_2) \times \left(\prod_{i=3}^{d-1} (1 - a_i) \right) \times a_d \times (1 - a_{self}) & \text{if } 3 < d \leq P+1 \\ f(d-1) \times a_{self} & \text{if } d > P+1 \end{cases}$$

For the case $a_1 = 0$ and $a_1 > 0$, the estimated distribution and the actual one are depicted in <fig.20>. Although the shapes of the estimated densities between $a_1 = 0$ and $a_1 > 0$ do not differ much, both the accuracy and computational efficiency for $a_1 = 0$ are improved.

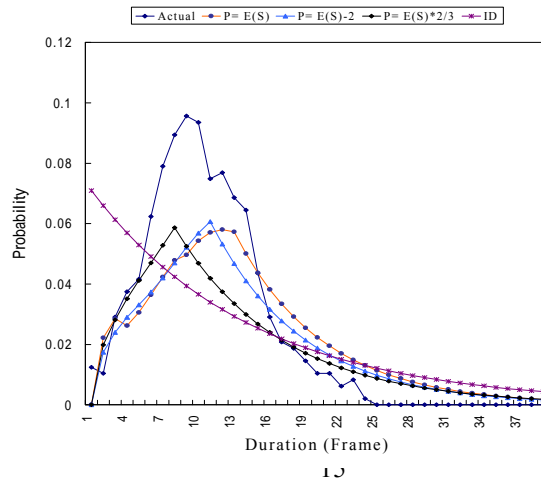
<fig.20> The duration distributions for 2 types of hybrid duration models



Further examination of the experimental results reveals that many recognition errors occur because that some states have short duration. This suggests a model with minimum duration $m < P$. This can be easily done by setting $a_1 = \dots = a_m$. The case that $a_1 = 0$ corresponds to $m = 1$. By choosing m and P carefully, both the recognition rate and computation time are improved.

In <fig.21> we compare the estimated densities for $m = E(S)/2 \times 3$, $E(S) - 2$, and $E(S)$.

<fig.21> The duration distribution for hybrid duration models plus minimal duration m



10. Experiments and A Prototype System

To validate the approaches proposed here, several experiments have been done.

The baseline system is based on 147 CD-Initials / 77 CI-Finals, the state number (N) for each model is 3 for the Initial, and 3 multiply the number of phones for the Final. The mixture number (M) is set to 1 here for the purpose of computation efficiency in real-time implementation. The word error rate (WER) for the case is 15.42% and the number of states (NS) in the searching space is 981, where NS is approximately proportional to the computational time. By adding the 105 inter-syllabic models, The WER is down to 11.54% while NS is up to 1296.

Replace the Initial/Final modeling with the phone modeling, i.e., using 208 inside-syllabic RCD phone models, where each model has the same configuration as that in Initial/Finals. The WER is up from 11.54% to 12.00%. The increase in WER is because that the inter-syllabic modeling has not been applied yet. However, because of the significantly reduce in the state number (NS is down from 1296 to 624), such minor increase in WER is worth. By adding the inter-syllabic RCD phone models, the number of total states expands from 624 to 3,087, while the WER is down significantly from 12.00% to 7.89%. Adopting model clustering on state 1 of the model by acoustic decision tree will reduce the WER to be 7.18% and the NS to be 2,501 simultaneously. When the clustering is put on both state 1 and state 2, the WER is 7.22% and the NS is 1,902. When the clustering is put on all state 1, state 2 and state 3, the WER is 8.54%, while the NS is reduced to be 1,368.

Furthermore, replace the left-to-right model with the finite duration model will increase the WER to be 7.32%, and increase the NS to be 6,793. If we use the proposed hybrid duration model to replace the finite duration model, then the NS is significantly reduced to be 2,140, while the WER is up again to be 7.51%. Finally, if the minimum duration constraint is put on the hybrid duration model, the WER is reduced to be 6.94%, the best result of this paper, and the NS is kept at a moderate level to allow real-time implementation. All the experimental results are summarized in <table.22>.

<table.22> The summary of the experimental results

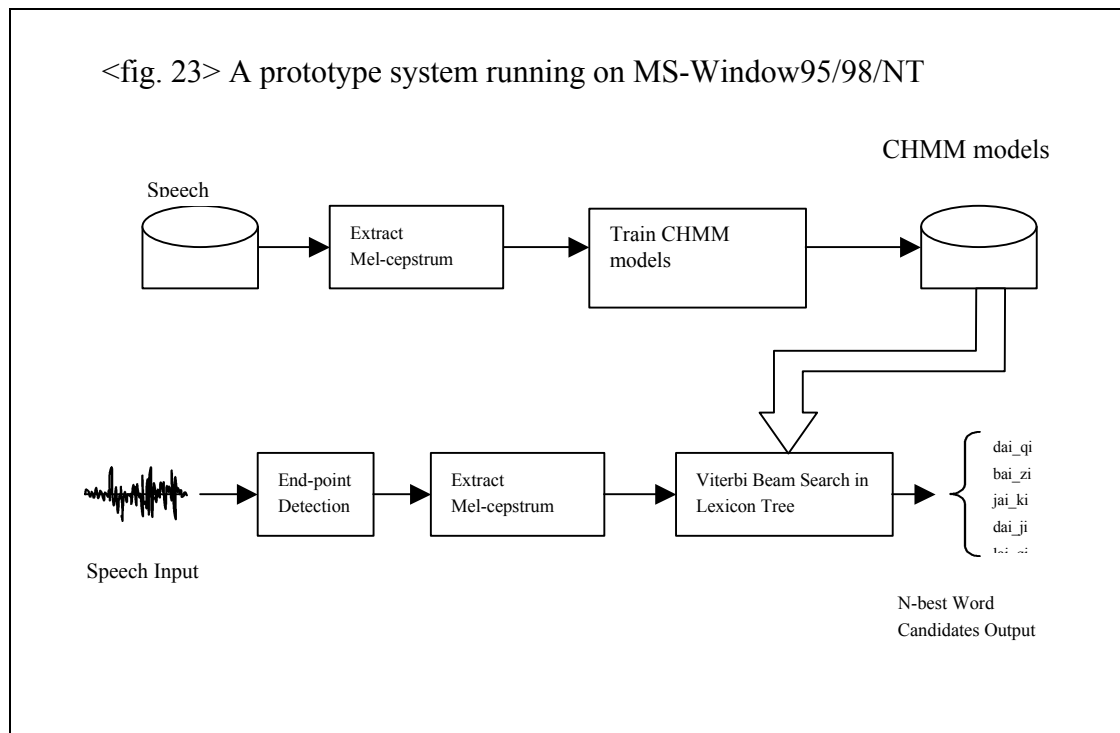
WER : the word error rate

NS : the number of states, approximately proportional to computation time

	WER %	NS
147 CD-Ini/77CI-Final	15.42	981
Plus 105 ISRCD	11.54	1,296
208iRCD	12.00	624
1029oRCD	7.89	3,087
1029oRCD/ S1	7.18	2,501
1029oRCD/ S1,S2	7.22	1,902
1029oRCD/ S1,S2,S3	8.54	1,368
Finite Duration Modeling	7.32	6,793
Hybrid Duration Model	7.51	2,140
Plus Minimum Duration	6.94	2,140

To validate the approaches proposed in this paper, a prototype system was implemented on a Pentium-II personal computer running MS-Windows95/98/NT, by using the best configuration as described in <table.22>, considering the trade-off between speed and accuracy. The overall block diagram of the system is shown in <figure.23>, while the

graphic user interface (GUI) is shown as in <fig.24>, where a short paragraph consisting of multi-syllabic words are demonstrated.



<fig.24> The GUI of the prototype system implemented in MS-Windows95/98/NT



11. Reference

- [Ethnologue96] <http://www.sil.org/ethnologue/top100.htm>
- [Lyu95] Ren-Yuan Lyu, et al. "Golden Mandarin (III)-User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary", ICASSP-95, pp57-60
- [Lyu98] **Ren-yuan Lyu**, I-Chung Hong, Jia-Lin Shen, Ming-Yu Lee, Lin-Shan Lee, "A New Approach for Isolated Mandarin Syllable Recognition Based Upon Segmental Probability Model (SPM)", *IEEE Transactions on Speech and Audio Processing*, pp. 293-299, Vol.6, No.3, May, 1998
- [Wang57] 王育德, "台灣語常用語彙", 永和語學社, 1957.
- [Chiang94] 江永進, "台音式輸入法 version4.1", 臺灣新竹清華大學統計所, 1994
- [Chinag97] 江永進, "台音式調記順序 e 選擇理由", 台灣研究通訊, 第十期, 1997
- [Chiang98] 許世楷等, 江永進執筆, "口語調自然調形", 台灣世界 12 期, 1997
- [Cheng97] 鄭良偉, "台語的語音與詞法", 遠流, 1997
Robert L. Cheng, "Taiwanese and Mandarin Structures and Their Developmental Trends in Taiwan--I: Taiwanese Phonology and Morphology", 1997
- [C.H.Lee89] C.H.Lee, .. etc, " A frame-synchronous network search algorithm for connected Word recognition", IEEE Trans. ASSP, pp. 1649-1658, Nov. 1989
- [L.Lee93] Lin-shan Lee, etc, 'Golden Mandarin (II) - An Improved Single-chip Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary', *ICASSP-93*, Apr. 1993, pp. II-503-506
- [H.Wang97] Hsin-min Wang, et al., "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data", IEEE Trans. on Speech and Audio Processing, vol. 5, no. 2, pp.195-200, March 1997.
- [Odell95] J.J. Odell, " The Use of Context in Large Vocabulary Speech Recognition", PHD Dissertation of Cambridge University, 1995
- [Yu99] 余伯泉等, "台灣通用拼音方案", 中研院民族所, 1999