A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System Based on Inter-syllabic Initial-Final Modeling and Lexicon-Tree Search

Ren-yuan Lyu[1], Yuang-jin Chiang[2], Ren-zhou Fang[2], Wen-ping Hsieh[2]

[1]Chang Gung University   [2]National Tsing Hua University

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

Tel: 886-3-3283016#5677

Abstract

In this paper some preliminary work about Taiwanese (Min-nan) speech recognition research has been done and described. Also, we report some pioneer experimental results on an initial study about a large-vocabulary (with 20 thousand words) Taiwanese multi-syllabic word recognition system. For the speaker dependent case, 9.4% word error rate is achieved. A real-time prototype system implemented on a Pentium-II personal computer running MS-Windows95/NT is also shown to validate the approaches proposed in this paper.

1. Introduction

Taiwanese, one of the major Chinese dialects, 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-Hokkian), 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, according to the 1996 Ethnology. In the past few decades, scientists in Taiwan do speech recognition research on Mandarin speech. Some achievements have been achieved in recent years.[Lyu95] Since Taiwanese is another major language spoken in this land, and Taiwan is basically a multilingual society, we decided to develop a similar large-vocabulary speech recognition system for Taiwanese speech.

In this paper, some preliminary work has been done, including the study of Taiwanese phonetics, setting up a Taiwanese 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.

The basic technology adopted here is the continuous Hidden Markov Model (CHMM) because of its success in speech recognition in the past decades. We adopt CHMM to model the Taiwanese Initial/Final phonetic units, considering both the inside- and inter-syllabic coarticulation. A promising result, with the error rate being 9.4%, for the speaker dependent case was obtained.

Additionally, a real-time prototype system in a Pentium-II personal computer running MS-Windows95/NT was implemented for further study and to validate the approaches we proposed here.

This paper is organized as follows: First, the Taiwanese Phonetics was summarized in section 2. The scripts, lexicon, and database were described in section 3. The front-end signal processing was then described in section 4. Section 5 was about the selection of the speech units. Section 6 was about the search in the lexicon tree. In section 7, we reported the results of the experiments and some discussions. A prototype system was shown and several concluding remarks were finally given in section 8.

2. Taiwanese Phonetics

Taiwanese, like Mandarin as a member of Sino-Tibetan language family, is a tonal, monosyllabic language. Traditionally, a Taiwanese syllable is decomposed into three parts, namely an Initial, a Final and a tone. Take the syllable "dan4" (to wait) as an example, where /d/ is an Initial, /an/ is a Final, and /4/ represents a high-falling tone. There are 18 Initials (including one null Initial), 47 Finals, and 7 lexical tones in Taiwanese. [Wang54] [Cheng97] An Initial is equivalent to a consonant, but a Final can be further decomposed into 1 to 3 vowels plus possible consonants. In particular, for each Final, there is a corresponding "entering-tone" Final, which is ended with an unreleased /p/, /t/, /k/ or /h/. All the phonemes are listed in <table.1>, each of which has a corresponding Chinese character with that phoneme as part of its pronunciation. The phonemes are represented by 3 alternative symbolic systems, including the International Phonetic Alphabet (IPA), the Mandarin Phonetic Alphabet (MPA), and a set of specially designed phonetic alphabet called Daiim, where Daiim is especially convenient to encode Taiwanese speech, and is adopted in the following parts of this paper.[Chiang94] All the Initials and Finals are thus listed in <table.2> and

<table.3> with their corresponding Chinese characters and Daiim representations. (Note that: Some Taiwanese syllables used in daily conversation have no widely accepted Chinese characters, and we use "○" to represent such syllables).

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) and the other five are carried in those not ending with final /p, t, k, h/ (called non-entering tone). An example of these 7 tones with one corresponding Chinese character for each tone is shown in <table.4>. [Chinag97] Some acoustic characteristics, including the waveform, the contour of fundamental frequency, the description of relative frequency level, and the traditional phonological order [Zhou97][Cheng97] are also shown in <table.4>. The tone sandhi issue is even more complex and beyond the discussion of this paper.

Since the task we are considering here is the recognition of multi-syllabic words, which have relatively few homonyms even when the tones are disregarded. In this initial study, we decided not to deal with the issues of tones and then reduce the 1683 phonologically allowed tonal syllables to 714 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.

3. Training Script, Lexicon, Testing Script and Database

To initiate the study of the large-vocabulary speech recognition of a new language, like Taiwanese we are studying here, one of the most important preliminary jobs is to construct a pronunciation lexicon. For this, we have set up a Taiwanese pronunciation lexicon of about 20 thousand words, each of them has a corresponding string of phonetic symbols encoded in Daiim phonetic alphabet. [Chiang94] In this lexicon, there are 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 much phonetic variety 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 1029 distinct Right-Context-Dependent (RCD) phonemes found in the lexicon were selected. In addition, several extended sets of words, which contain as many distinct RCD phonemes as possible, were also selected to enhance the phonetic variety. Furthermore, a set of single-syllabic words, containing all 2874 phonologically possible syllables, was picked out, too. The statistics of all the sets of words used in the training session is listed in <table.5>.

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.6>.

The speech database 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.5> and <table.6>.

4. 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-dimentional log energy was extracted to form a 13-dimentional 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-dimentional 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.

5. Selection of Speech Units

In this paper, we adopted Initial-Final's, 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 and all Finals are right context independent. There are thus 147 RCD Initial models and 77 CI Final models.(Note that there seem to be 47*2=94 Finals by observing <table.3>. But since some of them do not exist in the lexicon, they are not chosen in our training script and not used as the basic speech units.) 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 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 following experiments, such an explicit consideration about the inter-syllabic coarticulation does decrease the word error rate at a little cost of additional computation.

6. 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 roughly estimate of speed improvement is more than 10 times! A sub-tree is shown in <fig.1>. A widely used Viterbi beam search is then used to find N best paths and then the N candidates of the recognized words. [Lee89]

7. Experiments and Discussions

The experimental results for the testing corpus are listed in <table 7>. The word error rates we achieved in this initial study in average for 2 speakers are 11.4% for inside-syllable modeling and 9.4% for inter-syllable modeling. The speed for each case is approximately the same.

From <table.7>, it is observed that the word error rate is lower when the average length of each word is longer. Also one can observe that the average length of words in the 4 testing sets is very close to the average length of words in the whole 20K-lexicon, as shown in <table.6>, where 2.49 and 2.52 syllables per word for the 4 testing sets and the whole lexicon respectively. It is thus safe to claim the recognition rate for the testing sets can represent well the recognition rate for the whole 20K-lexicon.

It is so surprising to observe that there is almost no increase in computation when we added 105 additional inter-syllabic RCD units! The reason is because the width of the beam of the Viterbi search was set to be constant, and thus there was almost the same number of states active in each forward calculation.

8. A Prototype System and Concluding Remarks

To validate the approaches proposed in this paper, a prototype system was implemented on a Pentium-II personal computer running MS-Windows95/NT. The block diagram of the system is shown as in <fig 2>, and the graphic user interface (GUI) is shown as in <fig.3>.

Compared with the speech recognition systems for the major languages in the world, such as English or Mandarin, the Taiwanese speech recognition research is still in the baby stage. However, since Taiwan is famous with her computer industry, and Taiwanese is so popular in Taiwan, we hope there are more and more researchers in Taiwan devote themselves in the study of this language. I hope in some day my old grandmother can talk to the computer in Taiwanese, which is the only one language for her to communicate.

144

9. Reference

[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

[Wang57]   王育德, "台灣語常用語彙", 永和語學社, 1957.

[Chinag94]   江永進, "台音式輸入法 version4.1", 臺灣新竹清華大學統計所, 1994

[Chiang97]   許世楷等, 江永進執筆, "口語調自然調形", 台灣世界 12 期, 台中市,1997

[Zhou97]   周長揖 康啓明 "台灣閩南語教程", 1997

[Cheng97]   Robert L. Cheng, "Taiwanese and Mandarin Structures and Their Developmental Trends in Taiwan--I: Taiwanese Phonology and Morphology", 1997

[Lee89]   C.H. Lee, etc, " A frame-synchronous network search algorithm for connected Word recognition", IEEE Trans. ASSP, pp. 1649-1658, Nov. 1989

10. Tables and Figures

<table.1>   A List of Phonemes in Taiwanese

Consonant				Consonant				Vowel			
IPA	Chinese Character	MPA	Daiim	IPA	Chinese Character	MPA	Daiim	IPA	Chinese Character	MPA	Daiim
p	保	ㄅ	b	ts	資	ㄗ	z	a	阿	ㄚ	a
p'	坡	ㄆ	p	ts'	此	ㄘ	c	i	伊	一	i
m	冒	ㄇ	m	s	思	ㄙ	s	u	有	ㄨ	u
b	帽		v	z	如		r	ɛ	鞋	ㄝ	e
t	刀	ㄉ	d	x	好	ㄏ	h	ɔ	烏	ㄛ	o
t'	討	ㄊ	t	ø	英			ə	蚵	ㄜ	or
n	怒	ㄋ	n					ã	餡		ann
l	路	ㄌ	l					ĩ	嬰		inn
k	糕	ㄍ	g					ũ	樣		unn
k'	科	ㄎ	k					ɛ̃	嬰		ehn
ŋ	雅	ㄥ	ng					ɔ̃	惡		onn
g	鵝		q								

IPA: the International Phonetic Alphabet

MPA: the Mandarin Phonetic Alphabet widely used in Taiwan

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

<table.2>   18 Initials of Taiwanese syllables.

	Chinese Character	Daiim		Chinese Character	Daiim
1.	保	b	10.	科	k
2.	坡	p	11.	雅	nq
3.	冒	m	12.	鵝	q
4.	帽	v	13.	資	z
5.	刀	d	14.	此	c
6.	討	t	15.	思	s
7.	怒	n	16.	如	r
8.	路	l	17.	好	h
9.	糕	g	18.	英	(null initial)

<table.3> 47 finals and their counterparts for entering-tone in Taiwanese

	Chinese Character	Dai-im	Chinese Character	Dai-im (entering-tone)		Chinese Character	Dai-im	Chinese Character	Dai-im (entering-tone)
1.	阿	a	鴨	ah	25.	鴦	iunn	◯	iunnh
2.	會	e	窄	eh	26.	妙	iaunn	◯	iaunnh
3.	伊	i	裂	ih	27.	碗	uann	◯	uannh
4.	烏	o	◯	oh	28.	妹	uenn	◯	uennh
5.	蚵	or	學	orh	29.	黃	uinn	◯	uinnh
6.	有	u	◯	uh	30.	橫	uainn	◯	uainnh
7.	愛	ai	◯	aih	31.	姆	m	◯	mh
8.	後	au	◯	auh	32.	秧	ng	◯	ngh
9.	野	ia	頁	iah	33.	暗	am	盒	ap
10	腰	ior	藥	iorh	34.	安	an	扎	at
11	優	iu	◯	iuh	35.	紅	ang	沃	ak
12	邀	iau	◯	iauh	36.	蔘	om	◯	op
13	娃	ua	活	uah	37.	汪	ong	惡	ok
14	話	ue	喂	ueh	38.	音	im	立	ip
15	威	ui	挖	uih	39.	因	in	一	it
16	歪	uai	◯	uaih	40.	英	ing	益	ik
17	餡	ann	◯	annh	41.	鹽	iam	葉	iap
18	嬰	enn	脈	ennh	42.	煙	en	拽	et
19	院	inn	物	innh	43.	央	iang	◯	iak
20	惡	onn	膜	onnh	44.	勇	iong	育	iok
21	哼	ainn	◯	ainnh	45.	溫	un	熨	ut
22	貌	aunn	◯	aunnh	46.	彎	uan	越	uat
23	影	iann	◯	iannh	47.	◯	uang	◯	uak
24	薑	ionn	◯	ionnh					

Note: "◯" represents the syllable which has no widely accepted Chinese character

<table.4> The 7 lexical tones of Taiwanese

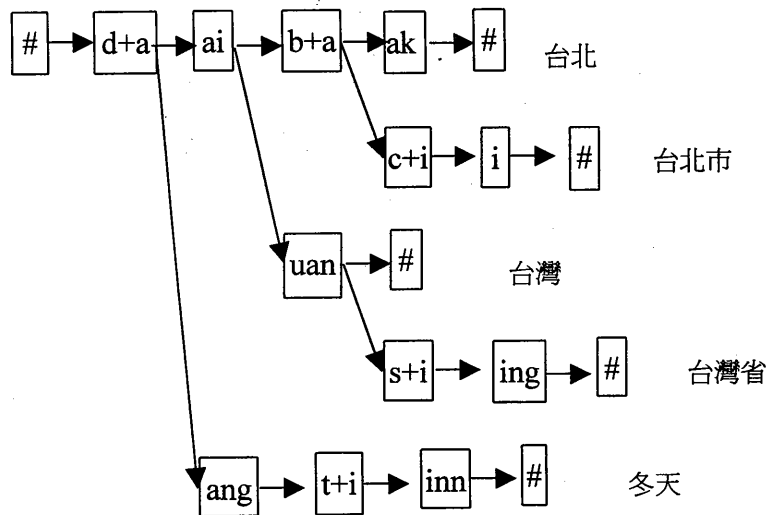| 漢字	東	洞	棟	黨	同	獨	督
Waveform							
Fundamental Frequency							
Relative Frequency	High level	Mid Level	Low falling	High falling	Rising	High Stop	Low stop
Traditional Tone order	1	7	3	2	5	8	4

147

<table.5> Statistics of the Lexicon and the Training Word Sets

		Number of words	Number of distinct RCD phonemes	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
Lexicon		19,152	1,029	N/A	N/A

<table.6> 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
Lexicon		19,152	2.52	N/A	N/A

<fig.1> A sub-tree of the lexicon tree

→ d+a → ai → b+a → ak → # 台北

c+i → i → # 台北市

uan → # 台灣

s+i → ing → # 台灣省

ang → t+i → inn → # 冬天

148

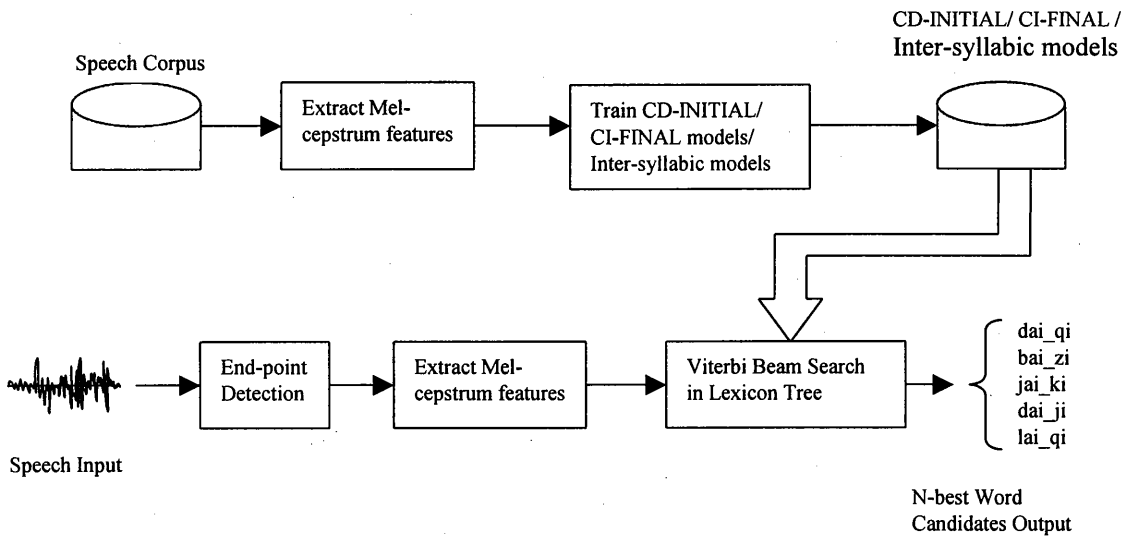<table. 7> Experimental results for the large-vocabulary Taiwanese word Recognition by using Inside-syllable modeling v.s. Inter-syllable modeling

	Inside-syllable modeling		+ Inter-syllable modeling	
	Error Rate%	CPU time	Error Rate%	CPU time
R1000	9.5	3.19 x realtimes	7.2	3.20 x realtimes
H500	16.8		13.8	
N407	13.8		12.1	
P396	6.9		6.4	
Average	11.4	3.19 x realtimes	9.4	3.20 x realtimes

<fig 2> A prototype system running on MS-WindowNT

<fig.3> The GUI of the prototype system implemented in MS-Windows95/NT