A Bi-lingual Mandarin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Tong-yong Phonetic Alphabet (TYPA)

Ren-yuan Lyu [1], Chi-yu Chen [1], Yuang-chin Chiang [2], Min-shung Liang [1]
[1.] Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2.] Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
Email: rylyu@mail.cgu.edu.tw, rylyu@ms1.hinet.net; Tel: 886-3-3283016ext5677

Abstract

In this paper, we describe the first Mandarin/Taiwanese (Min-nan) bi-lingual, continuous speech recognition system for large vocabulary or vocabulary-independent applications. A phonetic transcription system called Tong-yong Phonetic Alphabet (TYPA) is described and used to transcribe the bi-lingual Mandarin/Taiwanese lexicons. The Right-Context-Dependent (RCD) phonetic continuous-density Hidden Markov Models (CHMM) based on TYPA are used as the acoustic models. A lexicon tree containing 40 thousand bi-lingual words is used as a searching net to evaluate the performance of the recognizer. A 92.55% word accuracy is achieved on a speaker dependent case. Furthermore, we construct a continuous-speech real-time demonstration system based on the vocabulary-independent RCD models for a specific application domain of automated hospital appointment arrangement, where Mandarin/Taiwanese mixed speech is very possible to happen.

1. Introduction

Mandarin is the most important language in Chinese societies, it has been widely studied in the field of speech recognition for decades. [8] Recently, it has received even more attention because of its huge population and the potentially huge market. However, most people in Chinese societies speak at least 2 languages in their daily lives, i.e., Mandarin and their mother tongues. including Taiwanese (Min-nan, Southern Hokkianese), Hakka, Yue (Cantonese), Wu (Shanhaiese),....etc. To make the future speech interface more friendly, the bi-lingual or even multi-lingual capability of the speech recognizer is highly desired.

Taiwanese is the mother tongue of more than 75% of the population in Taiwan. It belongs to a even larger Chinese dialectical family called Min-nan, which is not only used in Fu-jian Province of China but also used by many overseas Chinese living in Singapore, Malaysia, Philippine, Indonesia and the other areas of Southern-East Asia, one of the fastest economically growing areas in the world. Since 1997, we have conducted a series of projects about Taiwanese speech recognition [9] and text-to-speech. In this paper, we report the recent progress about our efforts towards multilingual speech recognition, including Mandarin and Taiwanese.

This paper is organized as follows: section 2 describes pronunciation dictionaries and the Tong-yong Phonetic Alphabet (TYPA), which is the startup of our research; section 3 is about the unit selection and speech corpus; section 4 summarizes the speech recognition technologies used here; section 5 reports the experimental setup and result; section 6 describes an application system of a specific domain; finally, section 7 is a conclusion.

2. The Pronunciation Dictionaries and the TYPA

One of the preliminary jobs to construct a large-vocabulary speech recognition system is to construct a pronunciation lexicon. We have set up both a Taiwanese and a Mandarin pronunciation lexicon of more than 60 thousand (60K) words for each language. Each lexicon item has a corresponding string of phonetic symbols encoded in TYPA, which will be described in the following paragraph.

Many symbolic systems have been developed for labelling the sounds of all languages in the world. One of the most popular systems is the International Phonetic Alphabet (IPA) developed originally more than 100 years ago. However, since many IPA symbols are not defined in the ASCII code set and are hard to manipulate in modern digital computers, many ASCII-coded IPA symbolic sets have been proposed recently. Two popular systems are SAMPA [2] and WorldBet [3]. It's claimed that one can select parts of these phone sets for his own language. However, since these systems are designed for all languages in the world. They may be too detail or too complex to be used for some local languages like what will be addressed here. In the other hands, the Mandarin Phonetic Alphabet (MPA, Zhu-in-fu-hao) and Pinyin (Han-yu-pin-yin) have been officially used in Taiwan and Mainland China respectively for a long history. They are good in the syllabic level for Mandarin, but insufficient for the other Chinese dialects like Taiwanese or Hakka, also spoken by many inhabitants in Taiwan. To begin with the multilingual speech data collection and labeling, it's necessary to design a more suitable phoneme set which is at least easier to learn.

It is well known that phonemes can be divided almost arbitrarily, depending on the level of detail desired. The labeling philosophy here adopted is that when faced with the choice, we prefer not to divide a phoneme into distinct allophones, except in cases when the sound is clearly different to the ear or the spectrogram to the eye. [5] Since the labeling is often performed by engineering students and researchers (as opposed to professional phoneticians), it is generally safer to keep the number of units as small as possible, assuming that the recognizer will be able to learn any finer distinctions that might exist within any context.

The whole phone set for 3 major languages used in Taiwan are listed in <table.1>. Some comments and abbreviations about the table are listed as follows: There are 5 broad classes of phoemes, i.e., Consonant (***C***), Vowel (***V***), Null Vowel (***Nu***), Nasalized Vowel (***Na***), and Stop Vowel (***St***).In additional to the ***TYPA***, we list here several other symbolic systems for comparison, including (1) Mandarin Phonetic Alphabet (***MPA***); (2) ***Pinyin***, also called Hanyu-pinyin; (3) International Phonetic Alphabet (***IPA***); (4) WorldBet (***WB***); (5) SAMPA-T (***SP-T***) [1].

Some symbolic systems are designed only for Mandarin (***M***), e.g. MPA and Pinyin; others are suitable for all languages in Taiwan, including Mandarin, Taiwanese and Hakka (***MTH***),

e.g., IPA, SP-T and TYPA. For each phone in TYPA , there is an example Chinese (***Ex***) character with the pronunciation as a Mandarin syllable or Taiwanese syllable.

In TYPA, there are several major differencies other than the others. These differencies include:

(1) Pinyin's discriminative pairs like "z,j" , "c,q", and "s,x" are merged into one for each pairs, denoted in TYPA as "z", "c", and "s", respectively. This is because the right context of the phone in each discriminative pair will never be the same, e.g., Pinyin's "z", "c", and "s" can only be followed by Pinyin's "a", "u", "e", "ê"; while Pinyin's "j", "q", and "x" can only be followed by Pinyin's "i" and "ü". Since almost all speech recognition systems, including ours, use context dependent phones as their basic acoustic units, to discriminate "z" and "j" is unnecessary in the mono-phonetic level.

(2) Pinyin's "e", "ê", and "er" are mapped to TYPA's "er", "e" and "err" because "ê" is not in the range of ASCII code and hard to manipulate in computers. Although TYPA's "err" has 3 symbols for one phone at the first glance, this phone is always standing alone as a syllable, i.e., it is never concatenated in its left or right context to form a syllable with more than 3 symbols.

(3) Pinyin discriminates the pairs of vowels and its semi-vowel counterparts "i,y", "u, w" and "ü, yu". In TYPA, we integrate them into one for each pair, denoted as "i", "u" and "yu", respectively.

(4) Pinyin denotes the Null Vowel as "i", like "**資** zi", which is confused with the vowel "i", like "**基** ji". In TYPA, we design a distinct symbol "ii" to denote the Null Vowel.

(5) In Taiwanese and Hakka, there are additional Nasalized Vowels and Stop Vowels, which do not exist in Mandarin. For such phones, the postfixes -nn, -p, -t,-k, -h are tailed in normal vowels, e.g., "ann", "ap", "at", "ak" and "ah".

(6) Motivated by Taiwanese's specific Nasalized Vowels, it's suggested that Mandarin's vowels buried in nasalized syllables are transcribed to Nasalized Vowels, such as Pinyin's "**媽** ma", "**那** na", or even "**他** ta" being transcribed as TYPA's "mann","nann", and "tann". Note here that in Taiwan the syllable "**他** ta" is for some reason nasalized by almost all people with Taiwanese as their mother tongues.

(7) Mandarin's syllables with the 5^{th} tone (so called neutral tone) are suggested to be transcribed to the "-h" Stop Vowels, e.g., Pinyin's "**的** de" being transcribed as TYPA's "derh"

3. Unit Selection and Speech Corpus Construction

Based on the two pronunciation dictionaries transcribed in TYPA, we extract the sets of distinct syllables for the two languages. We get a set of 407 base syllables for Mandarin and a set of 825 base syllables for Taiwanese. Between them, there are 183 common base syllables for both languages, and thus a set of 1049 distinct base syllables are obtained when considering both languages simultaneously. It is well known that when considering large-vocabulary, continuous speech recognition using Hidden Markov Models (HMM), syllables are too large to be well modeled due to insufficient data and coarticulation effect. Therefore, phone-like units are widely used in previous studies. We further decompose the sets of base syllables into sets of phones and sets of inside-syllabic right context dependent (RCD) phones. There are 36 phones for Mandarin and 49 phones for Taiwanese. A set of 60 phones are necessary to transcribe both languages. Similarly, there are 186 RCD phones for Mandarin, 444 RCD phones for Taiwanese. A set of 529 RCD phones are necessary for both languages. All the statistics of the phonetic units considered here are listed in <table.2>. The set of phones for both languages are also listed in <table.3>

Since all units considered here are derived from the sets of base-syllables, it seems enough to collect the speech database composed of all possible base-syllables uttered in an isolated mode. However, at least 2 reasons make us do more than this. One reason is that to utter speech in an isolated syllabic mode is very unnatural and boring for most speakers. The other reason is that to collect speech data in such a way will loose much information about coarticulation effect of continuous speech. Therefore, in addition to all base-syllables, phonetically abundant word sets are extracted from the pronunciation dictionaries for collecting speech corpus. The requirements for selecting such a word set is that it should contain all possible phonetic units to be used in HMM and keep the size of the set as small as possible. However, to make the speech database as useful as possible for future research, the phonetic units include not only base-syllables, phones, and RCD phones mentioned in this paper, but also Initial-Finals, RCD Initial-Finals, tonal-syllables, and even the inter-syllabic RCD phones. The selection of such a word set is actually a set-covering optimization problem [6], which is NP-hard. Here we adopt a simple greedy heuristic approximate solution as described in [7]

A growing set of words is obtained as we put the requirements of the word set to cover the following phonetic units sequentially: (1) phones; (2) Initial-Finals; (3) RCD phones; (4) RCD Initial-Finals (5) Base-syllables; (6) Tonal syllables; (7) Inter-syllabic RCD phones. As shown in <table.4>, one can see that a set of only 19 words is enough to cover all phones, and a set of 3515 words can cover all 7 different categories of phonetic units.

4. Phonetic Modeling and Word Tree Searching

We adopted one of the most popular speech recognition schemes based on the RCD phonetic CHMM with the following configuration:

- 16KHz sampled microphone speech
- 16 ms Hamming windowing with 8 ms window shifting
- 26-dimentional feature vector, including 12-dimentional MFCC, 1-dimentional energy, 12-dimentional delta MFCC, and 1-dimentional delta energy
- Inside-syllabic RCD phonetic CHMM with 3 states and 3 mixtures
- 444 models for Taiwanese; 186 models for Mandarin and 529 models for both languages.

The searching space is constrained by a word tree which uses base-syllables as the nodes as shown in <fig.1>. Each base-syllable is directly concatenated with its corresponding RCD phonetic CHMM's.There are 90496 nodes, 128877 links for 40K T/M bilingual word tree, 49666 nodes, 69664 links for 20K Mandarin word tree, and 41736 nodes, 60393 links for 20K Taiwanese word tree respectively.

5. The Experimental Setup and Word Accuracy

For the initial study, we have a specific speaker who is fluent in both Mandarin and Taiwanese and records a large bilingual speech database, including isolated syllables and the phonetically abundant word set. The statistics of the speech

corpus is listed in <table.5> The testing data is a set of 2000 Mandarin/Taiwanese words randomly selected from 40K-word Mandarin/Taiwanese lexicon.

When Taiwanes specific models are trained only using Taiwanese training speech, the word accuracy for Taiwanese testing speech is 95.98%. Similarly, when Mandarin models are used to recognize Mandarin speech, 94.67% word accuracy is achieved.

It's possible to use Mandarin models to recognize Taiwanese speech and the bilingual speech, if a little Taiwanese speech is available. This is meaningful because there are a large quantity of speech corpus distributed in many places for the mainstream languages like Mandarin or English. However, languages used by fewer people have little speech database available. Therefore, many researchers try to adapt acoustic models of one language to those of a new target language.

We also made an initial try of language adaptation as follows: first, using Mandarin models trained by Mandarin speech as the kernel acoustic models; second, borrowing some Taiwanese specific models from Taiwanese models trained by Taiwanese speech. This preliminary approach achieves 85.06% word accuracy for Taiwanese and 89.26% for bilingual speech. Although there are many other more reasonable approaches, we do not trapped ourselves here too long at the present time but leave it in future research.

Finally, all speech data for both languages are used to train a bilingual Mandarin/Taiwanese global phone set, a word accuracy 92.55% can be achieved, which can be a good start for further research in future. All the experimental results are listed in <table.6>.

6. An Application to a specific domain

In this section, we describe a real-time system of a specific domain based on the bilingual Mandarin/Taiwanese, vocabulary independent, RCD phonetic CHMM's described previously. The system can be applied to be a speech driven scheduling application which will allow patients to arrange appointments by speaking to an automated virtual attendant. The speech queries are constrained by a grammar net as shown in <fig.2>, where it is allowed to speak a sentence containing the information about the names of departments and doctors, the date, the time and the register number of personal anamnesis. The queries can be in Mandarin/Taiwanese bi-lingually mixed speech. The prototype system is now constructed in a Pentium!!!-500 PC and the real-time response with a word accuracy more than 95% achieved. The user interface of the system is shown in <fig.3>. The telephone speech version of this system is now under construction and will be field-tested in Chang Gung Memorial Hospital, which is the largest hospital in Taiwan.

7. Conclusions

In this paper, we have described the first Mandarin/Taiwanese bi-lingual, continuous speech recognition system for large vocabulary based on the TYPA, a phonetic transcription system specially designed to be suitable for all languages used in Taiwan. At present, only speaker-dependent experiments have been conducted and reported due to the lack of speaker independent speech database. In the near future, when an under-constructing multi-speaker speech database becomes available, more experiments should be performed to test the performance of the system. Furthermore, there are recently many papers talking about language adaptation, which use speech corpus of main-stream languages such as English or Mandarin to train language independent acoustic models and adapt them to any specific, local languages. Such a trend will be noticed in our future work.

Refference

[1] Chiu-yu Tseng, Fu-chiang Chou, " Machine Readable Phonetic Transcription System for Chinese Dialects Spoken in Taiwan", Lec4.1~Lec4.8, the 1998 Symposium on Speech Signal Processing, May, 1998, Hsinchu, Taiwan

[2] J.C. Wells, " Computer-coding the IPA: a proposed extension of SAMPA", http:// www. phon. ucl. ac. uk / home / sampa / ipasam-x.pdf

[3] J.L.Hieronymus, "ASCII Phonetic Symbols for the World's Language: Worldbet", AT&T Bell Labs, Technical Report, 1994

[4] Frank Seide, Nick Wang, " Phonetic Modelling in the Philip Chinese Continuous-Speech Recognition System", p.54~p.59, The 1998 International Symposium on Chinese Spoken Language Processing, Dec, 1998, Singapore

[5] http:// webserver. pue. udlap. mx/ ~sistemas/ tlatoa/ documentation/ labels.html

[6] Jia-lin Shen, Hsin-min Wang, Ren-yuan Lyu, and Lin-shan Lee, "Automatic Selection of Phonetically Distributed Sentence Sets for Speaker Adaptation With Application to Large Vocabualry Mandarin Speech Recognition", ***Computer Speech and Language***, vol. 13, no. 1, pp. 79-97, Jan. 1999.

[7] T.H Cormen, etc, "Chapter 37: Approximation Algorithms", *Introduction to Algorithms*, p.974~p.978

[8] Lin-shan Lee, "Voice Dictation of Mandarin Chinese", IEEE Signal Processing Magzine, July, 1997,p.63~p.101

[9] Ren-yuan Lyu, Yuang-chin Chiang, etc, "A Large-Vocabulary Speech Recognition System for Taiwanese (Min-nan)", *Journal of the Chinese Institute of Electrical Engineering*, Vol7, No.2, p.123~p136, May, 2000

Tables and Figures

<table.1> Tong-yong Phonetic Alphabet (**TYPA**) and others

		MPA	***Pinyin***	***IPA***	***WB***	***SP-T***	***TYPA***			
		M	***M***	***MTH***	***M***	***MTH***	***M***	***T***	***H***	***Ex***
C	1	ㄅ	b	p	p	b	b	b	b	八 ba
	2	ㄆ	p	p'	ph	p	p	p	p	趴 pa
	3	ㄇ	m	m	m	m m^	m	m	m	媽 ma 梅 m
	4	ㄈ	f	f	f	f	f		f	發 fa
	5	ㄉ	d	t	t[	d	d	d	d	搭 da
	6	ㄊ	t	t'	t[h	t	t	t	t	他 ta
	7	ㄋ	n	n	n	n	n	n	n	那 na
	8	ㄌ	l	l	l	l	l	l	l	拉 la
	9	ㄍ	g	k	k	g	g	g	g	嘎 ga
	10	ㄎ	k	k'	kh	k	k	k	k	喀 ka
	11	ㄏ	h	x	x	h	h	h	h	哈 ha
C	12	ㄓ	zh	tå	thsr	dz`	zh		zh	渣 zha
	13	ㄔ	ch	tå'	tsr	ts`	ch		ch	差 cha
	14	ㄕ	sh	å	sr	s`	sh		sh	殺 sha
	15	ㄖ	r	û	r+	z`	rh		rh	然 rhan

	16	ㄗ ㄐ	z j	ts t¦	ts cC	dz dz\	z	z	z	匝 za 機 zi
	17	ㄘ ㄑ	c q	ts' t¦'	tsh cCh	ts ts\	c	c	c	擦 ca 七 ci
	18	ㄙ ㄒ	s x	s ¦	s C	s s\	s	s	s	撒 sa, 西 si
	19			z Í		DZ		r		如 ru 字 ri
	20			b		B		v	v	肉 va
	21			g		G		q		我 qua
	22	ㄤ	-ng	k	N	N J n^		ng	ng	骯 ang 雅 nga 黃 ng
V	23	ㄚ	a	a	@ A	a	a	a	a	阿 a
	24	ㄛ	o	Ë	>	o	o	o	o	喔 o
	25			o				or		蚵 or
	26	ㄜ	e	Ă	2	@	er	er	er	鵝 er
	27	ㄝ	ê	b	E	e	e	e	e	也 ie
	28	ㄦ	er	Ē	&r	@`	err			而 err
	29	ㄧ	y, i	§	j I i:	i	i	i	i	一 i 金 zin 機 zi
	30	ㄨ	w u	u	w & u	u	u	u	u	吳 u 蹲 dun 度 du
	31	ㄩ	yu ü	y	jw y	y	yu			原 yuan 淤 yu
Nu	32		i	ï	If 4r	U, U`	ii		ii	之 zhii 資 zii
Na	33					~		-nn		餡 ann
St	34			-p		p}		-p	-p	壓 ap
	35			-t		t}		-t	-t	握 at
	36			-k		k}		-k		沃 ak
	37			-h		?		-h	-k	鴨 ah

<table.2> The statistics of phonetic units of 2 languages

	Syllables	Phones	RCD Phones
M	407	36	186
T	825	49	444
M∪T	1049	60	529
M∩T	183	25	101

M: Mandarin; T: Taiwanese; ∪: union; ∩: intersection

<table.3> 60 M/T Bilingual Phones

a	ah	ak	ann	annh	annp	ap	at	b	c
ch	d	e	eh	enn	ennh	er	erh	err	et
f	g	h	i	ih	ii	iih	ik	inn	innh
ip	it	k	l	m	n	ng	ngh	nh	o
oh	ok	onn	onnh	op	p	q	r	rh	s
sh	t	u	uh	unn	ut	v	yu	z	zh

<table.4> The numbers of words selected sequentially to cover all 7 different categories of phonetic units

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No. of Words	19	187	409	440	886	1847	3515

<table.5> The statistics of the bilingual speech corpus

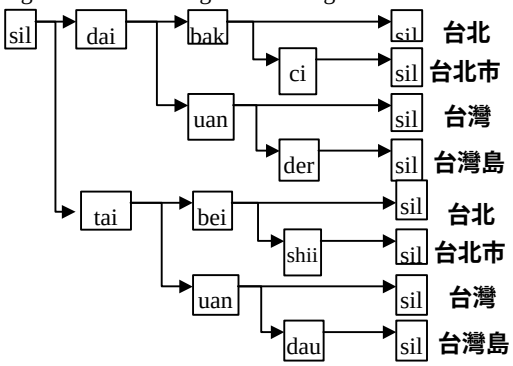
	Lang	Script	Speech Length (min)
Train	T	Syll	91.35
		Word	134.65
	M	Syll	26.95
		Word	94.43
	M/T	Total	411.73
Test	T	Word	1000 words
	M	Word	1000 words

Syll: The isolated syllables; Word: The phonetically abundant word set

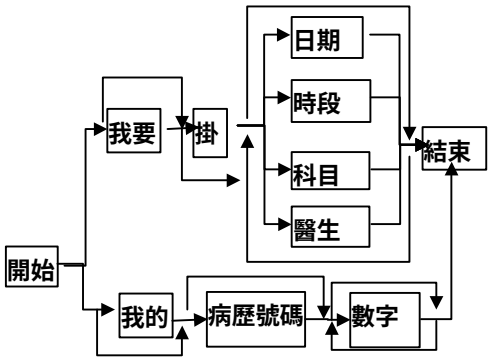
<table.6> The word accuracy (%) of the bilingual Mandarin/Taiwanese speech recognition

		Testing speech		
		T	M	M/T
Trained models	T	95.98		
	M	85.06	94.67	89.26
	M/T			92.55

<fig.1 > The searching net of bilingual word tree



<fig.2> the finite-state grammar for a specific domain



<fig.3> The user interface of the system