MODELING CONTEXT-DEPENDENT PHONETIC UNITS IN A CONTINUOUS SPEECH RECOGNITION SYSTEM FOR MANDARIN CHINESE

Jim Jian-Xiong Wu[1,2], *Li Deng*[2], *and Jacky Chan*[2]

[1]Nortel Technology, 16 Place du Commerce, Nuns' Island, Verdun, Quebec, Canada H3E 1H6.
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
deng@crg5.uwaterloo.ca

ABSTRACT

We study the problem of phonetic modeling for continuous Mandarin speech recognition by providing a systematic performance comparison for systems based on following primitive speech units: syllable, demi-syllable (Initials and Finals), context-independent phones, left-or-right context-dependent phones (diphones), and left-and-right context-dependent phones (triphones). In our speaker-dependent continuous speech recognition experiments, a generalized triphone system has achieved the best performance among all. Our best system contrasts most other Mandarin speech recognition systems which have been based on demi-syllable units.

1. INTRODUCTION

Despite sizable efforts in the past devoted to developing dictation machines for Mandarin Chinese [4][5][6], the research on continuous speech recognition for the Chinese language is only of recent interest[8][3]. Most of the existing systems use demi-syllable units (Initials and Finals) as the primitive speech units for acoustic modeling. Since the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances, it is important to study to what degree the modeling of context-dependent phonetic units, which has been demonstrated to be highly successful for English speech recognition [2, 9], but largely ignored in Chinese speech recognition research, is effective for Mandarin continuous speech recognition. The issue of modeling unit selection is particularly important for speaker-dependent recognition (reported in this paper) because the variabilities in the speech data are largely attributed to contextual factors (vs. speaker variations). This importance is further accentuated by the fact that modeling unit selection is closed linked to the requirement of training data, which are necessarily sparse for speaker-dependent recognition because of the practical difficulty of collecting a large amount of training data for each speaker.

In this work, we will report a systematic study on how the performance of a speaker-dependent Mandarin continuous speech recognizer is effected by the amount of contextual information utilized in the acoustic modeling. In particular, we will compare recognition performances of various systems we developed which are based on the following primitive speech units: syllable, demi-syllable, context-independent phones, left-or-right context dependent allophones (diphones), left-and-right context dependent allophones (triphones) under various training conditions. Our experimental results obtained so far indicate that generalized triphone units are most effective, among the types of units studied, for speaker-dependent Mandarin continuous speech recognition. These triphone units give the best recognition performance and the speech recognition performance based on the triphone units appears to be less sensitive to the amount of training data compared with other types of units evaluated under the otherwise identical experimental conditions.

2. BASIC PHONETIC STRUCTURE OF MANDARIN CHINESE

Mandarin is a tone language. Among the 1254 distinct syllables, there are a total of 408 toneless base-syllables and most (not all) of the base-syllables are associated with four different tones. Since the tonality of a syllable is largely characterized by the pitch pattern of the syllable and may be recognized separately, in this paper we only deal with the problem of recognizing 408 base-syllables. Traditionally, a Mandarin (base) syllable is decomposed into the Initial and Final structures according to the following rules:

$$\begin{aligned}
Syllable &\rightarrow [Initial] \quad Final \\
Initial &\rightarrow Consonant \\
Final &\rightarrow [Medial] \quad Vowel \quad [Coda] \\
Medial &\rightarrow Vowel \\
Coda &\rightarrow \{Vowel, \quad Nasal\}
\end{aligned}$$

The lists for Initials and Finals used in our experiments are provided below in terms of their corresponding Pin-Ying notations:[1]

$$\begin{aligned}
Initials: \quad & b, p, m, f, t, d, n, l, g, k, h, j, q, x, r, \\
& c, z, s, ch, sh, zh, y, w, sil. \\
Finals: \quad & a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, \\
& iang, iao, ie, in, ing, iong, iu, o, ong, ou, u, ua, \\
& uai, uan, uang, uen, ueng, ui, un, uo, \ddot{u}, \ddot{u}an, \ddot{u}e.
\end{aligned}$$

Note that we treat semi-vowels $/y/$, $/w/$ as Initials. In addition, we use a pseudo-Initial, $/sil/$, if there is no Initial in the syllable.

[1]Pin-Ying is an alphabetic writing system for Chinese characters.

We have noted from our data analysis that the Pin-Ying notation has been inconsistent with the acoustic-phonetic observations of the Mandarin Chinese speech. (The inconsistency is much less than that between the orthographic form and the phonemic form of English.) For example, acoustic-phonetic observations indicate that the same phone /eh/ should be used for both "a" in "bi**a**n" and in "e" in "bi**e**", completely different symbols in the Pin-Ying notation. Following is the list of phones we have adopted for Mandarin Chinese in our experiments:

$$\begin{aligned} Consonant: &\quad b, p, m, f, d, t, n, l, g, k, h, j, q, x, \\ &\quad c, z, s, ch, zh, sh, ng, r. \\ Vowel: &\quad I, a, e, eh, i, i1, i2, o, u, ux \\ Semivowel: &\quad w, y. \end{aligned}$$

In Appendix we give the phonetic labels for all Mandarin base syllables expressed conventionally in the Pin-Ying notation. We used these labels as the pronunciation of the syllables in our recognition system. Each of the labels we used has one-to-one correspondence to the IPA symbol.

3. RECOGNITION SYSTEM DESCRIPTION

In this study, we have built seven different recognizers for speaker-dependent Mandarin continuous speech recognition, and compared their respective performances. Each recognizer is associated with use of a distinct set of speech units. The architectures of all the seven recognizers are the same: they all use the hidden Markov model (HMM) technique for acoustic modeling; Gaussian mixtures are employed as the state-conditioned output probability distributions; the HMM states are arranged in a left-to-right, no-state-skipping topology; the segmental k-means algorithm is used for training and the Viterbi algorithm is used for decoding; and identical speech preprocessors are used to create inputs (MFCCs and delta MFCCs) to the recognizers. As a first step towards studying phonetic modeling for Mandarin speech recognition, we further limit ourselves at this time to consider only within-syllable co-articulations whenever context-dependent models are used. In this way, we have simplified our decoding algorithm for search only at the syllable level. At the expense of losing some performance gain, this simplification keeps the search space of all the recognizers to be constant and provides a fair basis for performance comparison.

Details of the seven recognizers we have built and evaluated are:

1. **Recognizer using syllable models**. Each toneless syllable is modeled by an eight state HMM.
2. **Recognizer using demi-syllable (Initial and Final) models**. Since the most important co-articulation effect within a syllable occurs from the first vowel in Final to Initial, most of the existing systems use context-independent Final models and context-dependent Initial models (which depend only on the first vowel of the following Final)[4][5][6]. In our system, each (context-independent) Final is modeled by a six-state HMM and each context-dependent Initial by a three-state HMM.
3. **Recognizer using context-independent phone models**. A three-state HMM is used for each of the phones we identified (not Pin-Yin symbols) for the Mandarin Chinese.
4. **Recognizer using left-context-dependent diphone models**. Each of left-context-dependent diphones is modeled by a three-state HMM.
5. **Recognizer using right-context-dependent diphone models**. Each of right-context-dependent diphones is modeled by a three-state HMM.
6. **Recognizer using generalized triphone models A**. Each of the generalized triphones is modeled by a three-state HMM. The middle state of the HMM is made to depend only on the center phone of the triphone context, while the left or right state depends only on the left or right contexts, respectively. This generalized triphone structure has been used in the past for classification of English phonemes[7].
7. **Recognizer using generalized triphone models B**. Each of this kind of generalized triphones is modeled by a four-state HMM. The first two states of the HMM depend only on the left context, while the last two states depend only on the right context. A similar generalized triphone structure has been used by C. Chan [1], with a different phonetic label set and including cross-syllable context-dependent models.

Both Recognizer 6 and Recognizer 7 have the ability of predicting *unseen* triphones in a straightforward manner. For example, if both triphone units /L1 C1 R1/ and /L2 C1 R2/ have been trained, Recognizer 6 will predict the model of an unseen triphone /L1 C1 R2/ by using the same first and second states as in /L1 C1 R1/ and the same third state as in /L2 C1 R2/, while Recognizer 7 will predict the model of /L1 C1 R2/ by using the same first two states as in /L1 C1 R1/ and the same last two states as in /L2 C1 R2/.

It is well known that the performance of a recognizer depends not only by the accuracy of the acoustic modeling but also by the number of free parameters in the recognition system (given a fixed amount of training data). Since our primary goal in this study is to examine the relative qualities of the various acoustic modeling approaches, we have attempted to keep the number of free parameters roughly the same in different systems by adjusting the number of Gaussian mixtures per state. Table 1 gives the number of mixtures for each of the seven recognizers described above.

System	No.States	No.Mix./state	Total No.Mix.
1	503	7	3,521
2	554	7	3,878
3	107	33	3,531
4	506	7	3,542
5	3,282	1	3,282
6	389	9	3,501
7	706	5	3,530

Table 1. Number of states, number of Gaussian mixtures per state, and the total number of Gaussian mixtures for each of the seven recognizers.

4. EXPERIMENTAL RESULTS

We use the HKU93 Chinese speech database[10] developed at University of Hong Kong in our experiments. The database includes speaker-dependent isolated syllables and continuous speech in read style from a total of 20 speakers (10 males and 10 females). It has been designed such that all Mandarin syllables in all tones are balanced and all possible phone transitions are included. The speech material is recorded in a normal office environment and is digitized at 16 KHz by a Sound-Blaster-16 DSP board in PC. The preliminary results reported in this paper are obtained using the speech material from only one female speaker. The acoustic features used in the experiments are normalized log energy, 12-order MFCCs and their first-order time derivatives. These features are calculated with a frame length 32*ms* and frame rate 10*ms*.

The test set contains 361 continuously spoken sentences which have 4781 syllables in total. To examine the effect of the amount of training data on the performances of different systems, we designed three sets of training data which are all disjoint with the test set. Table 2 gives, for each data set, the number of isolated syllables, the number of continuous sentences, the number of syllables in these sentences as well as the total number of frames available for training. Combining Table 1 and Table 2, one can see that there are roughly 102, 70 and 44 training frames per Gaussian mixture for these three data sets, respectively.

TrainSet	Iso.Syll.	Cont.Sent.	Cont.Syll.	Totframes
1	1,615	915	11,483	362,426
2	0	915	11,483	248,488
3	0	561	8,613	155,357

Table 2. Number of isolated syllables, number of continuously spoken sentences, number of continuously spoken syllables, and the total number of speech frames in three training sets used in the experiments.

Tables 3-5 show the syllable recognition performances, in terms of percent correct, percent accurate, percent substitute error, percent delete and insertion errors, for the seven recognizers described in Section 3.

Recognizer	Corr.	Acc.	Sub.	Del.	Ins.
1	82.54%	80.15%	15.77%	1.69%	2.38%
2	85.61%	83.54%	12.70%	1.69%	2.07%
3	75.28%	66.03%	23.57%	1.15%	9.24%
4	81.30%	76.18%	17.51%	1.19%	5.12%
5	85.09%	80.36%	13.60%	1.32%	4.73%
6	84.00%	79.42%	14.56%	1.44%	4.58%
7	85.69%	83.79%	12.53%	1.78%	1.90%

Table 3. Comparative syllable recognition performance using training set one, containing the greatest amount of training data among the three sets.

The results of Tables 3-5 demonstrate three interesting things. First, when there is a large amount of training data as for data set one, the conventional demi-syllable based approach (Recognizer 2) and the generalized triphone models (Recognizers 6 and 7) give about the same recognition performance. Second, as the amount of training data reduces, the generalized triphone models become superior to all other recognizers. This is likely due to the triphones' better mechanism of data sharing than other units. Third, the generalized triphone models in Recognizer 7 perform consistently better than those in Recognizer 6 in that significantly fewer insertion errors are made. Note that in interpreting the results of Tables 3-5 we have used the fact that all seven recognizers have roughly the same number of model parameters.

Recognizer	Corr.	Acc.	Sub.	Del.	Ins.
1	77.89%	69.02%	21.23%	0.88%	8.87%
2	84.48%	78.10%	14.41%	1.11%	6.38%
3	73.88%	58.00%	25.50%	0.63%	15.88%
4	80.82%	72.20%	18.18%	1.00%	8.62%
5	82.41%	68.14%	16.88%	0.71%	14.26%
6	85.19%	76.36%	13.80%	1.00%	8.83%
7	84.67%	81.36%	13.70%	1.63%	3.30%

Table 4. Comparative syllable recognition performance using training set two, about two thirds of training data of set one.

Recognizer	Corr.	Acc.	Sub.	Del.	Ins.
1	77.06%	67.22%	22.07%	0.88%	9.83%
2	80.69%	74.25%	17.95%	1.36%	6.44%
3	74.08%	57.08%	25.12%	0.79%	17.00%
4	80.03%	70.19%	19.16%	0.82%	9.83%
5	80.28%	64.42%	19.10%	0.63%	15.85%
6	82.74%	72.14%	16.38%	0.88%	10.60%
7	82.66%	78.44%	15.92%	1.42%	4.23%

Table 5. Comparative syllable recognition performance using training set three, containing the least amount of training data.

5. CONCLUSIONS

In this paper we present a systematic performance comparison among various levels of acoustic modeling for continuous Mandarin speech recognition. Although the conventional (demi-syllable) approach to Mandarin speech recognition has been based on a combination of Final-dependent Initials and context-independent Finals, we found in our experiment that the generalized triphone models can provide equal or better performance (speaker-dependent recognition evaluated so far); this is so especially when the amount of training data is small. A greater advantage of the triphone-based approach over the demisyllable-based one will be its ability for acoustic modeling of cross-syllable context-dependence. This is more difficult to achieve with demisyllable units. In our future research, more extensive evaluation experiments will be needed to further confirm the findings reported in this paper and we will investigate the impact of incorporating cross-syllable context-dependence for Mandarin continuous speech recognition.

6. REFERENCES

1. C. Chan. Personal Communication, 1995.
2. L. Deng, M. Lennig, F. Seitz and P. Mermelstein. “Large vocabulary word recognition using context-dependent allophonic hidden Markov models,” *Computer Speech and Language*, Vol.4, No.4, December, 1990, pp. 345-357.
3. H.T. Ho, H.M. Wang, L.F. Chien, K.J. Chen and L.S. Lee. “Fast and accurate continuous speech recognition for Chinese language with very large vocabulary ,” *EUROSPEECH’95*, pp. 211-214, 1995.
4. H.W. Hon, B. Yuan, Y.L. Chow, S. Narayan and K.F. Lee. “Towards large vocabulary Mandarin Chinese speech recognition ,” *ICASSP’94*, pp. I545-548, 1994.
5. Lin-Shan Lee et. al. “Golden Mandarin (II) - An improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary,” *ICASSP’93*, pp. II503-506, 1993.
6. Ren-Yuan Lyu et. al. “Golden Mandarin (III) - A user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary,” *ICASSP’95*, pp. I57-60, 1995.
7. C. Rathinavalu and L. Deng. “Use of generalized dynamic feature parameters for speech recognition,” *IEEE Trans. Speech and Audio Processing*, 1996 (to appear).
8. Hsin-min Wang et. al. “Complete recognition of continuous Mandarin speech for Chinese language with vary large vocabulary but limited training data,” *ICASSP’95*, pp. I61-64, 1995.
9. S. Young. “Large vocabulary continuous speech recognition: a review,” *Proc. IEEE Workshop on Automatic Speech Recognition*, Snowbird, Utah, 3-28, 1995.
10. Y.Q. Zu, W.X. Li and C. Chan. “HKU93 - A Putonghua Corpus,”, 1994

Acknowledgments

Discussions with Chorkin Chan and Jenny Wang on Chinese speech recognition and on Chinese phonology/phonetics are gratefully acknowledged. Part of this work was done while the second author was visiting Hong Kong University of Science and Technology 1994-1995.

Appendix: Syllable Pronunciation Dictionary

a	a	ai	a I	an	a n	ang	a ng
ao	a o	e	e	ei	e I	en	e n
eng	e ng	er	er	o	o	ou	o u
ba	b a	bai	b a I	ban	b a n	bang	b a ng
bao	b a o	bei	b e I	ben	b e n	beng	b e ng
bi	b i	bian	b y eh n	biao	b y a o	bie	b y eh
bin	b I n	bing	b I ng	bo	b o	bu	b u
pa	p a	pai	p a I	pan	p a n	pang	p a ng
pao	p a o	pei	p e I	pen	p e n	peng	p e ng
pi	p i	pian	p y eh n	piao	p y a o	pie	p y eh
pin	p I n	ping	p I ng	po	p o	pou	p o u
pu	p u	ma	m a	mai	m a I	man	m a n
mang	m a ng	mao	m a o	me	m e	mei	m e I
men	m e n	meng	m e ng	mi	m i	mian	m y eh n
miao	m y a o	mie	m y eh	min	m I n	ming	m I ng
miu	m y u	mo	m o	mou	m o u	mu	m u
fa	f a	fan	f a n	fang	f a ng	fei	f e I
fen	f e n	feng	f e ng	fo	f o	fou	f o u
fu	f u	da	d a	dai	d a I	dan	d a n
dang	d a ng	dao	d a o	de	d e	dei	d e I
den	d e n	deng	d e ng	di	d i	dia	d y a
dian	d y eh n	diao	d y a o	die	d y eh	ding	d I ng
diu	d y u	dong	d u ng	dou	d o u	du	d u
duan	d w a n	dui	d w I	dun	d w n	duo	d w o
ta	t a	tai	t a I	tan	t a n	tang	t a ng
tao	t a o	te	t e	tei	t e I	teng	t e ng
ti	t i	tian	t y eh n	tiao	t y a o	tie	t y eh
ting	t I ng	tong	t u ng	tou	t o u	tu	t u
tuan	t w a n	tui	t w I	tun	t w n	tuo	t w o
na	n a	nai	n a I	nan	n a n	nang	n a ng
nao	n a o	ne	n e	nei	n e I	nen	n e n
neng	n e ng	ni	n i	nian	n y eh n	niang	n y a ng
niao	n y a o	nie	n y eh	nin	n I n	ning	n I ng
niu	n y u	nong	n u ng	nou	n o u	nu	n u
nuan	n w a n	nue	n ux eh	nuo	n w o	nuu	n ux
la	l a	lai	l a I	lan	l a n	lang	l a ng
lao	l a o	le	l e	lei	l e I	leng	l e ng
li	l i	lia	l y a	lian	l y eh n	liang	l y a ng
liao	l y a o	lie	l y eh	lin	l I n	ling	l I ng
liu	l y u	lo	l o	long	l u ng	lou	l o u
lu	l u	luan	l w a n	lue	l ux eh	lun	l w n
luo	l w o	luu	l ux	za	z a	zai	z a I
zan	z a n	zang	z a ng	zao	z a o	ze	z e
zei	z e I	zen	z e n	zeng	z e ng	zi	z i2
zong	z u ng	zou	z o u	zu	z u	zuan	z w a n
zui	z w I	zun	z w n	zuo	z w o	ca	c a
cai	c a I	can	c a n	cang	c a ng	cao	c a o
ce	c e	cen	c e n	ceng	c e ng	ci	c i2
cong	c u ng	cou	c o u	cu	c u	cuan	c w a n
cui	c w I	cun	c w n	cuo	c w o	sa	s a
sai	s a I	san	s a n	sang	s a ng	sao	s a o
se	s e	sen	s e n	seng	s e ng	si	s i2
song	s u ng	sou	s o u	su	s u	suan	s w a n
sui	s w I	sun	s w n	suo	s w o	zha	zh a
zhai	zh a I	zhan	zh a n	zhang	zh a ng	zhao	zh a o
zhe	zh e	zhei	zh e I	zhen	zh e n	zheng	zh e ng
zhi	zh i1	zhong	zh u ng	zhou	zh o u	zhu	zh u
zhua	zh w a	zhuai	zh w a I	zhuan	zh w a n	zhuang	zh w a ng
zhui	zh w I	zhun	zh w n	zhuo	zh w o	cha	ch a
chai	ch a I	chan	ch a n	chang	ch a ng	chao	ch a o
che	ch e	chen	ch e n	cheng	ch e ng	chi	ch i1
chong	ch u ng	chou	ch o u	chu	ch u	chuai	ch w a I
chuan	ch w a n	chuang	ch w a ng	chui	ch w I	chun	ch w n
chuo	ch w o	sha	sh a	shai	sh a I	shan	sh a n
shang	sh a ng	shao	sh a o	she	sh e	shei	sh e I
shen	sh e n	sheng	sh e ng	shi	sh i1	shou	sh o u
shu	sh u	shua	sh w a	shuai	sh w a I	shuan	sh w a n
shuang	sh w a ng	shui	sh w I	shun	sh w n	shuo	sh w o
ga	g a	gai	g a I	gan	g a n	gang	g a ng
gao	g a o	ge	g e	gei	g e I	gen	g e n
geng	g e ng	gong	g u ng	gou	g o u	gu	g u
gua	g w a	guai	g w a I	guan	g w a n	guang	g w a ng
gui	g w I	gun	g w n	guo	g w o	ka	k a
kai	k a I	kan	k a n	kang	k a ng	kao	k a o
ke	k e	kei	k e I	ken	k e n	keng	k e ng
kong	k u ng	kou	k o u	ku	k u	kua	k w a
kuai	k w a I	kuan	k w a n	kuang	k w a ng	kui	k w I
kun	k w n	kuo	k w o	ha	h a	hai	h a I
han	h a n	hang	h a ng	hao	h a o	he	h e
hei	h e I	hen	h e n	heng	h e ng	hong	h u ng
hou	h o u	hu	h u	hua	h w a	huai	h w a I
huan	h w a n	huang	h w a ng	hui	h w I	hun	h w n
huo	h w o	ji	j i	jia	j y a	jian	j y eh n
jiang	j y a ng	jiao	j y a o	jie	j y eh	jin	j I n
jing	j I ng	jiong	j y u ng	jiu	j y u	ju	j ux
juan	j ux eh n	jue	j ux eh	jun	j ux n	qi	q i
qia	q y a	qian	q y eh n	qiang	q y a ng	qiao	q y a o
qie	q y eh	qin	q I n	qing	q I ng	qiong	q y u ng
qiu	q y u	qu	q ux	quan	q ux eh n	que	q ux eh
qun	q ux n	xi	x i	xia	x y a	xian	x y eh n
xiang	x y a ng	xiao	x y a o	xie	x y eh	xin	x I n
xing	x I ng	xiong	x y u ng	xiu	x y u	xu	x ux
xuan	x ux eh n	xue	x ux eh	xun	x ux n	ran	r a n
rang	r a ng	rao	r a o	re	r e	ren	r e n
reng	r e ng	ri	r i1	rong	r u ng	rou	r o u
ru	r u	rua	r w a	ruan	r w a n	rui	r w I
run	r w n	ruo	r w o	wa	u a	wai	u a I
wan	u a n	wang	u a ng	wei	u e I	wen	u e n
weng	u e ng	wo	u o	wu	u	ya	y a
yan	y eh n	yang	y a ng	yao	y a o	ye	y eh
yi	i	yin	y I n	ying	y I ng	yo	y o
yong	ux ng	you	y o u	yu	ux	yuan	ux eh n
yue	ux eh	yun	ux n				