SUB-SYLLABLE ACOUSTIC MODELING FOR CANTONESE SPEECH RECOGNITION

K. F. CHOW, Tan LEE and *P. C. CHING*

Department of Electronic Engineering,
The Chinese University of Hong Kong
Tel. +852 2609 8271, FAX: +852 2603 6868,
E-mail:{kfchow, tlee1, pcching}@ee.cuhk.edu.hk

ABSTRACT

This paper presents a pioneer study on acoustic modeling for continuous Cantonese speech recognition. It starts from the context-independent modeling of sub-syllabic units, namely INITIALs and FINALs, and then moves on to examine a number of context-dependent models that characterize intra-syllable co-articulation. The acoustic models are trained with a large database of Cantonese polysyllabic words and evaluated with a general syllable recognition task in which no lexical or grammatical constraints are incorporated. A syllable recognition accuracy of 67.68% is attained using continuous-density HMM with 4 Gaussian mixtures.

1. INTRODUCTION

Cantonese is one of the major Chinese dialects, being spoken by tens of millions of people in Hong Kong and Southern China. It is also widely used in many overseas Chinese communities. However, strategic research of Cantonese speech recognition began only a few years ago. Most of the previous works were focused on either isolated syllable recognition [1,2] or speaker-dependent continuous speech recognition [3]. The acoustic modeling for continuous Cantonese speech in speaker-independent applications is still an untouched area.

Since Cantonese is a monosyllablic language, syllable modeling seems to be the most intuitive approach for speech recognition. However, the success of this approach depends greatly on whether sufficient training data are available for each and every single syllable. To facilitate better utilization of limited training data, sub-syllable modeling is often desired [4].

In this paper, we will describe our study, which is first of its kind, on the selection and training of sub-syllable-level models for speaker-independent recognition of Cantonese continuous speech. The first known result with sub-syllable (INITIAL and FINAL) models for Cantonese speech recognition will be reported. The performance of recognizer using syllable modeling will also be provided for comparison purpose.

In the next section, we will briefly describe the Cantonese syllable structure. The methodology being used to define and train the acoustic models will be elaborated in the subsequent sections. Preliminary experimental results based on our newly developed Cantonese speech databases will be included.

2. BASIC PHONETIC STRUCTURE OF CANTONESE CHINESE

The total number of phonologically valid Cantonese syllables is 1,800. Each syllable carries a specific lexical tone. There are a total of six acoustically distinguishable tones in spoken Cantonese. Tone recognition is usually performed separately using non-spectral information, namely the temporal pitch movement. This is based on the assumption that vocal tract parameters are only slightly influenced by the glottal excitation. If tonal difference is disregarded, the total number of so-called base syllables is reduced to about 620.

Conventionally, Cantonese base syllable can be divided into an optional INITIAL and a FINAL. The INITIAL corresponds to a consonant onset C_1 while the FINAL is typically formed by a diphthong or a vowel nucleus V, followed by an optional consonant coda C_2. Thus Cantonese syllables have the general structure of [C_1] V [C_2]. (The brackets [ ] means optional.) This hierarchical syllable structure is shown as in Table 1.

Base syllable (620)		
INITIAL (19)	FINAL (53)	
[C$_1$]	V	[C$_2$]

Table 1. The Cantonese base syllable structure

INITIALs (C$_1$)	
Plosive:	
b	[p]
d	[t]
g	[k]
gw	[k^{w}]
p	[p^{h}]
t	[t^{h}]
k	[k^{h}]
kw	[k^{wh}]
Approximant:	
l	[l]
w	[w]
j	[j]
Nasal:	
m	[m]
n	[n]
ng	[ŋ]
Fricative:	
f	[f]
h	[h]
s	[s]
S	[ʃ]
Affricate:	
z	[ts]
Z	[tʃ]
c	[ts^{h}]
C	[tʃ^{h}]

FINALs (V[C$_2$])			
Vowel		Diphthong	
i	[i]	ui	[ui]
yu	[y]	ei	[ɛi]
u	[u]	eoi	[øy]
e	[ɛ]	oi	[ɔi]
oe	[œ]	ai	[ɐi]
o	[ɔ]	aai	[ai]
aa	[a]	iu	[iu]
		ou	[ɔu]
Vowel-nasal:		au	[ɐu]
im	[im]	aau	[au]
in	[in]		
Ing	[ɪŋ]	Vowel-stop:	
yun	[yn]	ip	[ip]
un	[un]	it	[it]
Ung	[ʊŋ]	Ik	[ɪk]
eng	[ɛŋ]	yut	[yt]
eon	[øn]	ut	[ut]
oeng	[œŋ]	Uk	[ʊk]
on	[ɔn]	ek	[ɛk]
ong	[ɔŋ]	eot	[øt]
am	[ɐm]	oek	[œk]
an	[ɐn]	ot	[ɔt]
ang	[ɐŋ]	ok	[ɔk]
aam	[am]	ap	[ɐp]
aan	[an]	at	[ɐt]
aang	[aŋ]	ak	[ɐk]
		aap	[ap]
Syllabic Nasal:		aat	[at]
M	[m]	aak	[ak]
NG	[ŋ]		

Table 2. A list of Cantonese INITIALs and FINALs using LSHK system. The corresponding IPA systems are also given for reference.

Phonologically speaking, there are 19 INITIALs and 53 FINALs in Cantonese. In this work, we split the fricative INITIAL /s/ and the affricate INITIALs /c/ and /z/ into finer sub-classes which correspond to their typical allophonic variations. (i.e., /s/ and /S/, /c/ and /C/, and /z/ and /Z/.) The 22 INITIALs and 53 FINAL s are listed in Table 2.

3. CANTONESE SPEECH DATABASES

One of the major difficulties for Cantonese speech recognition research is the lack of properly constructed speech databases for system training and performance evaluation. Recently, a series of large-scale Cantonese speech databases have been collected [5]. They provide the essential resources to facilitate effective acoustic modeling. In the present work, our acoustic models are trained and evaluated with one of these corpora, CUWORD, which is a large speech database of Cantonese polysyllablic words. Their performance for speech recognition will be analyzed and comparison is made between sub-syllable and syllable modeling.

The CUWORD database was designed to provide speech data for acoustic modeling at syllable or sub-syllable level. The basic corpus consists of 2,527 polysyllabic words, each of which is composed of 1 – 7 syllables. Thirteen male and fifteen female speakers were recorded. Each of them read through the basic corpus once in a moderately quiet room and a sampling frequency of 16 KHz was used. As a result, about 70,000 utterances were obtained and manually transcribed. The speech data from twenty two speakers (10 male and 12 female) were designated for the training of speech recognition systems while the remaining data (3 male and 3 female) were used for performance evaluation purpose.

It is noticed that 46 out of the 620 syllables occur less frequently in CUWORD. Indeed, these syllables are rarely used in daily life as well. Therefore, we remove utterances that contain these 46 syllables and essentially we are dealing with the problem of recognizing 573 base syllables only. Table 3 gives some statistics of the CUWORD database.

	Training Set	Testing Set
# of Recording	51,854	15,124
# of Speaker	22	6
# of Syllables	151,711	41,807
# of FINALs	151,711	41,807
# of INITIALs	150,582	41,003

Table 3a. The details of acoustic data in CUWORD

# of base syllables (out of 573)	Distribution
573	< 1,000
544	< 800
476	< 400
223	< 200
120	< 100
44	< 50
0	< 10

Table 3b. The distribution of base syllables in training set of CUWORD

# of INITIALs (out of 22)	Distribution
22	< 13,000
13	< 8,000
4	< 4,000
1	< 1,000

Table 3c. The distribution of INITIALs in training set of CUWORD

# of FINALs (out of 53)	Distribution
53	< 8,000
39	< 4,000
23	< 2,000
8	< 1,000
4	< 500

Table 3d. The distribution of FINALs and FINALs in training set of CUWORD

4. METHODOLOGY

Continuous-density Hidden Markov models (CHMM) are used for acoustic modeling. They have standard left-to-right model topology without state skipping. The CHMMs are trained by the modified segmental K-means training algorithm and Baum-Welch re-estimation method. Viterbi algorithm is used for decoding during recognition.

In all the experiments, each acoustic feature vector has 26 dimensions including 12-order Mel-Frequency Cepstral Coefficients (MFCC), the log energy and their first derivatives. The analysis window is 25*ms* with 10*ms* frame shift.

The following abbreviations are used for different speech units under consideration:

1. CI-I: context-independent INITIALs;
2. CI-F: context-independent FINALs;
3. CD-I: context-dependent INITIALs;
4. CD-F: context-dependent FINALs;
5. BS: base syllables.

4.1 Context Independent INITIALs and FINALs

As each Cantonese base syllable can be transcribed into INITIAL-FINAL format, we purposely create a set of generic INITIALs and FINALs, and then cascade them to form base syllables for recognition.

We first train 22 INITIAL HMMs and 53 FINAL HMMs without considering their phonetic context. The number of states used in each model roughly corresponds to the number of sounds (phonemes) within the phonetic unit being modeled. The details are shown in Table 4.

# of state	CI-I / CI-F
3	b, d, g, p, t, k, gw, kw, l, w, j, m, n, ng, f, h, s, S, z, Z, c, C
3	ap, at, ak, aa, aap, aat, aak, o, ot, ok, e, ek, eot, Ik, i, ip, it, Uk, u, ut, oe, oek, yu, yut, m, ng
5	ai, au, am, an, ang, aai, aau, aam, aan, aang, oi, ou, on, ong, ei, eng, eoi, eon, Ing, iu, im, in, Ung, ui, un, oeng, yun

Table 4. Total number of state for CI-I and CI-F

Generally, INITIAL models have 3 states and FINAL models have either 3 or 5 states. In this case, each phoneme (except the final stop consonants) will occupy 2 to 3 states. Since the final stop consonants of Cantonese are unreleased, the average duration of FINALs ending with stop consonants are much shorter than the other FINALs. Thus even these FINALs have two phonemes, only 3 states are assigned.

For non-speech modeling, a three-state HMM and an one-state HMM are used for silence and short pause respectively.

4.2 Context-Dependent Modeling

Considering the monosyllabic structure of Cantonese speech, we define two categories of phonetic context: intra-syllable versus inter-syllable.

Generally, the co-articulation effects within syllables are much significant than those across syllables. As a first step towards studying acoustic model unit selection for Cantonese speech recognition, we consider only intra-syllable co-articulations for context-dependent modeling.

Two further assumptions have been made for context-dependent modeling:

First, the dependency of INITIAL in intra-syllable context is primary related to the beginning phone of the FINAL only. Since the definition of FINAL has already taken the nucleus-coda transition into account, the intra-syllable context refers to the onset-nucleus transition only. For example, in the base syllable /leng/, the transition between consonant onset /l/ and vowel nucleus /e/ is the intra-syllable context being considered. In this way, we divide FINAL into 11 groups as shown in Table 5a. Consequently, a total of 156 right-context-dependent (RCD) INITIALs need to be modeled [6]. For example, the INITIAL /b/ has 8 RCD counterparts, denoted as /b_aa/, /b_e/, /b_i/, /b_I/, /b_o/, /b_u/ and /b_U/.

Secondly, we further assume that the intra-syllable context affects the INITIAL to a much greater extent than to the FINAL as the FINAL in a Cantonese syllable is usually much longer than the INITIAL. Besides, as these RCD INITIALs have accounted the onset-nucleus co-articulation, the left context dependency of FINAL in intra-syllable context is ignored.

A 3-state HMM is used to model each RCD INITIAL while the FINAL are still modeled context-independently as described in the previous section.

Obviously, the training data for each RCD-INITIALs must be less than that of CI-I. Their distribution is given in Table 5c. The average training samples for RCD INITIALs is around 900.

group	member	beginning phone
1	ai, au, am, an, ang, ap, at, ak	a
2	aa, aai, aau, aam, aan, aang, aap, aat, aak	aa
3	o, oi, ou, on ong, ot, ok	o
4	e, ei, eng, ek	e
5	eoi, eon, eot	eo
6	Ing, Ik	I
7	i, iu, im, in, ip, it	i
8	Ung, Uk	U
9	u, ui, un, ut	u
10	oe, oeng, oek	oe
11	yu, yun, yut	yu

Table 5a. The 11 groups of FINALs for context dependent modeling

	a	aa	o	e	eo	I	i	U	u	oe	yu
b	1	2	3	4		5	6	7	8		
d	9	10	11	12	13	14	15	16		17	18
g	19	20	21	22	23	24	25	26	27	28	29
p	30	31	32	33		34	35	36	37		
t	38	39	40	41	42	43	44	45		46	47
k	48	49	50	51	52	53	54	55	56	57	58
m	59	60	61	62		63	64	65	66		
n	67	68	69	70	71	72	73	74		75	76
ng	77	78	79					80			
gw	81	82	83								
kw	84	85	86								
w	87	88	89			90			91		
j	92	93		94	95	96	97	98		99	100
f	101	102	103	104				105	106		
l	107	108	109	110	111	112	113	114		115	116
h	117	118	119	120	121	122	123	124		125	126
z	127	128	129	130	131	132		133			
c	134	135	136	137	138	139		140			
s	141	142	143	144	145	146		147			
Z							148			149	150
C							151			152	153
S							154			155	156

Table 5b. A list of 156 RCD-INITIALs in which the vertical scale is the 22 basic INITIAL, and the horizontal scale is the beginning phones of the following FINAL in the 11 groups in Table 5a.

# of RCD-INITIALs (out of 156)	Distribution
156	< 4,000
143	< 2,000
95	< 1,000
45	< 500
1	< 50

Table 5c. The distribution of RCD-INITIALs in training set of CUWORD

4.3 Syllable Modeling

Syllable modeling is the most direct approach to the recognition of a monosyllablic language. Besides, it covers most of the context within syllables. In order to compare the recognition performance between base syllable modeling and sub-syllable modeling, we create HMMs for the 580 base syllables. In our training set, the number of occurrences for each base syllable is 312 on average.

Again, the number of states in BS models is assigned based on phonetic composition: three states (0+3), six states (3+3) or eight states (3+5) (Table 4).

5. EXPERIMENTAL RESULTS

The recognition performance with different modeling units is summarized in Table 6. The context-independent INITIAL and FINAL models give syllable accuracy of 60.70%. With the introduction of RCD-I models, the syllable recognition accuracy is significantly improved to 67.68%. As for the BS models, the best accuracy is 69.01%, which slightly outperforms the sub-syllable models.

	Base syllable accuracy (%)			
	1 mix	2 mix	3 mix	4 mix
CI-I + CI-F	50.73	57.76	58.01	60.70
CD-I + CI-F	59.77	66.45	66.53	67.68
BS	56.98	65.81	67.54	69.01

Table 6. The summary of recognition results

It is observed that many errors are caused by some easily confused Cantonese phonemes.

- Long vowel /aa/ and short vowel /a/: According to researches of Cantonese phonology [7], the /aa/ and /a/ are acoustically very close indeed, especially when they are followed by the same coda. This may explain why the confusion between /aa/ FINALs and /a/ FINALs accounts for 1/3 of recognition errors of FINALs.

- Nasal ending /n/, /ng/: More than 32% of /aang/ is classified as /aan/.

- Plosive INITIALs, /d/ and /g/.

In all our experiments, many insertion errors occur where deletion errors are few. Most of the insertions errors is due to 3 BS's, namely /Uk/, /M/ and /NG/. As these kinds of syllables contain a single short phoneme and do not have INITIALs, they have the smallest number of states and thus are highly probable to be erroneously inserted between syllables.

It is also observed that recognition errors are frequently observed at the coda-onset junction like /i/-/j/ in /zi/-/jin/. This is because the semi-vowel /j/ is phonetically very similar to the vowel /i/ so that it tends to be wrongly classified as /i/. (That is also the reason why /j/ and /i/ are most frequently mis-recognized over all INITIALs and FINALs respectively.) To reduce these typical errors and attained better recognition performance, context-dependent modeling for cross-syllable co-articulation is necessary.

The experimental results obtained so far indicate that sub-syllable models can produce recognition performance comparable to that of syllable models if context-dependent modeling is carried out properly. Since syllable models inherently cover all intra-syllable context, the experimental results give strong support to our earlier assumption that most of the intra-syllable context can be handled by the right-context-dependent INITIAL models. Furthermore, as sub-syllable models have smaller total number of states, they are fewer parameters and less sensitive to the amount of training data.

	Total number of models	Total number of states
CI-I + CI-F	75	279
CD-I + CI-F	209	681
BS	573	4,386

Table 7. The summary of total number of states for different modeling approaches.

6. ACKNOWLEDGEMENT

This work is supported by a RGC research grant (CUHK 368/96E) awarded by the Hong Kong Research Grant Council.

7. REFERENCES

[1] Tan LEE, Automatic Recognition of Isolated Cantonese Syllables Using Neural Networks, *Ph.D. thesis*, The Chinese University of Hong Kong, 1996.

[2] Tan LEE and P.C. CHING, "A neural network based speech recognition", *Proceedings of ICASSP-97*, vol.4, pp. 3269-3272.

[3] Y.P. Alfred NG, L.W CHAN and P.C. CHING, "Automatic recognition of continuous Cantonese speech with very large vocabulary", *Proceedings of EUROSPEECH-97*, vol.3, pp.1551-1554.

[4] H.Wang, et. al, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data", IEEE Transactions on Speech and Audio Processing, Vol.5 March 1997.

[5] W.K. LO, Tan LEE and P.C. CHING, "Development of Cantonese Spoken Language Corpora for Speech Applications", *ISCSLP-98*.

[6] K.F. CHOW, An HMM based Connected Speech Recognition System for Cantonese, *M.Phil. thesis*, The Chinese University of Hong Kong, 1998.

[7] Eric ZEE, "Formant Frequency and Vowel Categorization in Cantonese", *The Proceedings of the Conference on Phonetics of the Languages in China*, pp.137-140, 1998.