CHAPTER 2

Speech Segmentation

2.1 Acoustic Model Training

2.1.1 Speech Corpora

In this thesis, we use three different corpuses to train and test acoustic model, 1) The TIMIT corpus. The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT has resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO). Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and has been maintained, verified, and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). TIMIT contains a total of 6300

sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The text material in the TIMIT prompts consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. The dialect sentences were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences and each text was spoken by 7 different speakers. The phonetically-diverse sentences were selected from existing text sources - the Brown Corpus and the Playwrights Dialog - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts. Each speaker read 3 of these sentences, with each sentence being read only by a single speaker. 2) & 3) the second and third corpus are derived from the EAT (English Acoustic-model for Taiwanese) corpus. The EAT corpus recording project is originated by ITRI, C.C Lab. (Industrial Technology Research Institution，Computer and Communication Research Lab.) in 2004. Corpus speech recording was a joint effort among the National Taiwan University (NTU), National Tsing-Hua University (NTHU), National Chiao-Tung University (NCTU) and National Cheng-Kung University (NTHU). Every school is in charge of speech recording of 240 persons, and every person uttered 80 English sentences. Half of the persons are students from the department of foreign language (FL), and the others

are not. The collection of speech spoken by FL students is called EAT_ENG and the other are called EAT_NONENG.

2.1.2 Acoustic Model Design

To train a phone-level acoustic model, we need a word-to-phoneme dictionary to translate the word into phonemic units. There are 60 different phones in TIMIT dictionary design. However, for Chinese, some phones are not differentiated easily and are usually mis-pronounced in a similar way. Therefore, in this work, we use the dictionary under development at Carnegie Mellon University (CMU). The phone set conversion from the TIMIT to CMU is listed in Table 2.1. There are only 40 different phones of CMU dictionary. The complexity of acoustic models is reduced and training data for each model increase. Table 2.2 shows the phone models of CMU.

2.2 Recognition Network Generation

2.2.1 Word-Internal and Cross-Word Expansion

The pronunciation of English for native speakers and L2 learners differ in many ways, especially for the tempo of their pronunciation. The boundary between two words is often ignored by native speakers, so cross-word models are important for native speakers. For this reason, word-internal models may not cover all phonemes pronounced by natives. However, the cross-word phenomenon does not appear in the student's pronunciation, and they have tendency toward separating two words very obviously. So cross-word models may not so critical to

Deletion				
BCL,DCL,GCL,EPI,KCL,PAU,PCL,TCL (Closure interval of stops)				
Substitution				
TIMIT	Meanings	Word	Phn. list	Cmu
AX	devoiced Vowel	about	AX bcl aw tcl t	AH
AX-H	devoiced-schwa	suspect	S AX-H s pcl p eh kcl l tcl t	AH
AXR	devoiced Vowel	butter	Bcl b ah dx AXR	ER
DX	Flap /d/	muddy	M ah DX iy	D
HV	Voiced /h/	ahead	Ax HV eh dcl d	HH
IX		debit	Dcl d eh bcl b IX tcl t	IH
NX	nasal flap /n/	winner	W ih NX axr	N
Q	vowel-vowel boundary	button	b ah Q en	T
UX	Fronted /u/	toot	Tcl t UX tcl t	UW
Split				
ENG->IH NG EL->AH L EM->AH M EN->AH N				

Table 2.1 Modification from TIMIT's dictionary to CMU's dictionary

model	K.K.	model	K.K.	model	K.K.	model	K.K.	model	K.K.
AA	*Z*	**D**	*D*	**IH**	*I*	**OW**	*o*	**UH**	*U*
AE	*G*	**DH**	*F*	**IY**	*I*	**OY**	*W*	**UW**	*u*
AH	*O*	**EH**	*A*	**JH**	*P*	**P**	*p*	**V**	*v*
AO	*R*	**ER**	*S*	**K**	*k*	**R**	*r*	**W**	*w*
AW	*aU*	**EY**	*E*	**L**	*l*	**S**	*s*	**Y**	*y*
AY	*aI*	**F**	*F*	**M**	*m*	**SH**	*B*	**Z**	*z*
B	*b*	**G**	*G*	**N**	*n*	**T**	*t*	**ZH**	*N*
CH	*Q*	**HH**	*H*	**NG**	*E*	**TH**	*L*		

Table 2.2 Phone models in CMU's dictionary

L2 learners. The following example should clarify the process. It corresponds to the network which presents all sequences consisting of the words “bit” and “but” starting and ending with “sil”, so this grammar allows speakers to say “bit but bit but but bit …etc”. Firstly, assume that the dictionary contains simple monophone pronunciations, in this case, there will no expansion and network generator will directly generates the network shown in Fig 2.2.

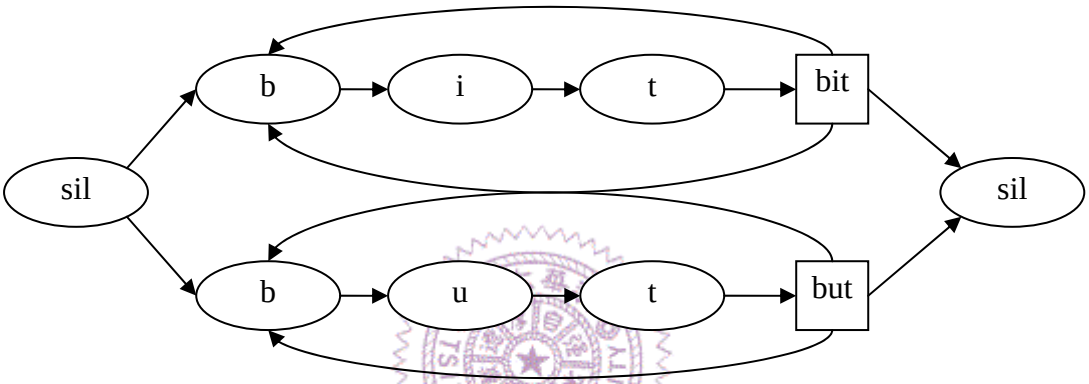

Figure 2.1 Monophone Expansion of Bit-But Network

Similarly, if the dictionary contained word-internal triphone pronunciations, then again, it will generate the network shown in Fig 2.3

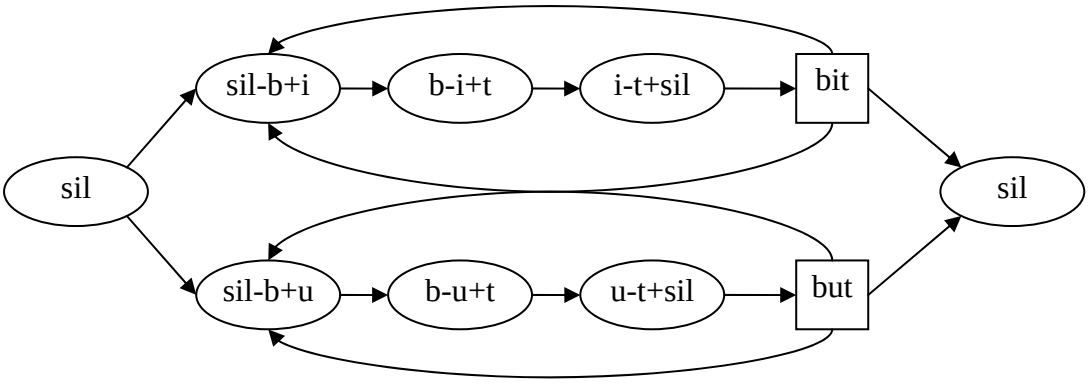

Figure 2.2 Word Internal Triphone Expansion of Bit-But Network

If, however, we use the cross-word network expansion, it will generate the network shown in Fig 2.4.

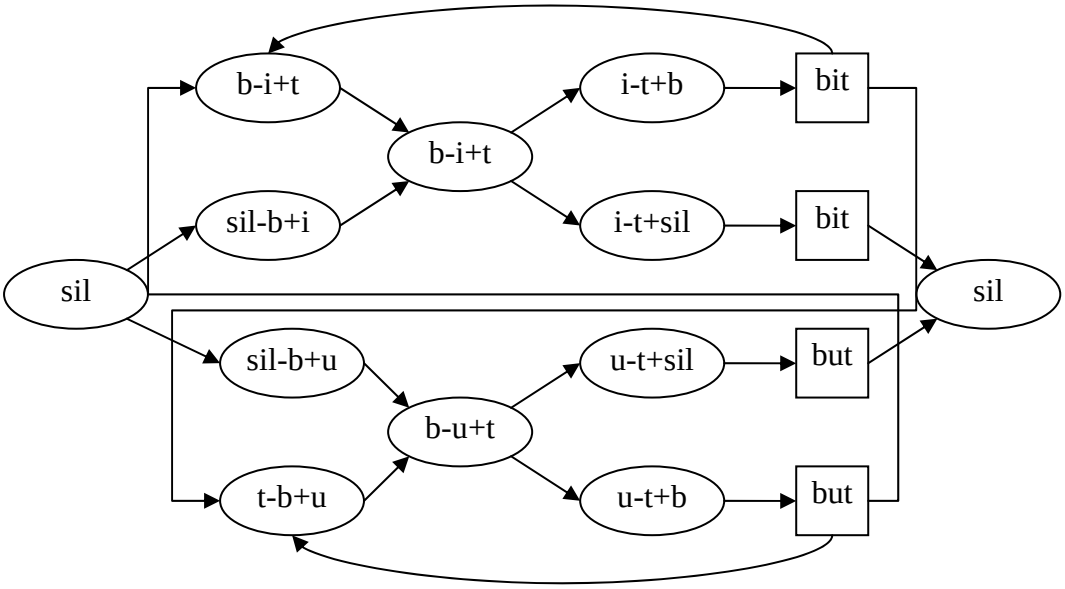

Figure 2.3 Cross-Word Triphone Expansion of Bit-But Network

When developing acoustic models, we use both of two kinds of network expansion methods to test the recognition rate and alignment rate.

2.2.2 Pronunciation Confusion Network (PCN)

To predict pronunciation errors, we need error patterns of Chinese students according to the linguistic literature. Table2.3 and Table2.4 lists all of the common pronunciation errors for Chinese in Taiwan, including 22 error patterns. There are 10 patterns concerning vowels substitution (shown in Table 2.3). In addition, there are 12 patterns concerning confusion between consonant pronunciations (shown in Table 2.4). To detect these errors, the patterns are applied to a PCN (pronunciation confusion network). For example, the PCN of the word “husband” is shown in Fig 2.5. The solid lines indicate the correct

sequence of the pronunciation and the dotted lines provide alternative paths to detect the possible pronunciation errors by L2 learner. The "sil" nodes represent the start and the end section of the silence. To align the usually longer and influent utterance of a L2 learner, a dynamic insertion approach is also used here.

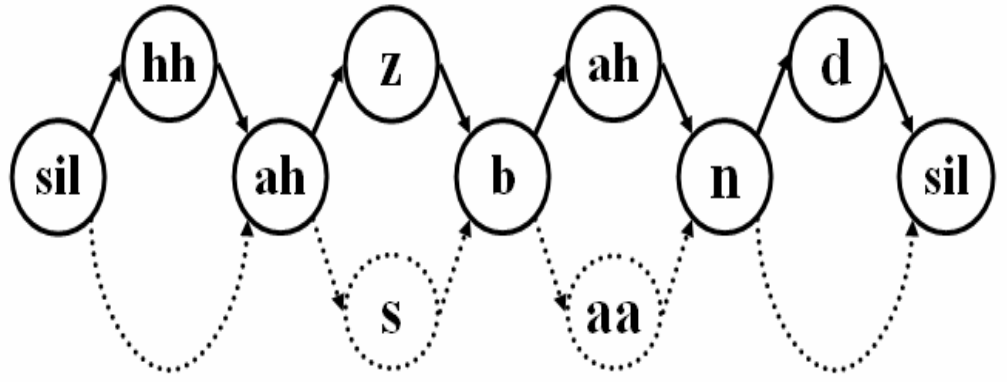

Figure 2.4 The PCN of the word "husband"

To determine if our method can detect the error pronunciation effectively, we use a dataset containing 135 utterances from two L2 learners, both of them are males, with various levels of proficiency in English. Two English experts manually transcribed the phones in each utterance, and labeled the type of the error pronunciation. In the process of transcribing, they added some new reference rules and modified the common pronunciation errors mentioned above. Table2.5 shows the modifications they proposed.

2.3 Speech Segmentation & PCN Approach

The goal of speech segmentation is to translate the incoming speech signal into a sequence of phonetic units based on the trained acoustic

models. When performing this process, a recognition network with the corresponding acoustic context-dependent model in the network is needed. In section 2.1.2, we have described the design of acoustic models, while in section 2.2, two kinds of network expansion methods and the network for predicting pronunciation error are also presented.

In the PCN approach, for an incoming utterance, we convert the sampled waveform of each window to 39-dimensions MFCC with various dimensions formant, and apply the viterbi decoding algorithm in the utterance alignment. The path with higher log probability is chosen when there are multiple paths in the recognition network generated by PCN. PCN setup the possible pronunciation error into the network and if error path is chosen, the system supposes the pronunciation is incorrect, and formant-level assessment will apply. This process is the first step of pronunciation error spotting and it try to detect the possible pronunciation error by recognition network.

1	/ *I* / and /i/are often confused.
Examples	is, him, if, seat, need, teacher
2	/e/ in pre-consonantal position is often pronounced [æ]
Examples	taken, made
3	/ *o* / is often realized as [*R*]
Examples	no, so
4	The diphthong / *aI* / is often simplified to the monophthong [*a*]
Examples	nice, I
5	/ *A* / is often replaced by [æ]
Examples	weather, next
6	/*O*/ is often realized as [*a*]
Examples	husband, months, funny
7	/ *U* / is often replaced by [u]
Examples	look, should, good
8	/ *U* / is often replaced by [o]
Examples	talk, long
9	/ *a* / is often replaced by [*R*] when orthographically written as o
Examples	not, John, Tom.
10	/z/ often realized as [s] when written as s in the orthography
Examples	is, days, shoes, those, husband

Table 2.3 Common pronunciation errors for vowel

11	/ks/, when represented by x in the orthography, is often simplified to [s]
Examples	excuse
12	/θ/ is often replaced by [s]
Examples	thank
13	/h/ is often realized as [x]
Examples	him, husband, how
14	Syllable-final /n/ is often deleted, leaving only a nasalized vowel before it
Examples	mine, the one in the stupid green sweater
15	Word-final / *H* m/ is often realized as [*H* n]
Examples	system, wisdom
16	Final voiced stops are usually completely devoiced, if they are pronounced at all
Examples	made, jog
17	Final stops are often deleted, even when the word ends with a grammatical /s/ ending
Examples	stupid, good, at, bought, like, pants, it's
18	Dark /l/ , together with the preceding vowel if there is one, is often realized as [o*U*]
Examples	also, cold
19	Postvocalic /r/ is often dropped
Examples	are, warm, person, learn, first, for
20	Epenthetic [*H*] sometimes added before the approximant in consonant clusters
Examples	black, England
21	Initial /ð/ often replaced by [l] or [d]
Examples	they, them, the
22	Nasals are often determined allphonetically by the backness of the preceding vowel; a nasal after the back vowels [o*U*], [*R*], [*O*], / *Z* / tends to be realized as a velar nasal [ŋ]; a nasal after the non-high front vowels/*A* / or /æ/ tends to be realized as alveolar [n]; though a nasal after high or mid-high front/i/ or / *I* / is usually [ŋ], and a nasal after the high and mid-high back vowels /u/ or / *U* / is usually [n]
Examples	want, months, in, been, run, (sometimes) him

Table 2.4 Common pronunciation errors for consonant

Modifications	
item	Description
10	[z] often realized as [s]
15	final [m] is realized as [n] e.g. from, home
17	final stops are often dropped → final consonants are often dropped also, the when [s] follows a stop, the stop is often dropped. e.g. cats
New Reference Rules	
item	Description
28	the second vowel of diphthongs is dropped e.g. around
29	final syllable is dropped e.g. mother
30	[r] is replaced by [l]
31	add [D] after the final stop
32	unaspirated stops become aspirated
33	[e] is replaced by [G]
34	[I] is replaced by [dI]
35	[M] is replaced by [L]
36	one of the consonants in consonant clusters is dropped e.g. dr, department
37	the fricative sounds of [z] is unclear
38	past tense is confused e.g. ed [t] or [d]

Table 2.5 common pronunciation errors proposed by two English experts