CHAPTER 4

Experimental Results

4.1 Introduction

The purpose of this thesis is to find the error pronunciation of L2 learners, and presents the feedback information according to the error type and the level of confident measure. So, high accuracy rate of error detection is critical for achieving good quality in CAPT system. Because of the large variation of pronunciation in English native speakers and L2 learners, we use the corpus from native speakers and L2 learners, and design some experiments of the speech recognition, phoneme segmentation and error pronunciation detection to find the best speech feature and parameters for our system. In this chapter, we will perform at least one experiment for the method we proposed above, detailed experiment results will be listed also.

4.2 The Corpora

4.2.1 Corpus for Acoustic Model Training

In the section 2.2.1, we have had an overview of three corpus used. Table 4.1 shows these three corpuses in more detail.

4.2.2 Test data for Error Pronunciation Detection

The corpuses mentioned above are used to train the acoustic models, and then perform speech recognition and forced alignment on speech. However, to exam the ability of detecting the error pronunciation, we need another test data which exist some error pronunciation and the corresponding error type marks. As mentioned in section 2.2.4, two English experts had transcribed the dataset from two L2 learners, and this dataset is used for the purpose of detecting error pronunciation. Table 4.4 shows the number for each error pronunciation type, and the ID can be referred to Table2.3, 2.4 and 2.5.

4.3 Experiment 1:

Recognition Accuracy for TIMIT

- Purpose

  Combine formant coefficients with MFCC to HMM for TIMIT corpus, and test the free phone recognition rate .

- Content

  A. Overview

Corpus	TIMIT	EAT_ENG	EAT_NONENG
Channel	Microphone		
Language	English		
Formant	16KHz,16bit		
Speakers			
Male	438	109	109
Female	192	131	131
Total	630	240	240
Total Data	6300 Sentences 5.38 Hours	7415 Sentences 8.24 Hours	7868 Sentences 8.74 Hours
Training Data	4620 Sentences 3.64 Hours	4940 Sentences 5.49 Hours	5248 Sentences 5.83 Hours
Testing Data	1680 Sentences 1.43 Hours	2475 Sentences 2.75 Hours	2620 Sentences 2.91 Hours
Remarks	Spoken by English Native	Spoken by Foreign Language Department Students	Spoken by Non-Foreign Language Department Students

Table 4.1 Detailed information for TIMIT, EAT

ID	#	ID	#	ID	#	ID	#
1	10	11	1	21	36	31	2
2	2	12	1	22	0	32	8
3	4	13	0	23	5	33	1
4	0	14	2	24	15	34	1
5	1	15	3	25	19	35	4
6	15	16	1	26	0	36	3
7	0	17	11	27	4	37	1
8	1	18	0	28	7	38	0
9	1	19	13	29	7		
10	7	20	8	30	5		

Table 4.2 The number for each error pronunciation type

In general, 39-dimension MFCC is chosen to be the feature of speech when training the acoustic model. However, according to the description in Chapter 3, we use formant frequency as additional feature set in different stream with MFCCs, and we test the performance of MFCC-Only experiment to be the comparsion.

B. Data

TIMIT (training and testing data is described in previous section)

C. Configurations

General Parameters	
Pronunciation Dictionary	CMU
Recognition Type	Free Phone
Mode Type	Biphone
Training Start Type	Flat Start
Viterbi Decoding Parameters	
Word insertion log probability	-24
Pruning Threshold	1000
Grammar scalar factor	0

D. Result

TIMIT					
ID	Formant Addition	Dim. added	Acc. for Vowels	Acc. for Consonants	Overall Acc.
1	N/A	0	47.08%	60.38%	55.20%
2	(F2-F1)/F1、F3-F2)/F2	2	46.07%	60.22%	54.70%
3	Delta_F1、Delta_F2	2	47.06%	60.78%	55.44%
4	Delta_F1~Delta_F5	5	47.62%	60.41%	55.44%
5	Delta_F1~Delta_F5 (Mean Normalization)	5	46.37%	60.14%	54.79%
6	F1~F5	5	45.83%	60.09%	54.53%
7	Delta_F1~Delta_F5、DDelta_F1~DDelta_F5	10	44.52%	58.72%	53.19%

Table 4.3 Recognition accuracy rate for TIMIT

E. Analysis & Discussion

From the result of Experiment 1, there is a slightly improvement for vowels in the row of ID 4. The feature set of formant coefficients used here is delta_F1 to delta_F5, it represents first order deviation coefficients of formant

frequency. In particularly, it means slope of the formant coefficients according to time. The smooth curve of formant coefficients in vowel should be an important feature; this situation is shown in dashed line region in Figure4.1.

The slope of different phoneme will be different, so this feature set can improve the performance in recognition accuracy rate. However, the slope of the formant coefficients according to time is unstable for consonants, so the recognition accuracy for consonants is unpredictable.

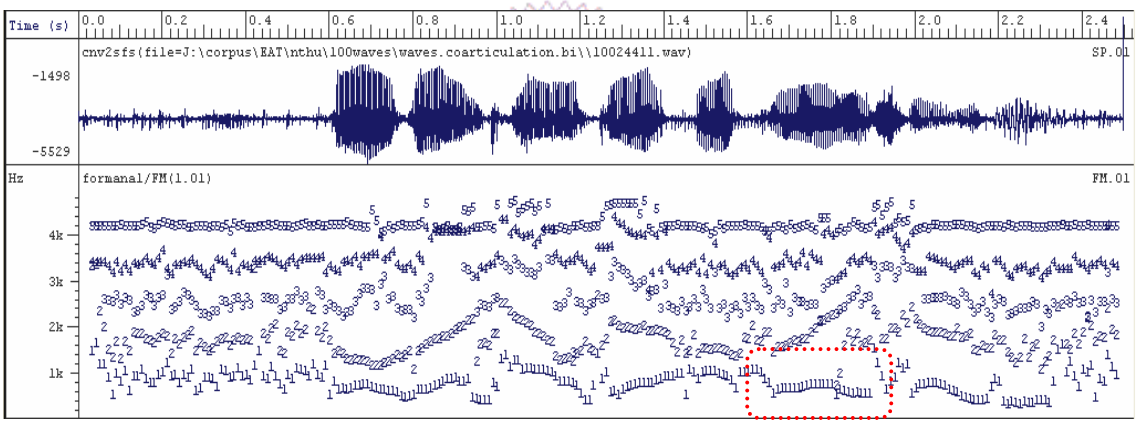

Figure 4.1 Formant curve for a utterance

4.4 Experiment 2:

Phonetic Segmentation accuracy for TIMIT

- Purpose

  Combine formant coefficients with MFCC to HMM for TIMIT corpus, and exam the performance of phonetic segmentation.

- Content

A. Overview

Different acoustic model leads to different result of phonetic segmentation. We use the 7 different feature sets in training in experiment 1, and compare the alignment accuracy with the manual transcription of TIMIT.

B. Data

TIMIT

C. Configuration

The same as experiment 1

D. Result

Formant Addition	<10 ms	<20 ms	<30 ms	<40 ms	<60 ms	<80 ms	<100 ms	>100 ms
N/A	33.2	63.2	82.0	89.6	95.1	97.0	98.0	1.99
(F2-F1)/F1、F3-F2)/F2	32.4	62.3	81.6	89.8	95.0	96.9	98.0	1.97
Delta_F1、Delta_F2	34.7	64.5	82.6	90.2	95.2	97.1	98.1	1.86
Delta_F1~Delta_F5	34.5	64.2	82.2	89.8	95.1	97.1	98.0	1.91
Delta_F1~Delta_F5 (Mean Normalization)	34.6	64.2	82.3	90.0	95.2	97.1	98.1	1.89
F1~F5	32.0	62.3	82.0	90.0	95.1	97.1	98.0	1.90
Delta_F1~Delta_F5、DDelta_F1~DDelta_F5	32.2	61.8	80.5	89.0	95.0	97.2	98.2	1.76

Table 4.4 Phonetic Segmentation Accuracy for TIMIT

E. Analysis & Discussion

The delta and double delta formant coefficients are more beneficial than MFCC only speech feature set, especially for the accuracy smaller than 30 ms. The same reason as experiment 1, we suppose high free phone recognition

accuracy will lead to high segmentation accuracy.

4.5 Experiment 3:

Recognition Accuracy for EAT

- Purpose

  Similar to the experiment 1, we want to exam how the formant coefficients influence the recognition accuracy rate for corpus spoken by L2 learners.

- Content

  A. Overview

  Because of the large variation among the pronunciation of L2 learners, the recognition rate for corpus spoken by L2 will be less than that spoken by native speakers. However, the property of formant frequency in vowel is still consistent, so this experiment studies the effect of formant coefficients to the recognition rate for EAT_NONENG and EAT_ENG.

  B. Data

  EAT_ENG & EAT_NONENG

  C. Configuration

  The same as experiment 1

  D. Result

EAT_ENG					
ID	Formant Addition	Dim. added	Acc. for Vowels	Acc. for Consonants	Overall Acc.
1	N/A	0	40.01%	42.20%	41.37%
2	(F2-F1)/F1、F3-F2)/F2	2	39.39%	42.83%	41.48%
3	Delta_F1、Delta_F2	2	39.51%	43.42%	41.88%
4	Delta_F1~Delta_F5	5	39.28%	43.78%	42.01%
5	Delta_F1~Delta_F5 (Mean Normalization)	5	39.07%	43.62%	41.83%
6	F1~F5	5	39.29%	43.01%	41.55%
7	Delta_F1~Delta_F5 、DDelta_F1 ~DDelta_F5	10	40.11%	43.83%	42.37%

Table 4.5 Recognition accuracy rate for EAT_ENG

EAT_NONENG					
ID	Formant Addition	Dim. added	Acc. for Vowels	Acc. for Consonants	Overall Acc.
1	N/A	0	35.23%	41.59%	39.09%
2	(F2-F1)/F1、F3-F2)/F2	2	35.79%	41.13%	39.03%
3	Delta_F1、Delta_F2	2	34.97%	41.46%	38/91%
4	Delta_F1~Delta_F5	5	35.91%	40.75%	38.85%
5	Delta_F1~Delta_F5 (Mean Normalization)	5	36.11%	40.84%	38.98%
6	F1~F5	5	35.01%	40.77%	38.50%
7	Delta_F1~Delta_F5 、DDelta_F1 ~DDelta_F5	10	35.03%	39.76%	37.90%

Table 4.6 Recognition accuracy rate for EAT_NONENG

E. Analysis & Discussion

Recognition accuracy for EAT_ENG is higher than that of EAT_NONEAT, because the variation of pronunciation of Foreign Language Department students is larger than that of students not belong to it. Different formant feature sets embed in 39-Ddimension MFCC improve accuracy rate according to different copra. From the result in Table 4.6, we can choose the most suitable formant feature set for students according to their pronunciation level.

4.6 Experiment 4: Recognition Accuracy Using Word-Internal & Cross-Word Network Expansion for TIMIT, EAT

- Purpose

Compare the recognition rate of acoustic models trained using word-internal & cross-word network expansion for the 3 different corpora.

- Content

A. Overview

As mentioned in section 2.2.3, we know that the different network expansion method in the process of training acoustic model will influence the overall recognition result. In this experiment, we want to determine the most suitable network

expansion for each corpus we used here, and compare the difference between them.

B. Data

TIMIT, EAT_ENG and EAT_NONENG

C. Configuration

The same as experiment 1

D. Result

E.

Corpus	Word-Internal	Cross-Word
TIMIT	53.77 %	55.16 %
EAT_ENG	38.65 %	37.92 %
EAT_NONENG	35.96 %	35.59 %

Table 4.7 Recognition rate of inter-word & cross-word network expansion

F. Analysis & Discussion

The recognition result listed in the Table4.5 is expected. It is clear that native speaker have much more fluent pronunciation than L2 learners, so the possibility of appearance of cross-word observation is also higher than L2 learners. It means the training data of cross-word models for corpus spoken by natives are larger in quantity, so the models are more consistent. However, the training data for cross-word models in corpus spoken by L2 learners will appear rarely. Worse still, it can lead to over-training since the models can become too closely matched to the training data and fail to

generalize well on unseen test data. When training the acoustic models, we can choose the most suitable network type according to the corpus.

4.7 Experiment 5: Determine the Number of Mixtures for GMM Used in Formant-Level Assessment

- Purpose

  Determine the number of mixtures in GMM of formant-level assessment, which will leads to the best score in average over each phoneme in TIMIT.

- Content

  A. Overview

  A GMM is needed in the process of formant-level assessment mentioned in section 4.4. We know that the number of mixtures is a critical parameter for Gaussian Mixture Model, so before performing assessment, we have to determine the number of mixtures for the GMM. We use the average RCM over all phonemes in TIMIT test data as criteria to decide the number of mixtures. We start at 2 mixtures, and then 4,8,16, etc, until the average RCM begin to drop.

  B. Data

  There are 117205 phonemes for training and 42620 phonemes for testing. All of them come from test data in TIMIT.

  C. Formula

$$\text{Average RCM} = \frac{\sum_{\text{all phoneme segmentin copurs}} RCM \text{ of Phoneme Segment } i}{\text{total \# of phoneme segments in coprus}} \quad \text{Eq. 4-1}$$

D. Result

# of mixtures	Average RCM
2	0.7336
4	0.7794
8	0.7912
16	0.8047
32	0.7958

Table 4.8 Average RCM with different number of mixtures

E. Analysis & Discussion

From the result listed in Table 4.8, when number of mixtures equal to 16, we get the highest average RCM. If we continue to increase the number, RCM is starting to drop, worse still, some Gaussian models is lost because of the training data for these context-dependent models is too less.

4.8 Experiment 5:

Determine the Threshold for RCM

- Purpose

Determine the threshold for RCM according to the equal error rate

- Contents

A. Overview

When an error pronunciation phoneme is detected in PCN approach, we measure the confidence of this phonetic segment.

If the RCM of this phoneme is less than the threshold, the feedback generation is carried out. Therefore, the decision of threshold is very important. High threshold leads to false rejection the correct sound, while low threshold fail to false accept the error one. In this experiment, from 0.1 to 1.0, and 10 points with 0.1 increasement are tested in the range.

B. Data

The dataset for detecting the error pronunciation mentioned in section 4.2.2.

C. Result

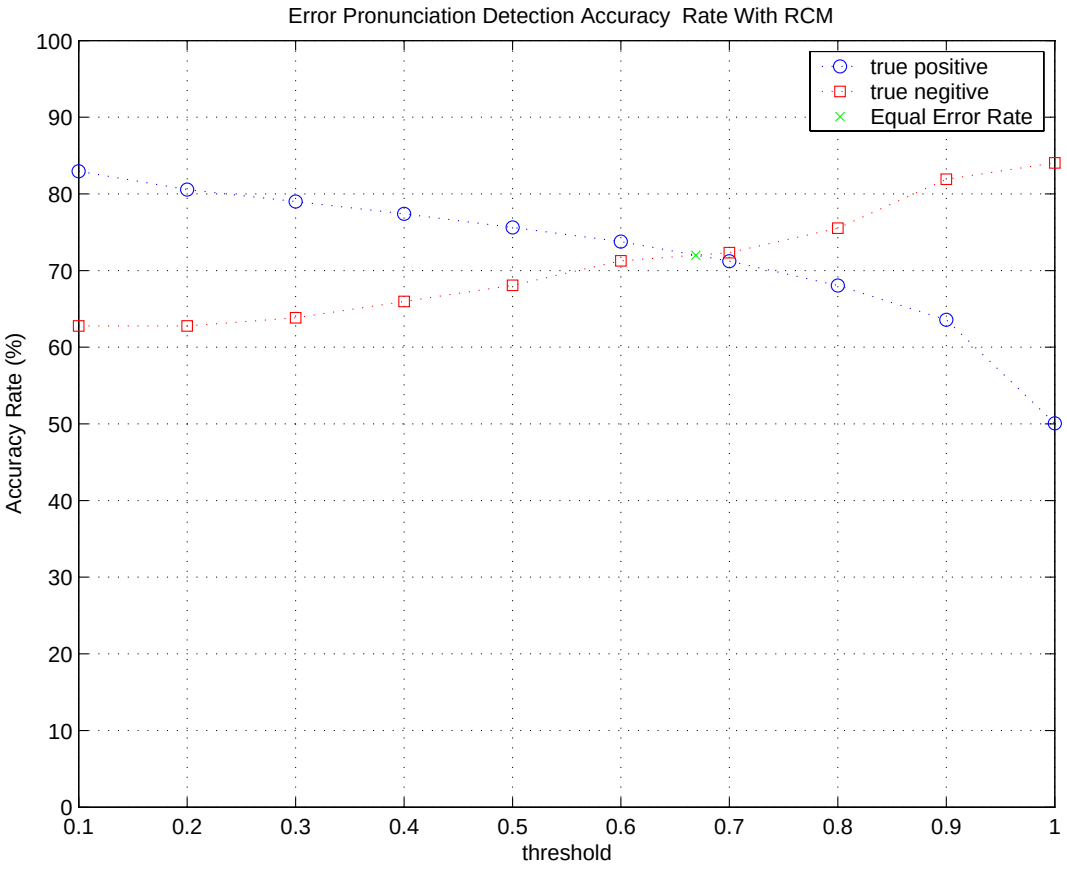

Figure 4.2 Error pronunciation detection accuracy with RCM using different threshold