第4章 實驗結果與討論

4.1 實驗簡介

本論文研究目的在於，自動偵測使用者個人的發音錯誤混淆型態，並動態新增至錯誤混淆規則的定義中，將一般系統使用的 PCN 加上 BPCN 及使用者個人化調適的 PPCN 而後產生 BPPCN，讓系統能更有效地抓出使用者的錯誤混淆發音。我們進行了以下實驗，首先找出中文音素和英文音素在發音上的對應關係，讓系統產生 BPCN，然後偵測產生使用者個人特殊的錯誤混淆規則以建立 PPCN，最後合併兩者組成 BPPCN，並比較 PCN 與 BPPCN 兩種不同辨識網路在偵測錯誤混淆發音上的準確率。

4.2 語料和聲學模型

4.2.1 聲學模型訓練語料

我們使用 EAT（English Across Taiwan）語料來訓練英文聲學模型[10]。EAT 語料是由台灣大學、清華大學、交通大學、成功大學和台灣師範大學五所學校的師生聯合錄製，包含市話錄音、手機錄音和麥克風錄音語料。我們實驗使用的麥克風錄音語料分為 ENG（EAT_ENG）和 NONENG（EAT_NONENG）兩部分。EAT_ENG 的錄製人員為英語系的師生，訓練所使用的人數為 553 位，總句數為 26752 句，總長度約為 15 時 43 分 47 秒。EAT_NONENG 的錄製人員則為非英語系的師生，其英文發音中含有中文口音的情形較明顯，訓練所使用的人數為 576 位，總句數為 26281 句，總長度約為 16 時 23 分 02 秒。

中文聲學模型是以 TCC-300 的語料進行訓練[11]，TCC-300 語料是由台灣大學、交通大學和成功大學的師生錄製，總人數為 300 人，總句數為 8913 句。

4.2.2 測試語料

測試聲學模型用語料是由 EAT 兩部份中抽取出來。EAT_ENG 的部份，人數為 552 位，總句數為 6856 句，總長度約為 3 時 55 分 56 秒；EAT_NONENG 的部份，人數為 571 位，總句數為 6790 句，總長度約為 4 時 05 分 45 秒。

4.2.3 聲學模型的建立

我們根據 CMU（Carnegie Mellon University）字典定義的拼音簡化成 39 種英文音素[12]，建立 40 個聲學模型，包含 39 個 Mono-Phones 和一個 Silence Model。表 4-1 為我們使用的 40 個英文聲學模型。

表 4-1：CMU 英文音素符號和 K.K.音標對應

Model	K.K	Model	K.K	Model	K.K.	Model	K.K.
AA	ɑ	**EH**	ɛ	**L**	l	**SIL**	sil
AE	æ	**ER**	ɚ	**M**	m	**T**	t
AH	ʌ	**EY**	e	**N**	n	**TH**	θ
AO	ɔ	**F**	f	**NG**	ŋ	**UH**	ʊ
AW	aʊ	**G**	g	**OW**	o	**UW**	ʉ
AY	aɪ	**HH**	h	**OY**	ɔɪ	**V**	v
B	b	**IH**	ɪ	**P**	p	**W**	w
CH	tʃ	**IY**	i	**R**	r	**Y**	j
D	d	**JH**	dʒ	**S**	s	**Z**	z
DH	ð	**K**	k	**SH**	ʃ	**ZH**	ʒ

中文的聲學模型部份，我們則是依照中文的 37 個注音符號，來建立 38 個中文聲學模型，包含 37 個 Mono-Phones 和一個 Silence Model。表 4-2 為我們使用的 38 個中文聲學模型。

表 4-2：中文音素符號和注音的對應

Model	Phoneme	Model	Phoneme	Model	Phoneme	Model	Phoneme
B	ㄅ	H	ㄏ	S	ㄙ	AU	ㄠ
P	ㄆ	J	ㄐ	I	ㄧ	OU	ㄡ
M	ㄇ	Q	ㄑ	U	ㄨ	ANN	ㄢ
F	ㄈ	X	ㄒ	YU	ㄩ	NN	ㄣ
D	ㄉ	ZH	ㄓ	A	ㄚ	ANG	ㄤ
T	ㄊ	CH	ㄔ	O	ㄛ	NG	ㄥ
N	ㄋ	SH	ㄕ	ER	ㄜ	ERR	ㄦ
L	ㄌ	RH	ㄖ	E	ㄝ	sil	sil
G	ㄍ	Z	ㄗ	AI	ㄞ		
K	ㄎ	C	ㄘ	EI	ㄟ		

聲學模型的訓練部分，我們使用三十九維的梅爾倒頻譜參數（Mel-Frequency Cepstral Coefficients, MFCC）做為訓練資料的特徵參數，包含十二階的頻譜值加上能量值，並取其一階微分和二階微分。模型格式為 3 個 States，每個 State 由 16 個 Mixtures 組成、Mono-Phone 的 HMM 模型，我們使用 HTK 來訓練聲學模型。

4.3 實驗一：尋找英文音素和中文音素的對應關係

利用 3.2 中介紹的流程，我們進行實驗找出英文音素和中文音素之間的對應關係。下表 4-3 列出實驗的對應結果。Model 和 K.K.欄位所標示的為英文聲學模型名稱和其對應的英文音素，1st、2nd 和 3rd Map 欄位所表示為累計統計各種中文音素對應到次數的前三名，以及該中文音素在該英文音素全部對應中的所佔比例。

表 4-3：尋找英文音素與中文音素對應關係的結果

Model	K.K.	1st Map / %	2nd Map / %	3rd Map / %
aa	/ ɑ /	ㄚ / 62%	ㄦ / 12%	ㄛ / 8%
ae	/ æ /	ㄝ / 27%	ㄦ / 21%	ㄜ / 16%
ah	/ ʌ /	ㄜ / 43%	ㄦ / 14%	ㄚ / 13%
ao	/ ɔ /	ㄛ / 44%	ㄡ / 38%	ㄦ / 6%
aw	/ aʊ /	ㄠ / 47%	ㄡ / 34%	ㄦ / 7%
ay	/ aɪ /	ㄞ / 71%	ㄚ / 11%	ㄤ / 4%
b	/ b /	ㄈ / 26%	ㄅ / 19%	ㄆ / 15%
ch	/ tʃ /	ㄑ / 35%	ㄔ / 27%	ㄒ / 9%
d	/ d /	ㄆ / 13%	ㄉ / 11%	sil / 9%
dh	/ ð /	ㄉ / 18%	ㄖ / 15%	ㄦ / 14%
eh	/ ɛ /	ㄝ / 40%	ㄦ / 17%	ㄜ / 17%
er	/ ɚ /	ㄦ / 41%	ㄜ / 23%	ㄥ / 7%
ey	/ e /	ㄟ / 58%	ㄝ / 21%	ㄦ / 9%
f	/ f /	ㄈ / 50%	sil/ 17%	ㄏ / 7%
g	/ g /	ㄍ / 23%	ㄆ / 16%	sil / 10%
hh	/ h /	ㄏ / 52%	ㄆ / 18%	ㄊ / 8%
ih	/ ɪ /	ㄧ / 45%	ㄩ / 23%	ㄟ / 8%
iy	/ i /	ㄧ / 73%	ㄩ / 9%	sil / 4%
jh	/ dʒ /	ㄓ / 26%	ㄑ / 17%	ㄒ / 11%
k	/ k /	ㄎ / 24%	sil / 23%	ㄆ / 20%
l	/ l /	ㄉ / 25%	ㄨ / 10%	ㄖ / 8%
m	/ m /	ㄇ / 38%	ㄣ / 18%	ㄥ / 12%

n	/ ɳ /	ㄣ / 51%	ㄥ / 21%	ㄋ / 12%
ng	/ ŋ /	ㄣ / 39%	ㄥ / 32%	sil / 9%
ow	/ o /	ㄡ / 59%	ㄛ / 24%	ㄨ / 11%
oy	/ ɔɪ /	ㄡ / 55%	ㄛ / 15%	ㄞ / 11%
p	/ p /	ㄆ / 56%	sil / 12%	ㄊ / 9%
r	/ r /	ㄦ / 19%	ㄖ / 17%	ㄨ / 15%
s	/ s /	ㄙ / 33%	ㄒ / 20%	ㄘ / 19%
sh	/ ʃ /	ㄕ / 50%	ㄒ / 34%	ㄑ / 7%
t	/ t /	ㄆ / 23%	ㄊ / 16%	ㄔ / 12%
th	/ θ /	ㄙ / 21%	ㄷ / 15%	sil / 12%
uh	/ ʊ /	ㄨ / 46%	ㄜ / 19%	ㄡ / 12%
uw	/ ʉ /	ㄨ / 31%	ㄙ / 12%	ㄜ / 10%
v	/ v /	sil / 20%	ㄏ / 13%	ㄌ / 12%
w	/ w /	ㄨ / 59%	ㄡ / 10%	ㄏ / 9%
y	/ j /	ㄧ / 38%	ㄩ / 34%	ㄝ / 8%
z	/ z /	sil / 19%	ㄒ / 17%	ㄙ / 15%
zh	/ ʒ /	ㄕ / 25%	ㄓ / 16%	ㄒ / 14%

從對應結果中可以發現，某些比例值高的對應組，英文音標唸起來確實和中文音標類似，例如：/ ɑ / 對應"ㄚ"（62%）、 / aɪ / 對應"ㄞ"（71%）；某些前二名或前三名的比例值相近的，英文音標唸起來似乎介於兩種中文音標的模糊地帶，與兩種音皆近似，例如：/ ɔ / 對應"ㄛ"（44%）和"ㄡ"（38%）、/ ɚ / 對應"ㄦ"（41%）和"ㄜ"（23%）；某些對應組前三名的比例值皆不高，感覺彼此間似乎並無相似關係，例如：/ dʒ / 對應"ㄓ"（26%）或"ㄑ"（17%）、/ v / 對應"sil"（20%）或"ㄏ"（13%）。由此發現，所佔比例的值似乎可以代表英文和中文對應組之間相似關係的程度，比例值越高代表兩種音標越相似，比例值越低則可能兩種音標毫無關聯。於是，我們利用下一個實驗，用來找出最適合判定對應組成立的所佔比例值。

4.4 實驗二：尋找對應關係的最佳比率值

要找出最適合判定對應組成立的比例值門檻，我們以對應組中的最低比率值（/ d / 對應"ㄉ", 13%）和最高比率值（/ I / 對應"一", 73%）為參考，取 10 和 60 為最低和最高門檻值，以 5 為區間，定出 11 種門檻值並分別進行實驗。在每次實驗中，以不同的門檻值為限，而唯有比例值超過該次實驗門檻值的對應組才算成立。實驗內容是以英文聲學模型對測試語料進行 Continuous Phone Recognition，在計算每種英文音素的模型機率值的同時，也計算該英文音素所對應到的中文音素的機率值，然後以機率值高的對應組為辨識結果。圖 4-1 為這項實驗的結果：

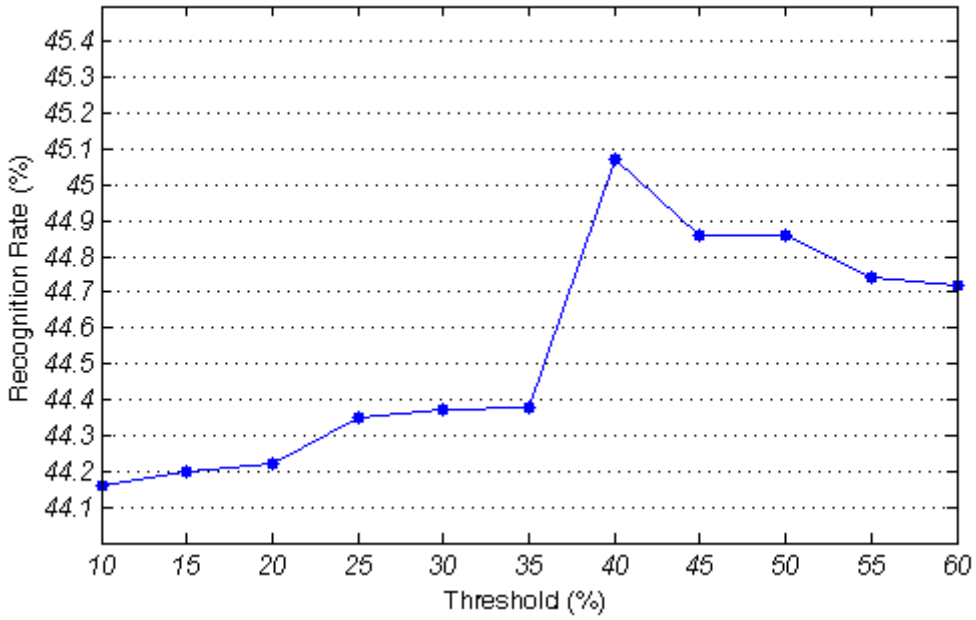

圖 4-1：尋找對應組最佳比率門檻值實驗結果

從實驗結果中得出，以 40%作為對應組比例值的門檻值來判定對應組，並用來輔助中文口語的英文辨識，能有最佳的辨識結果。

同時，從圖中可看出將門檻值自 35%調至 40%時能有較明顯的輔助辨識效果，

我們對應著前次實驗所得的對應組結果進行分析，發現對應組比例值為在 35%至 39%之間的對應組有 /ʧ/、/m/、/ŋ/和/j/，除了/m/之外的三組對應組其前兩名的對應比例值相差在 4%至 8%。再觀察比例值為 40%以上的對應組，每組的前兩名對應比例值則相差 13%至 64%之間。此外，/n/和/ŋ/皆對應到/ㄣ/，/j/、/i/和/ɪ/皆對應到/ㄧ/，而在將/m/和/j/列為不參考的對應組後，皆提升了/n/、/i/和/ɪ/的辨識率 3%至 5%。因此，我們認為當對應組中前兩名的比例值相差較低時，這些中英文音素的發音對應關係於是不明顯而不該成立。並且，若成立此類組的對應關係反而會導致降低其他對應組的輔助辨識效果。

最後，我們實驗比較利用對應關係輔助辨識的差異。一組以僅用英文聲學模型進行辨識，另一組則引用中英文音素對應關係，利用對應關係成立之中文模型輔助辨識。從實驗結果表 4-4 可以發現參考對應組能有效提升辨識率。之後在系統中，我們就以比例值為 40%以上的對應組用於建立 BPCN。

表 4-4：參考母語資訊辨識結果比較

	English Acoustic Models without Bilingual Mapping	English Acoustic Models with Bilingual Mapping	Improved Rate
Recognition Rate	34.48 %	45.07 %	10.59 %

4.5 實驗三：偵測和建立個人化錯誤混淆規則

4.5.1 實驗語者語料

我們從 EAT 語料中選出十位語者語料來進行下面的實驗，這十位語者語料皆未包含在聲學模型訓練與測試實驗中。其中五位語者選自 EAT_ENG 的部份，有一位男性和四位女性；另五位語者來自 EAT_NONENG 的部份，有三位男性和二位女性。此十位語者語料的總句數為 630 句，總長度約為 21 分 20 秒。

4.5.2 實驗內容

利用 3.3 中介紹的流程，對 10 位語者語料進行偵測個人化錯誤型態並建立錯誤混淆規則的實驗。在表 4-5 中列出實驗的部份結果，為每位語者部分常出現的個人錯誤混淆發音，其中有標示星號的規則是一般典型常見發音錯誤中所未見的，屬於使用者個人特殊的錯誤發音方式。沒有標示星號的規則，則是和一般典型常見發音錯誤中相同的錯誤規則。

表 4-5：語者個人常見錯誤混淆規則部份結果

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5
母音錯誤	/ æ /→/ ɛ /	/ ɚ /→/ ə / / ə /→/ ɑ /	*/ ə /→/ ɪ / / e /→/ ɑ /	*/ o /→/ ʊ / / e /→/ ɛ /	/ ɪ /→/ ɛ /
子音錯誤	/ m /→/ ŋ / */tʃ/→/dʒ/	/ s /→/ θ / / n /→/ ŋ /	/ n /→/ ŋ /	/ p /→/ b /	*/dʒ/→/tʃ/ /nd/→/ ŋ /
	Speaker 6	Speaker 7	Speaker 8	Speaker 9	Speaker 10
母音錯誤	*/ ʌ /→/ ɪ / */ ɚ /→/ ɑ / / ɔ /→/ ɑ /	/ o /→/ ɔ / / e /→/ æ / / æ /→/ ɛ /	/ ʌ /→/ ɑ / / o /→/ ɔ / */ ɛ /→/ ʊ /	/ e /→/ ɛ / / æ /→/ ɛ /	/ o /→/ ɔ / / æ /→/ ɛ /
子音錯誤	*/ ʒ /→/ ʃ / / g /→/gə/	*/ θ /→/ f / / b /→/ d /	/ɑr/→/ ɑ / /ks/→/gz/ / b /→/ p /	/ ɔr /→/ ɔ / / n /→/ ŋ / / t /→/ d /	/ s /→/ θ / / m /→/ ŋ / / t /→/ d /

於是我們從辨識結果中找出使用者語音發生錯誤發音的部份，確認規則內容是否正確，並找出這些特殊錯誤方式的發生範例。舉例，Speaker 1 的子音取代性錯誤規則：/tʃ/取代為/dʒ/的錯誤發生例子為單字 launch / lɔntʃ /，人工標示使用者的真實發音內容為/ lɔndʒ /；Speaker 3 的母音取代性錯誤規則：/ə/取代為/ɪ/的錯誤發生例子為單字 beautiful / bjutəfəl /，人工標示使用者的真實發音內容為/ bjutɪfəl /；Speaker 7 的子音取代性錯誤規則：/θ/取代為/f/的錯誤發生例子為單字 arithmetic / ərɪθmətɪk /，人工標示使用者的真實發音內容為/ ərɪfmətɪk /。

4.6 實驗四：比較 PCN、BPCN 與 PPCN 偵測錯誤發音的正確率

產生語者個人錯誤混淆規則後，系統即能建立語者個人化調適的 PPCN 作為系統辨識網路。此處進行實驗比較四種發音混淆網路，分別為一般系統所使用的固定規則 PCN 和本論文提出的 BPCN 和 PPCN，以及 BPCN 合併 PPCN，這四種不同辨識網路所能偵測語者錯誤混淆發音的正確率。

我們分別先以僅包含一般常見發音錯誤規則的 PCN 對實驗三所使用的十位語者語料進行辨識，再以包含中文和英文音素對應關係而建立的 BPCN 進行辨識。接著，利用上節實驗的方法產生十位語者的個人錯誤規則並建立每位語者的 PPCN 進行辨識。最後再以 BPCN 與 PPCN 合併成的發音混淆網路進行辨識。

實驗四中使用的語料如同實驗三，是自 EAT 語料庫中挑出的十位語者語料。不同在於，在此實驗之前，我們先以人工標記的方式標示出每句語音真實發音的音素，以此人工標記的內容作為系統應該產生的正確辨識結果。當系統辨識結果和人工標示內容完全相符合時，即表示系統能將正確發音和錯誤混淆發音內容都正確辨識。在此實驗中同樣藉由 LCS 演算法來找出辨識結果和人工標記的正確音素，這兩個字串的相同和不同處，用來計算辨識錯誤發音的正確率。計算辨識正確率的公式如下：

$$\text{Accuracy Rate} = \frac{\text{H - I}}{\text{T}}$$

公式中的 H 為兩字串的最大子字串的音素個數，即正確辨識的音素個數，I 為產生插入錯誤的音素個數，T 為人工標記正確音素字串中的音素總個數。

圖 4-2 為四種發音混淆網路的實驗結果。從結果圖中可以看出，使用 PPCN 作為系統辨識網路，其對於語者的錯誤混淆發音的辨識效果為最佳，這證實本篇論文提出的個人化發音混淆網路的實用效果。但令人意外的，包含母語發音對應的 BPCN 效果不如預期，十位語者的辨識率結果皆比使用固定 PCN 的辨識率低，平均下降約 3.81%。而包含母語發音對應和使用者個人發音規則的發音混淆網路

（BPCN+PPCN），其辨識率平均雖稍比使用固定 PCN 高，但仍比 PPCN 平均降低約 3.48%。

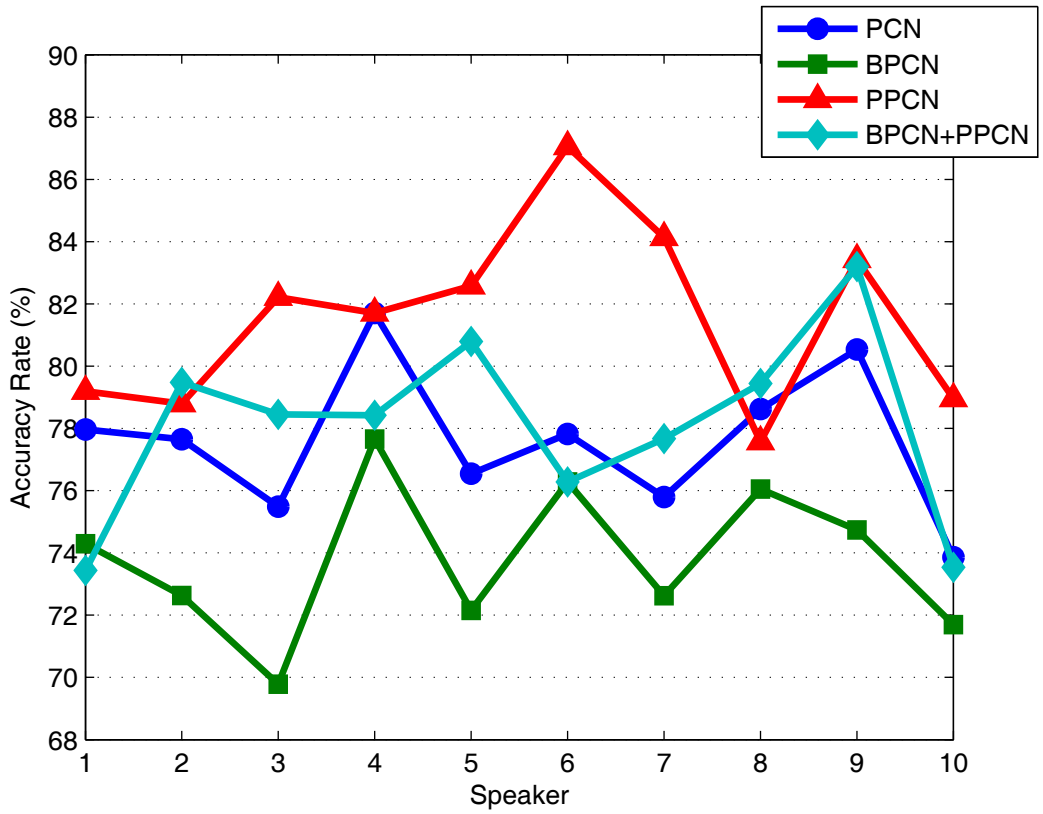

圖 4-2：四種發音混淆網路偵測錯誤混淆發音的辨識正確率

我們推測使用 BPCN 卻降低辨識率的原因，可能有以下幾點：

一、 兩套模型機率值未經過正規化：英文模型和中文模型是兩套不相同且分開訓練的模型，即便發音方式相似，但其整體平均的機率密度分布必不相同。實驗過程中並未先將兩套語言模行經過正規化，直接合併在一起進行辨識，可能導致機率值偏移向某一方平均值較高的語言模型，使得模型之間互補性喪失。

二、 尋找對應的方式太簡易：我們僅只以計算模型機率值決定英文音素和母語音素的對應，這種武斷認定的方式過於粗簡。英文和中文的發音方式上有明顯的不同，例如讀音時間的長短、讀音時的語調高低或聲調起伏等，這

些語音特性皆沒有考慮到。

由於決定對應的方式粗略，又未將聲學模型進行正規調適，於是將不適當的對應模型加入辨識網路，不但無法發揮協助系統辨認混淆發音的效果，反而徒增辨識網路的複雜度，而增加辨識錯誤的機會。實驗中英文音素的聲學模型數量共有 39 個，而 BPCN 中包含的母語發音對應有 21 組，於是有 21 個英文音素的路徑皆變為兩種，其建構的辨識網路的路徑數量為未考慮母語發音對應的 PCN 的 1.5 倍以上，於是可能因此導致 BPCN 和組合的 BPCN+PPCN 的辨識能力下降。

表 4-6：三種發音混淆網路相對 PCN 所改善的辨識正確率平均值

Recognition Rate (%)	PCN	BPCN	PPCN	BPCN+PPCN
EAT_ENG	77.87	73.30	80.90	78.12
EAT_NONENG	77.33	74.28	82.23	78.03
Average	77.6	73.39	81.57	78.08
EAT_ENG - EAT_NONENG	0.54	-0.98	-1.33	0.09

Recognition Rate (%)	BPCN - PCN	PPCN - PCN	(BPCN+PPCN)-PCN	(BPCN+PPCN)-PPCN
EAT_ENG	-4.57	3.03	0.25	-2.78
EAT_NONENG	-3.05	4.9	0.7	-4.2
Average	-3.81	3.97	0.48	-3.49
EAT_ENG - EAT_NONENG	-1.52	-1.87	-0.45	1.42

表 4-6 中分別列出實驗中 EAT 的英語系語者與非英語系語者使用四種發音混淆網路實驗所得辨識正確率的平均值，和不同方法之間的辨識率差值。從表 4-6 中可明顯看出使用 PPCN 進行辨識對於兩方語者皆為最佳辨識效果，平均辨識率為 80.9%和 82.23%，其中非英語系語者的平均辨識率較英語系語者高 1.33%。非英

語系語者使用 PPCN 的辨識率提升幅度達 4.9%，也較英語系語者高。這項結果符合我們的一般印象：非英語系語者的發音通常較不標準，較容易有個人特殊的口音和發音方式。而利用我們設計的流程所建立個人化錯誤規則，能將那些特殊錯誤情形偵測出來，輔助系統偵測錯誤混淆發音提升辨識率。英語系的語者則可能因發音相較正確，個人特殊的錯誤情形較少，因此辨識率提升程度相較為低。

另外可以觀察到的是使用 BPCN 後的差異。雖然兩方語者使用 BPCN 的結果皆比使用 PCN 的所得辨識率低，但非英語系語者的辨識率平均下降為 3.05%，而英語系語者的辨識率平均下降 4.57%，非英語系語者的辨識率下降幅度較小。可以確定的是非英語系語者的發音較不標準，有較明顯的母語口音情形。雖然使用 BPCN 後平均辨識率降低，但非英語系的降低幅度較小，這或許表示利用母語發音對應輔助偵測錯誤發音是有些許程度的效果。因此使用 BPCN 進行辨識，讓非英語系語者的平均辨識率比英語系語者的平均辨識率高。