應用錯誤型態分析於英語發音輔助學習

湯士民 莊則敬 吳宗憲
國立成功大學資訊工程學系
{ming, bala, chwu}@csie.ncku.edu.tw

摘要

語言教學方法主要是由以互動理論 (interactionist theories) 為基礎的溝通式教學法 (communicative language teaching) 所主導。因此，如果要針對學生個別的問題進行糾正，需要甚多的時間，很難採用雙向互動的教學方法。要解決這樣的問題，電腦輔助語言學習系統 (Computer Assisted Language Learning System, CALL) 是個可行的方案。利用語音辨識 (Automatic Speech Recognition, ASR) 技術的電腦輔助發音訓練系統 (Computer Assisted Pronunciation Training, CAPT) 不但可以提供一個沒有壓力的環境，讓學生反覆的練習，同時也能針對學生個別的發音問題，提供回饋與糾正的功能。本論文應用語音辨識、錯誤型態分析、及三維唇型動畫等技術，建立一套適合台灣人之發音輔助教學及矯正系統。本論文的主要技術包括：(1) 利用語音辨識技術，將使用者輸入的語音訊號轉變為音素序列，以進行發音錯誤分析。(2) 針對台灣學生可能的發音錯誤類型建立發音網路，偵測發音錯誤的位置及發音錯誤的型態，並針對錯誤的發音，進而提供適當的糾正。(3) 依據訓練語句之熵值 (entropy) 與使用者的個人發音錯誤類型動態的挑選測試句。(4) 運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫回饋系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。本論文研發之系統，未來將可提供於本國語者英語發音學習與發音矯正等範疇的實務應用。

1. 緒論

根據美國知名調查機構 IDC 統計，數位學習產業(E-Learning)的全球產值，將從 2003 年的 63 億美元，成長至 2004 年的 230 億美元，每年的複合成長率高達 54%。資策會市場情報中心統計也顯示去年台灣 E-Learning 市場有三・四五億的規模，粗估今年將達到六・二八億。近年來，由於政府對於英語學習的大力推動，使得市面上出現了琳琅滿目的相關書籍、補習班。以目前最熱門的全民英檢為例，不外乎分為聽、說、讀、寫四個部份。然而在「口說」這個部份卻較少有相關的方案可以自我評量。透過語音辨識技術的電腦輔助語言學習系統，使用者不僅可以在一個無壓力的環境下學習，更可以針對個別的發音問題，給予適當的糾正與回饋機制，讓使用者針對個人的錯誤反覆的練習，這不僅節省了人力、時間，同時也可達到較高的學習效果。因此，許多國外的學術單位或者一些商業軟體，都投入不少心力在 CALL System 的開發上。然而，這些市面上的商用軟體多數是套用現有的語音辨識引擎，例如 IBM 的 ViaVoice。而這些引擎原來都是針對母語為英語的使用者而設計的，所以如果針對母語為中文的使用者來說，其辨識率便會有所下降，而無法達到發音教學的目的[1][2]。由於目前大部份的系統針對發音的部份只是給定一個分數，然而我們希望能讓使用者可以得知其發音錯誤的型態，讓使用者知道自己到底發錯成什麼音。因此，本研究利用語音辨識的技術與錯誤型態的分析，建立一套適合台灣人的電腦輔助英語發音學習系統。在偵測發音錯誤類型的部份，首先利用本論文所找出的台灣大學生常犯的發音錯誤型態來建立辨識網路，藉由包含所有可能發音錯誤的辨識網路來找出發音錯誤的部份。且經由測試語句的挑選機制，希望能以較少量的句數歸納出使用者個人的發音錯誤型態。在回饋系統方

面，本論文則運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。

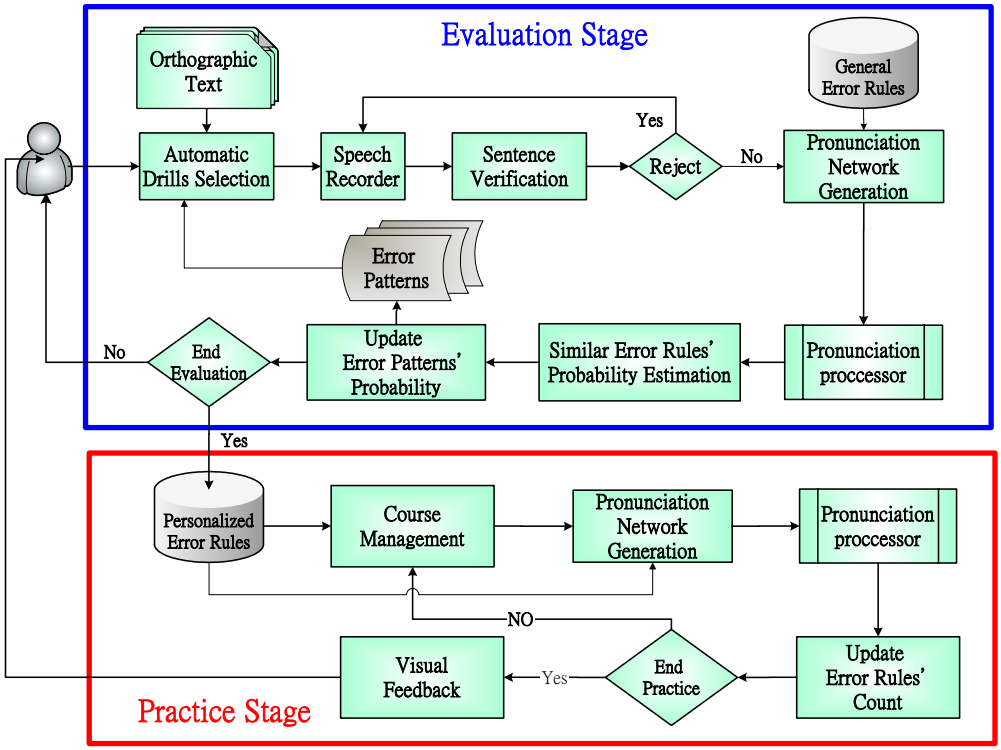
2. 相關研究

近幾年來相關的研究主要可分為發音評分與發音錯誤偵測兩部份。在發音評分的部份有 S.M. Witt 等人在 2000 年提出針對句子中的每個音素做評分，以 likelihood 為基礎的 Goodness of Pronunciation (GOP) [3]。另外，SRI EduSpeak System 則結合了 Phoneme posterior score、Duration score、及 Speech rate 等三種分數來對句子做評分[4]。Seiichi Nakagawa 等人在 2003 年則測試了 Log-likelihood、Likelihood ratio、Best log-likelihood、A posteriori probability、Phoneme recognition rate、Rate of speech 等各種評分方式來對句子評分，最後實驗發現結合 Log-likelihood、Best log-likelihood、Phoneme recognition rate、Rate of speech 與專家所評斷的結果有較高的相關性[5]。在發音錯誤類型的偵測方面，Yasushi Tsubota 等人在 2002 年利用 Pronunciation error network 來偵測日本學生發音錯誤的類型，並利用 LDA 針對發音錯誤的部份作驗證[6]。2004 年 Jong-mi Kim 等人則是根據韓國人的發音習慣建出一些可能的發音錯誤規則，辨識的時候利用這些發音規則來找出使用者發音與正確發音上的差異，並給與適當的建議[7]。

相關的 CALL application 在國外包括 SRI EduSpeak System[8]、ISLE system[9]、以及 PLASER system[10]等。EduSpeak System 主要是利用 speaker adaptation 技術結合 native 與 non-native 的語音，使得系統在辨識率方面有較好的效果。在功能方面主要就是結合 log-posterior score、duration score、及 speech rate 三個分數來對語音作評分。ISLE system 為一個針對義大利與德國人所設計的英語發音學習系統，此系統主要的功能為偵測發音錯誤的位置與發音錯誤的類型。發音錯誤偵測部份主要利用 HMM likelihood 對每個音素來做可信度分析。針對發音錯誤的音素，再利用事先定義好的錯誤規則來偵測錯誤類型。然而，此系統在錯誤類型偵測與回饋的部份效果較不理想。最後，PLASER system 則是針對母語為廣東話的中國人所設計。利用英文與廣東話的語料一起訓練聲學模型，在發音評估的部份則是計算之前所介紹過的 GOP 分數。根據評估的結果，75% 的使用者在使用此系統二至三星期後，在英文發音的正確性上均有所提升。在台灣較知名的 CALL application 有 My English Tutor (My ET)及 Train Speech。My ET 主要是針對發音、能量、音調、節奏四個部份分別給一個分數。並利用一個側面的舌位動畫與一些發音建議來提供使用者正確發音的回饋介面。Train Speech 主要的核心是利用 IBM Via Voice 的語音辨識器針對發音的部份來評分，且根據辨識的結果給使用者一些改正發音的建議。

3. 系統架構

本論文之整體架構，如圖一所示，主要分為“使用者發音錯誤類型評估”與“發音練習與視覺回饋”兩大部份。



圖一：系統流程

3.1 發音錯誤類型評估

這個部份的目的主要是要找出使用者常犯的發音錯誤類型。首先為了避免與標準語音內容差異過大，先針對輸入的語音做內容的驗證。本論文利用 log-posterior score 來對整句語音訊號做可信度分析，若分數小於門檻值則拒絕此語音的輸入。在發音錯誤類型偵測的部份，主要是透過語音辨識的方式，根據人工標記所找出來的發音錯誤規則將所有可能的發音(包含正確發音與錯誤發音)建立成對應的辨識網路，利用這樣的辨識網路來偵測發音錯誤的類型。由於我們希望能以較少的測試句來找出使用者個人的發音錯誤類型，利用計算句子的 entropy 與句子中還需納入考慮音素佔句子的比例來當做句子計分的準則。根據每一次的測試句所辨識出來的結果，我們可以計算出已測試過發音規則的發生機率，然而當測試語料量較少時，尚未出現在測試語料中之發音規則其機率則利用其它相關性較高的發音規則的機率來估計。每經過一次句子的測試，就需針對尚未被念過的句子重新計算其分數，然後挑選分數最大的句子當作下一次的測試句，直到每個音素的機率分布變化量小於我們所設定的門檻值時，即停止測試產生出個人的發音錯誤類型。

3.2 發音練習與視覺回饋

根據找出來的個人發音錯誤類型，我們挑選包含較多使用者常犯的發音錯誤的句子來讓使用者練習。在視覺回饋方面，本論文運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。運用 3D 虛擬人形動畫系統除了可增加趣味外，使用者也可以經由不同的角度來觀察唇型與舌位的變化。

4. 發音內容驗證

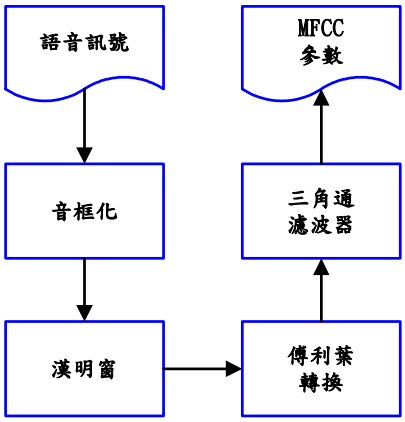
4.1 聲學模型

4.1.1 語料

針對母語為英語的聲學模型，我們使用 TIMIT 語料來訓練聲學模型。TIMIT 內容共 6300 句，由來自美國八個主要口音地區中的 438 位男性、192 位女性所錄製，每人錄製 10 句。我們以 TIMIT 建議的 4620 句做為訓練語料(語料總容量為 440 Megabytes、所有語料長度總和約為 3 小時 49 分 10 秒)來訓練母語為英語的聲學模型。由於本系統是針對母語為中文之使用者，因此我們也同時找了五位英語發音較佳的台灣人，錄製了一套台灣人口音的英語語料，來進行語者的調適。語料內容共 600 句，由 3 位男性、2 位女性大學生所錄製，每人錄製 120 句。

4.1.2 特徵參數擷取

要訓練聲學模型前必須先將訓練資料經過特徵參數的擷取。因此，對於處理語音這種高度差異的訊號時，需要找到能夠具有鑑別度特徵，這裡我們使用三十九維的梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC)，包含有十二階的頻譜值加上一階能量值，並取一階微分和二階微分。圖二為特徵參數的擷取流程圖：



圖二：特徵參數的擷取

4.1.3 聲學模型的建立

英文的一個音節是由一或多個音標所組成，每個音標均對應一種發音。TIMIT 語料定義了 62 個聲學模型，然而由於訓練語料的不足以及台灣人發音上準確度較低的情況下，我們不考慮同一個音標在不同位置下的重音情形。因此，我們定義了 42 個聲學模型，包含 40 個 monophones、1 個 silence model 與 1 個 short pause model。我們使用 HTK[10]來訓練聲學模型，表一為我們所定義的 40 個發音模型。

表 1：聲學模型與 KK 音標對照表

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

由於直接拿 TIMIT 訓練的聲學模型來辨識臺灣人口音的英文句，其辨識率便會有所下降，因此必須針對使用者的母語經過適當的調整。我們先使用 TIMIT 語料訓練初始的聲學模型，之後使用自行錄製的臺灣人口音的英文句，利用 MLLR (Maximum Likelihood Linear Regression)[11]調整使用 TIMIT 訓練出來的聲學模型。

4.2 語音內容驗證

在偵測使用者發音錯誤類型之前，我們希望使用者的語音內容與標準語音差異不致於過大，所以必須先針對使用者的語音做一個驗證的動作。我們的驗證機制主要是利用訓練好的 HMM Model，在已知使用者發音內容的情況下，做可信度的分析。

4.2.1 驗證機制

我們參考了兩個可信度分析的方法來建立本系統的驗證機制。其一是使用 LLR(log-likelihood ratio)[12]，然而此方法需要同時訓練 native speaker 與 non-native speaker 的聲學模型，因此需要有較大量的 non-native 語料。現階段受限於收集的台灣人口音語料的不足，於是我們使用 log-posterior probability score [13]來對整句語音做評分。假設 y_t 及 q_i 分別代表輸入語句中第 t 個 frame 的語音參數及其所對應的第 i 個音素。則事後機率 $P(q_i \mid y_t)$ 的計算如下式(1)：

$$P(q_i \mid y_t) = \frac{P(y_t \mid q_i)P(q_i)}{\sum_{j=1}^{M} P(y_t \mid q_j)P(q_j)}. \tag{1}$$

假設所有 model 出現的機率均相等即 $P(q_i) = P(q_j)$，因此上式(1)可近似為下式(2)：

$$P(q_i \mid y_t) = \frac{P(y_t \mid q_i)}{\sum_{j=1}^{M} P(y_t \mid q_j)}, \tag{2}$$

第 i 個音素之 log-posterior probability score ρ_i 便是計算此音素中所有對應 frame 之 log-posterior probability 平均：

$$\rho_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i \mid y_t), \tag{3}$$

其中 d_i 為此音素的 frame 總數。最後，整個句子的分數則為所有音素之 log-posterior probability 平均：

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\rho_i , \tag{4}$$

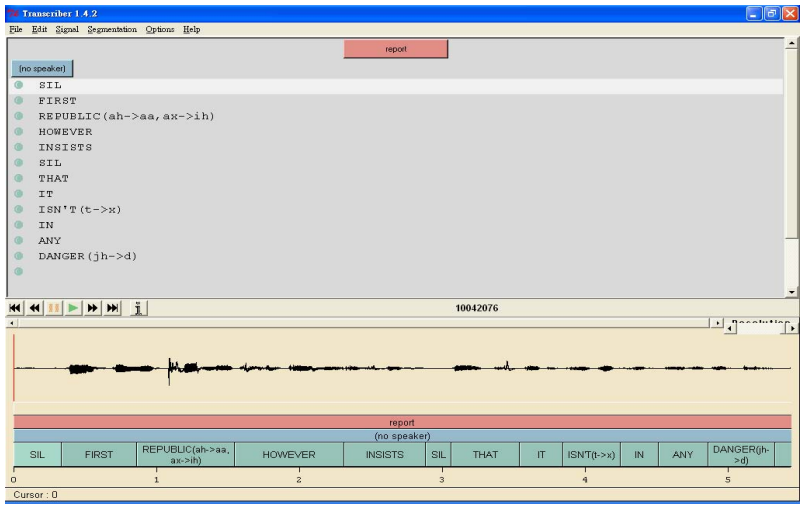
其中 N 為句子中音素的總數。計算出整個句子的分數後，將此分數與預先定好的門檻值比較，若分數小於門檻值，則拒絕此句語音的輸入，請使用者重新再輸入一次語音。相反的，若分數大於門檻值，則接受此句語音的輸入，接著進行發音錯誤類型的偵測。

5. 發音錯誤型態偵測

發音錯誤類型偵測主要是利用語音辨識的技術找出使用者發音錯誤的型態。首先，必須先定義出台灣人常犯的發音錯誤類型，在辨識時考慮此句子所有可能的發音錯誤的型態，建立其對應的辨識網路。透過此辨識網路，利用 Viterbi 演算法找出一條最佳的路徑，偵測出發音錯誤的類型。

5.1 台灣學生常犯之發音錯誤類型

我們從錄製的 2160 句台灣人口音的英文句中，盡可能的使每個音素出現的次數平均的情況下，挑選出了 1000 句英文句，其中包含 35 個男生、65 個女生，50 個非英語系學生、及 50 個英語系學生，語音內容約為 1 小時 6 分。我們將這 1000 句英文句由成大外文系 6 位受過轉寫訓練的學生做發音錯誤的標記。下圖是標記程式之介面[14]：



圖三：標記程式介面

根據標記的結果，我們整理出較常犯的錯誤類型。主要分為以下兩類：

A. 字轉音錯誤

這類型的錯誤主要由於字母拼字的關係，導致將英文字母轉成音標時發生錯誤。例如: cr**i**sis /k r aɪ s I s/，這個單字由於在字母上的拼字是 i，因此容易導致將/aɪ/這個發音念成/I/。下表列出幾個較常出現的錯誤類型：

表 2：字轉音錯誤

錯誤型態	Example
/ɑ/ -> /o/	T**o**m、J**o**hn
/z/ -> /s/	day**s**、hu**s**band
/ɔ/ -> /ɑ/	wr**o**ng、c**o**rporate
/aɪ/ -> /I/	cr**i**sis、d**i**versify
/æ/ -> /ɑ/	st**a**ff、**a**s

B. 發音錯誤

此類型的錯誤主要是由於母語的影響，導致發音的不正確。例如:full /f ʊ l/，/ʊ/這個音容易被念成/u/。這個發音錯誤主要是因為中文的母音並沒有長短之分，因此容易造成這類的錯誤。以下幾分別為母音與子音較常犯錯的類型：

表 3：母音發音錯誤型態

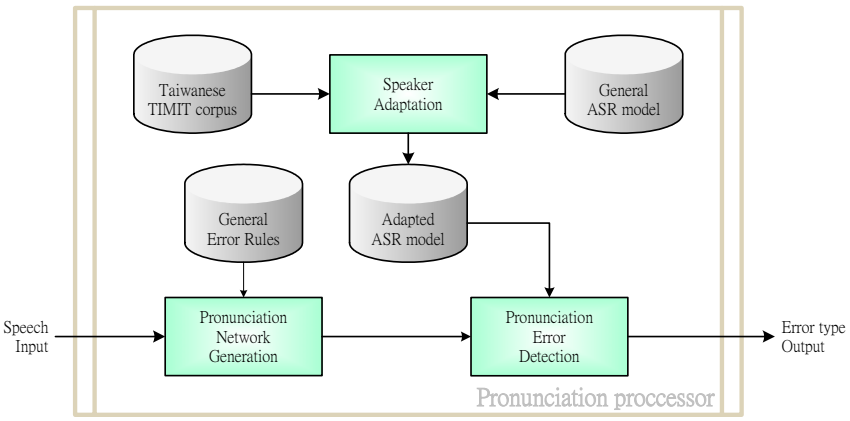
	錯誤型態	Example
短母音替代長母音	/i/ -> /ɪ/	seat、need
	/e/ -> /ɛ/	taken、made
	/u/ -> /ʊ/	fool
	/o/ -> /ɔ/	gold
長母音替代短母音	/ɪ/ -> /i/	year
	/ɛ/ -> /e/	weather、next
	/ʊ/ -> /u/	full、good
	/ɔ/ -> /o/	offer
/ɛ/替代/æ/	/æ/ -> /ɛ/	pan、matter
/ɑ/替代/ʌ/	/ʌ/ -> /ɑ/	husband、funny
非捲舌音替代捲音	/ɚ/ -> /ə/	either

表 4：子音發音錯誤型態

	錯誤型態	Example
非捲舌音替代捲舌音	/θ/ -> /s/	thank、think
	/ð/ -> /l/ or /d/	this、them
/ə/**替代節尾/r/**	/r/ -> /ə/	there、clear
/n/替代/ŋ/	/ŋ/ -> /n/	going
母音後面/r/省略	/r/ -> x	are、warm
母音後面/l/省略	/l/ -> x	almost、goal
音節節尾/n/省略	/n/ -> x	mine、one
停頓音節尾省略	/d/ -> x	stupid
	/t/ -> x	brought
	/k/ -> x	think
停頓音後增加/ə/	/d/ -> /də/	stupid
	/t/ -> /tə/	student
	/k/ -> /kə/	link

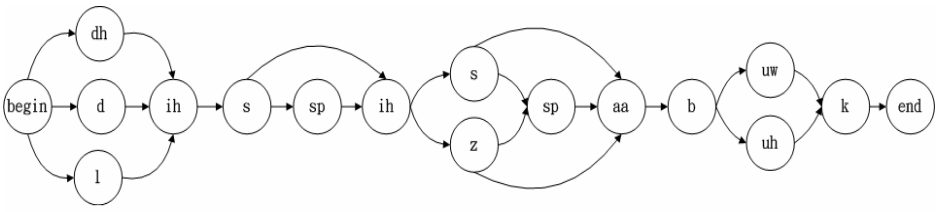
5.2 發音錯誤類型的偵測

發音錯誤的偵測主要是先針對句子建立對應的辨識網路，利用此辨識網路來辨識語音內容。其流程圖如圖四：



圖四：發音錯誤偵測流程

根據上一節介紹的發音錯誤類型，我們可以將所有可能的發音建立在辨識網路中(包括發音正確與所有可能的發音錯誤)，例如: This is a book 考慮所有發音的可能，其辨識網路如圖五：

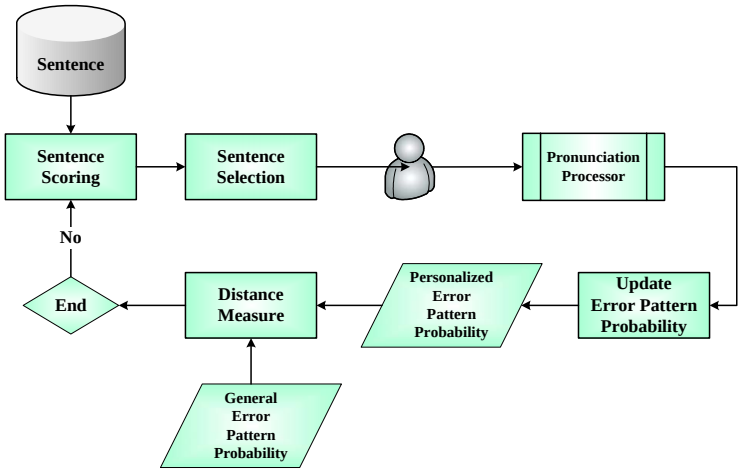


圖五：發音辨識網路

有了辨識網路後，利用 Viterbi 演算法找出一條最佳的路徑，將辨識結果與已知的語音內容比對，如此就能找出使用者該句發音錯誤的型態。

6. 最佳訓練資料之選取

最佳訓練資料選取之流程圖如圖六所示。



圖六：最佳訓練資料選取流程

在此我們希望能以最少量的測試句來找出使用者個人的發音錯誤型態。首先，針對資料庫中的測試句分別給予計分，挑選出分數最高的句子作為測試句。根據找出的錯誤型態，更新個人錯誤型態的發生機率。針對每個可能產生發音錯誤的音素，比對此音素與大量資料所統計出來的錯誤型態中機率分布的差異，倘若在連續兩測試句中所計算出的變動量已小於某個門檻值，則表示針對

這個音素使用者的發音錯誤機率已經達到一個穩定的狀態，所以在挑選下一句測試句時，便不需將此音素列入計分的考量中。依照這樣的流程，反覆的挑選測試句直到所有的音素均已不需再考量為止。

6.1 訓練語句之計分與挑選

本節首先介紹訓練語句的計分方式。對於語料庫中第 *i* 句訓練句，其 sentence score 計算方式如下式：

$$Sentece\_Score_i = ES_i \times TS_i, \tag{5}$$

其中 ES_i 表示第 *i* 句訓練句的 Entropy，計算方式如下：

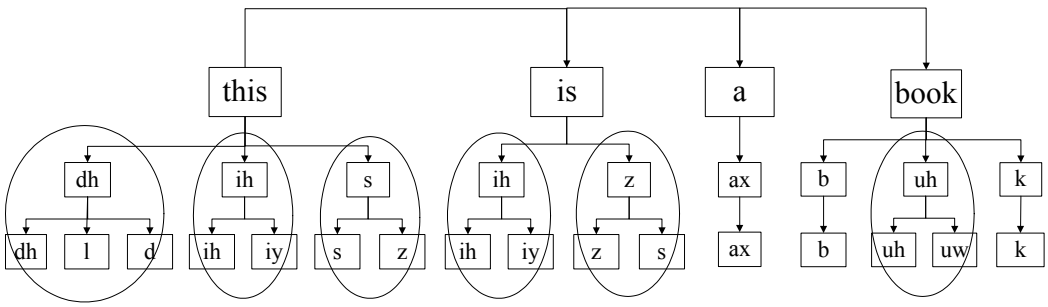
$$ES_i = -\sum_{j=1}^{m}\sum_{k=1}^{n} P_{jk}\log\left(P_{jk}\right), \tag{6}$$

m ：句子中可能發生錯誤的 phone 個數

n ：第 *j* 個 phone 中所有可能出現的情形

P_{jk} ：第 *j* 個 phone 的第 *k* 個情況的機率值

例如一訓練句 “this is a book”，其可能的發音型態如下圖所示：



圖七：句子發音型態範例

由這個例子可以得知，可能錯誤的音素個數有 6 個，因此分別對這 6 個可能錯誤的音素計算其 entropy，將此 6 個的 entropy 總和當成這整句的 ES_i 分數。我們使用 entropy 的目的在於找出最不能確定使用者會念對或會發生發音錯誤的句子。比如說以 “ih” 這個音素為例，當 “ih 念對成 ih” 與 “ih 念錯成 iy” 機率均為 0.5 的情況下，其 entropy 的值為最大，即表示我們較難以判定此音素是否較易被念錯或較易被念對。相反的當 “ih 念對成 ih” 與 “ih 念錯成 iy” 機率有一方為 1 的情況下，其 entropy 的值為最小，即表示我們可以判定此音素易被念錯或易被念對。基於上述的理由，我們考慮 entropy 越高的句子越優先讓使用者測試。

除了 ES_i 值外，第(5)式另一個變數為 TS_i。TS_i 表示計算句子中還需要納入計分考慮的音素佔句子的比例。其計算方式如下：

$$TS_i = \frac{NTC_i}{TCi}, \tag{7}$$

NTC_i：句子中尚需考慮的 phone 個數

TC_i ：句子中所有可能發生錯誤的 phone 個數

最後，根據我們人工標計發音錯誤的句子中，我們可以計算出每個音素的正確與錯誤發生機率。因此，我們將這些機率當做 general model。當使用者做測試時，我們也能根據測試結果計算出每個音素的正確與錯誤發生機率，之後利用 discrete KL distance 的方式來計算目前某個音素的機率

分布與 general model 的差距，假設累積到第 i+1 句測試句其與 general model 的差距相較於累積到第 i 句測試句其與 general model 的差距變動不大時，表示對於此音素而言，使用者發生對或錯的機率已趨於穩定，因此我們可以不需再將此音素納入句子挑選的考慮。以下將介紹對於某音素停止條件的計算方式。

若某個音素 a 有 N 種可能的唸法：$a_1 \sim a_N$，則定義 $p_{general}(a \to a_n)$ 表示在 general model 中音素 a 被唸成 a_n 的機率；$p_1^i(a \to a_n)$ 則表示在累積到第 i 句測試句時，音素 a 被唸成 a_n 的機率。則 KL distance 的計算如下：

$$KL_i = \sum_{n=1}^{N} p_1^i(a \to a_n) \log_2 \left(\frac{p_1^i(a \to a_n)}{p_{general}(a \to a_n)} \right), \tag{8}$$

$$KL_{i+1} = \sum_{n=1}^{N} p_1^{i+1}(a \to a_n) \log_2 \left(\frac{p_1^{i+1}(a \to a_n)}{p_{general}(a \to a_n)} \right), \tag{9}$$

上式中 KL_i 為累積到第 i 句與 general model 的差距，而 KL_{i+1} 則為累積到第 i+1 句與 general model 的差距。因此我們是以下式來做為停止考慮的參考標準：

$$\Delta KL = \left| KL_{i+1} - KL_i \right|, \tag{10}$$

當 ΔKL 小於某個 threshold 且此音素已被念過的次數大於五次以上時，即可停止考慮此音素。利用上述的計分方式，我們挑選出分數最高的句子當做下一次的測試語句，每經過一次測試均需根據使用者發音錯誤的結果，重新計算還未被測試句子的分數。因此，針對不同使用者之間發音錯誤類型的不同，同一測試句的分數變會有所不同。如此，便能依據個人化的發音錯誤習慣，挑選的句子變會有所差異。

6.2 發音錯誤類型機率的估計

由於使用者在測試的過程中，當某些音素在測試資料中尚未出現時，我們希望利用估計的方式來計算出其發生機率。因此，我們假設某些發音習慣會導致類似的錯誤發生。從我們人工標記的發音錯誤類型中，我們發現當某個測試者發生了/ð/念錯成/l/時，/ɚ/也容易被念錯成/ə/。由聲學的角度來看，這類的錯誤可能是由於捲舌音發的不好，導致念錯成非捲舌音。再舉例來說，由於在中文的在母音的部份沒有長短之分，所以長母音念錯成短母音的情況也容易同時出現。因此，從我們人工標記的發音錯誤的語料中，利用計算 Mutual Information 的方式找出不同發音規則間的關係(例如：/ð/→/l/表示一種發音規則)。假設 X、Y 分別代表兩個不同的發音規則，其 Mutual Information 計算方式如下：

$$I(X;Y) = \sum_X \sum_Y P(x_i, y_i) \log \frac{P(x_i, y_i)}{P(x_i) P(y_i)}, \tag{11}$$

下表列出幾個 Mutual Information 較高的發音規則：

表 5：相關性較高之發音規則

Rule X	Rule Y
/i/ -> /ɪ/	/e/ -> /ɛ/
/e/ -> /ɛ/	/o/ -> /ɔ/
/u/ -> /ʊ/	/o/ -> /ɔ/
/ð/ -> /l/	/ɚ/ -> /ə/
/o/ -> /o/	/e/ -> /e/
/e/ -> /e/	/u/ -> /u/
/l/ -> x	/r/ -> x
/k/ -> /kə/	/t/ -> /tə/
/æ/ -> /ɛ/	/l/ -> x
/ð/ -> /l/	/i/ -> /ɪ/

利用我們所設定的門檻值，我們從 80 個發音規則中(包含正確的發音規則)，找出了 53 組相關性較高的發音規則。由上述的例子中看出，大部份的相關性較高的發音規則可符合聲學方面的特性，然而因為是利用統計的方式，所以有些找出來的發音規則無法從聲學的角度來解釋。

經由上述的方式找出相關性較高的發音規則後，我們可以藉由下列的方式估計出機率值：

$$P(X,Y)=P(X\mid Y)P(Y)=P(Y\mid X)P(X), \tag{12}$$

$$P(X)=\frac{P(X\mid Y)}{P(Y\mid X)}P(Y) \quad , \quad P(Y)=\frac{P(Y\mid X)}{P(X\mid Y)}P(X) \quad , \tag{13}$$

其中 $P(X\mid Y)$ 與 $P(Y\mid X)$ 我們事先從經過人工標記過的大量語料中訓練出來，因此當使用者在測試語料中只出現 X 或 Y 其中之一，就可藉由上述的方式估算出另一方發生的機率。

假設與 X 相關的較高的發音規則有：$R_1, R_2, R_3 \ldots R_n$，因此 X 出現機率的計算方式如下：

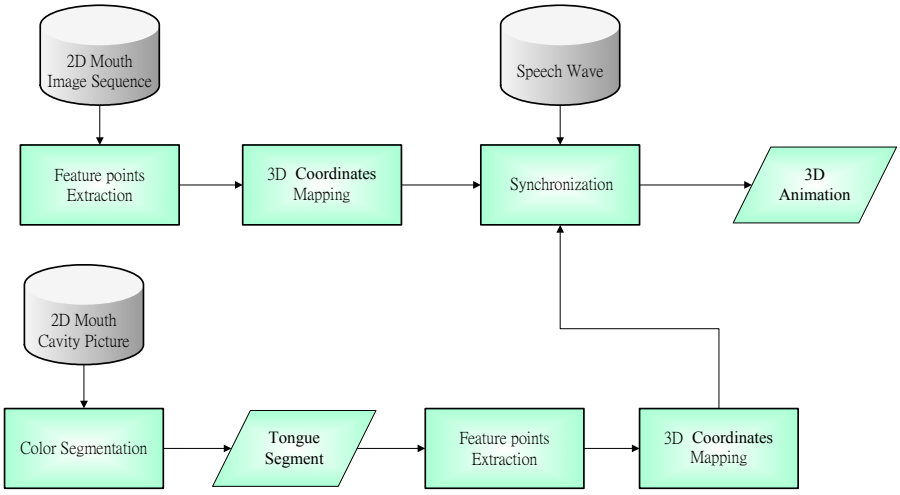
$$P(X)\approx\sum_{i=1}^{n}W_i\times\frac{\bar{P}(X\mid R_i)}{\bar{P}(R_i\mid X)}\times P(R_i), \tag{14}$$

$$W_i=\frac{\bar{P}(X,R_i)}{\sum_{i=1}^{n}\bar{P}(X,R_i)}, \tag{15}$$

上述的權重值也可事先從經過人工標記過的大量語料中訓練出來，因此，我們假設尚未出現在測試語料中的發音規則間的相關性與大量資料所訓練出來的 joint probability 是不變的，所以我們可以藉由這樣的方式估算出尚未出現的發音規則機率。

7. 視覺回饋

我們運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。圖八是 3D 動畫合成之流程：

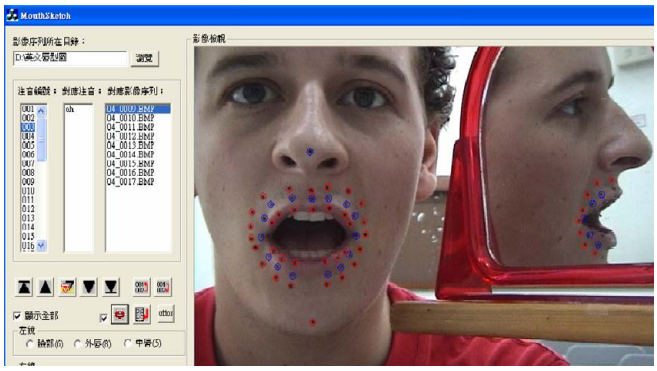


圖八：三維動畫合成

7.1 三維唇型動畫

7.1.1 唇型特徵點擷取

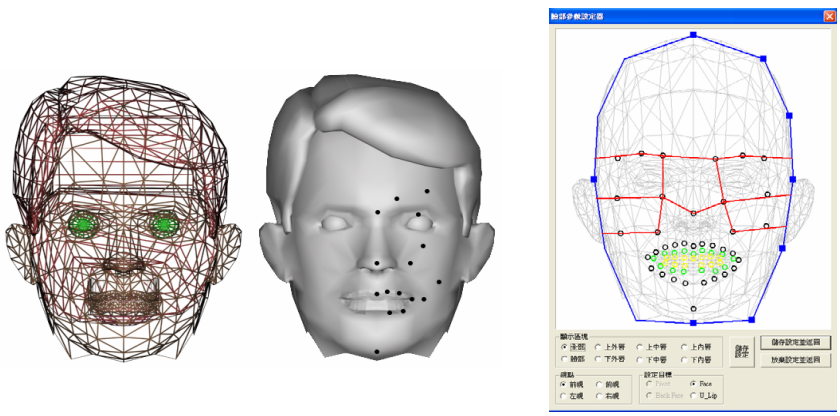
首先經由一組事先拍攝的唇型變化影片來擷取出 62 個唇形特徵點變化參數。因此，根據我們所定義的 40 個聲學模型，分別拍攝其發音之唇型影帶。接著利用 Optical Flow[15]動態偵測的方式自動偵測唇形週圍幾個特徵點的變化。圖九為唇型特徵點偵測之結果：



圖九：唇型特徵點偵測

7.1.2 三維座標轉換

擷取出唇形特徵點在三個座標軸中的位移之後，我們必須先在 3D 模型中，定義出此 62 個特徵點的位置，其界面如圖十所示。



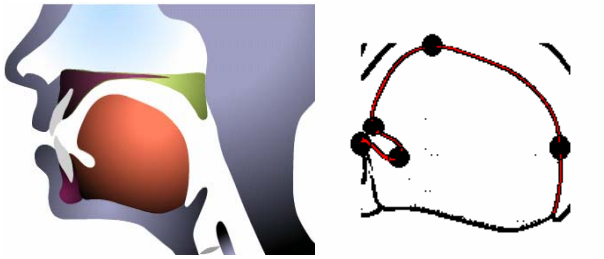
圖十：三維唇型控制點定義

其餘網格中的點則由鄰近的控制點來控制，其位移量為鄰近控制點位移量乘以個別的權重之總和，且控制點的權重與控制點到網格點距離的平方成反比。

7.2 三維舌位動畫

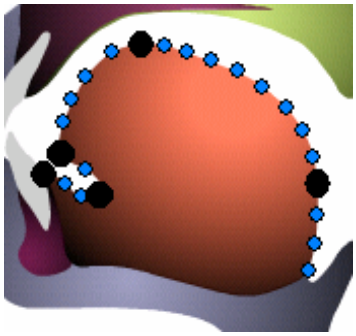
7.2.1 舌頭特徵點擷取

由於直接在舌頭上貼 Sensor 來偵測發音時舌位的變化是不容易的，因此我們實作一非侵入式 3D 舌位測量方法。首先我們由網路上的開放資源中蒐集了每個發音的口腔 2D 圖(如圖十一) [16]，由於要找出舌頭的部份，因此先將顏色由 RGB 轉換為 HSI 後，便可很容易的將舌頭的部份給切割出來。

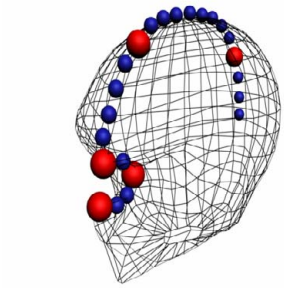


圖十一：發音口腔圖

由口腔的剖面圖我們發現通常轉折的地方(舌尖、舌根)變化較大，因此我們將這幾個轉折的地方定義為舌頭主要的特徵點(如圖十一中的黑色點)。根據不同的發音，我們必需記錄每個特徵點在不同時間下的位移量，所以利用影像處理的技術，自動偵測出舌頭的轉折點。首先將舌位圖經由 sobel operator 做邊界偵測，然後利用八鄰域的方式做邊界追蹤以擷取出特徵點的位置。為了讓動畫可以更精細，除了 5 個主要轉折點外我們還擷取了其他的特徵點(如圖十二)。



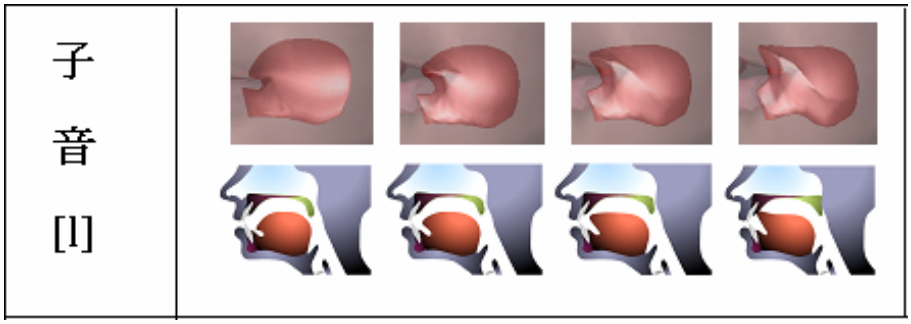
圖十二：舌頭特徵點偵側結果



圖十三：三維舌頭控制點定義

7.2.2 三維座標轉換

有了每個發音時的舌位特徵點位移變化量，接下來需將 2D 舌頭特徵點座標 map 到 3D 空間，並利用特徵點自動計算出 3D 模型網格點位移量。利用這些控制點座標與各控制點對相鄰網格點的權重(與網格點的距離成反比)，便可計算出 3D 模型中每個時間所有網格點的位移量。圖十四為一實作之舌頭 3D 動畫：



圖十四：舌位動畫與發音口腔圖比較

8. 實驗結果與討論

8.1 語音訊號切割實驗

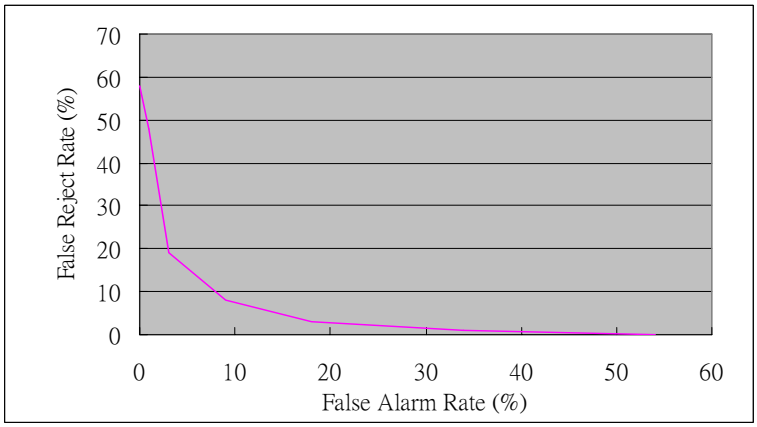
我們從所錄製的臺灣人口音之語音訊號中，挑選其中發音正確的 300 句語音來做訊號切割實驗，語音總長約 21 分 36 秒。平均每句訓練句包含 6.87 個 Word、35.26 個 phoneme。由於此 300 句有經過人工標記的動作，因此在我們可直接將利用 Forced Alignment 切割出來的時間點和人工標音出來的結果作比較。由於人工標音的部份只有標記到字，所以我們假設若切割出來字的時間區段和該字在人工標音下的時間區段前、後各相差在 0.1 秒以內，則稱此字的切割結果為正確。在此我們比較兩個不同的聲學模型：一個為使用 Native Speaker 所訓練出來的模型，另一個為使用台灣口音之英文語料經由語者調適所產生的模型。表 6 為英文語音訊號切割的正確率：

表 6：語音訊號切割正確率

模型 / 正確率	Model without Adaptation	Model with Adaptation
Word 時間正確率	84.63 %	87.93 %

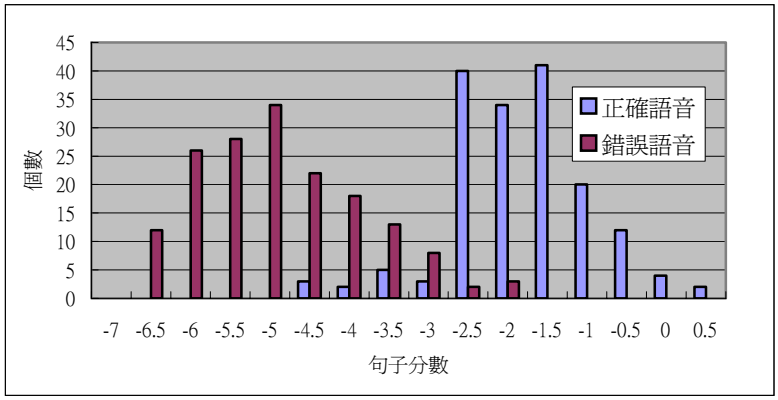
8.2 語音內容驗證實驗

將輸入的語音訊號做可信度的分析後，若分數高於門檻值則表示此輸入語音是可靠的，反之若小於門檻值，則拒絕此語音進入系統。因此，為了找到較佳的門檻值，我們取 166 句語音當作標準語音內容，語音的總長度約為 12 分 42 秒。此外取另外 166 跟標準語音內容不同的語音當作錯誤語音，錯誤語音總長度約為 10 分 38 秒。我們利用正確語料被拒絕(False Rejection, Type Ⅰ Error)及錯誤語料被接受(False Alarm, Type Ⅱ Error)的 ROC 關係圖(Receiver Operator Characteristic)來找出最佳的門檻值：



圖十五：False Rejection Rate 與 False Alarm Rate 之 ROC

我們利用正確拒絕加錯誤接受的和最小來找出門檻值。因此，當門檻值為-3.2 時有最佳的結果，其正確接受率為 93.4%，正確拒絕率為 6.6%，錯誤拒絕率為 95.2%，錯誤接受率為 4.8%。經由以上的實驗訂出門檻值之後，我們使用另外的 166 句正確語音與 166 句錯誤語音作測試，正確接受率為 92.2%、正確拒絕率為 7.8%、錯誤拒絕率為 96.9%、錯誤接受率則為 3.1%。圖十六為正確語音與錯誤語音經由計算可信度後的分數分布圖：



圖十六：語音內容驗證結果分布圖

8.3 發音錯誤偵測實驗

我們從所錄製的臺灣人口音之語音訊號中，挑選其中的 300 句語音來做發音錯誤偵測實驗，語音總長約 22 分 33 秒。將利用發音網路所辨識出來的結果與人工標記的答案比較來計算正確率，結果如下表所示：

表 7：發音錯誤偵測實驗結果

模型 / 正確率	Model without Adaptation	Model with Adaptation
Phoneme 正確率	72.59 %	78.18 %

8.4 最佳訓練語句挑選評估

目前我們所使用的測試句共 166 句，我們找了 5 位測試者每人均唸完所有的 166 句。以下針對三個不同的挑選句子方法做評估：

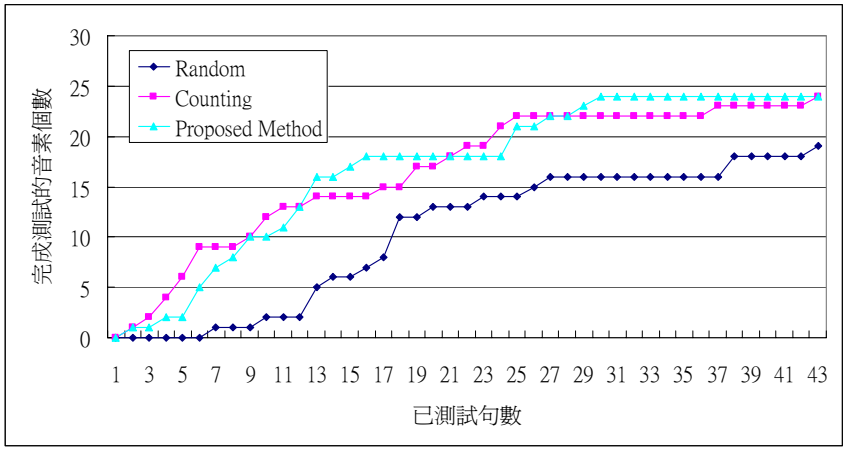
Random：隨機挑選測試語句

Counting：只考慮句子中尚需納入挑選的音素總數

Proposed Method：本論文所提出的句子計分方式

8.4.1 句子總數的評估

所需測試的音素總數共 24 個，根據我們所定義的音素完成測試的條件下，比較此三種方式所需測試的句子總數。

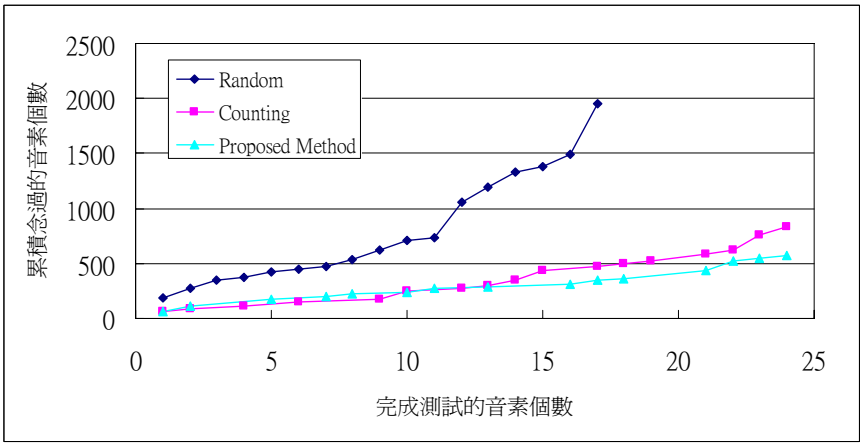


圖十七：句數評估實驗結果

由實驗結果可知，我們所提出的方法明顯比用 Random 的方式所需的句數較少。除此之外，與 Counting 比較之下雖然在完成 13 個音素測試前 Counting 所需句子數較少，然而比較所有完音素均完成測試時所需的句子數，我們所提出的方法需要 30 句，相較 Counting 所需的 43 句，所需的總句數較少。由以上的實驗結果顯示，在完成所有的音素測試下，我們所提出的方法有較佳的結果。

8.4.2 累積唸過音素總數的評估

所需測試的音素總數共 24 個，根據我們所定義的音素完成測試的條件下，比較此三種挑選句子的方式下，所累積唸過的音素總數。



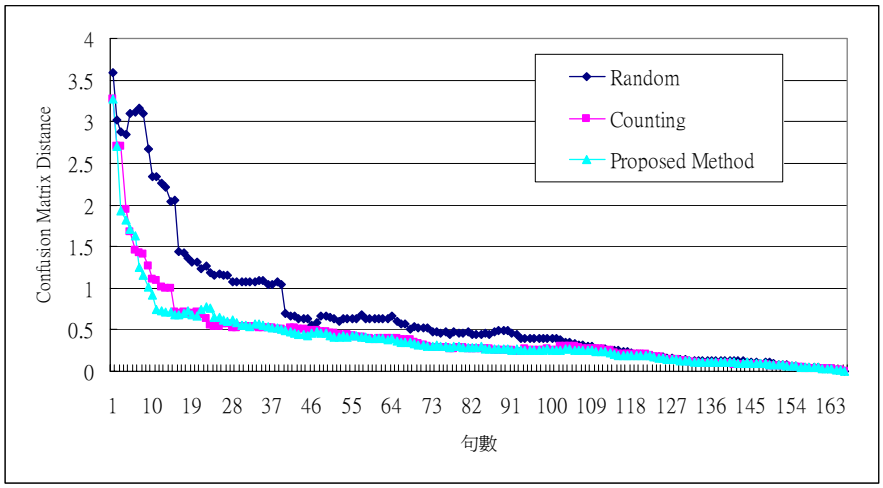
圖十八：音素總數實驗結果

由實驗結果可得知，我們所提出的方法明顯比 Random 所累積唸過的音素個數較少。在完成 10 個音素測試後，Counting 所累積唸過的音素個數為 251 個，我們所提出的方法所累積唸過的音素個數只有 239 個。在此之後，我們所提出的方法均較 Counting 所需唸過的音素總數少。當所有的音素均完成測試的情況下，Random 所累積唸過的音素個數為 1956 個，Counting 所累積唸過的音素個數為 829 個，我們所提出的方法為 570 個。由以上的實驗結果顯示，我們所提出的方法句子的平均長度相較其他兩個方法短。

8.4.3 錯誤型態機率的評估

我們由使用者所唸的 166 個句子中，可以計算出所有發音規則的出現機率。因此，我們可以建立一個 40 乘 40 的 Confusion Matrix。在此我們比較三種不同的挑選句子的方式，觀察其 Confusion

Matrix 在每經過一句測試資料後，我們將 Confusion Matrix 看成一個維度為 1600 的 vector，利用簡單的 Euclidean Distance 的方式計算與 166 句所計算出來的 Confusion Matrix 距離。如圖十九所示：



圖十九：使用三種不同方式時 Confusion Matrix 收斂結果

由圖中我們可以看出，用隨機的方式挑選句子 Confusion Matrix 的機率值收斂的最慢。運用我們所提出的挑句子的方式，可以發現少量的句子下，Confusion Matrix 機率值收斂的較快速，可以較快逼近接近真正的機率值。

9. 結論與未來展望

本論文提出一個以分析語者發音錯誤型態來輔助英語學習之方法。根據我們所錄製的台灣人口音之英文句，我們統計出了台灣大學生在英語發音上較常見的發音錯誤類型。我們利用包含可能的發音錯誤類型所建立的發音網路來偵測使用者發音錯誤的型態，且利用統計方法依據訓練語句之熵值(entropy)與使用者的發音錯誤類型動態的挑選測試句。由實驗中我們可以發現所提之方法可以有效降低訓練語句之數量，提高學習者之學習成效，充份顯示本論文所提之方法在實用上具有一定之效果。除此之外，在回饋系統的部份，我們實作了一個 3D 虛擬人物動畫，透過這樣的 3D 動畫能夠讓使用者以多個不同的角度來觀察發音時唇型與舌位動作，更可清楚的呈現發音的完整過程。在未來研究方向方面，可以從以下幾個地方來著手：(1) 自動新增錯誤型態：目前由於我們的錯誤類型是事先定義好的，因此無法動態偵測出不在定義中的發音錯誤。因此，若能根據系統不斷的使用的過程中，自動新增出一些個人化的發音錯誤類型，如此便能更有效的改正使用者發音錯誤。(2) 英文發音錯誤類型與中文發音的關係：由本論文所找出來的發音錯誤類型中，不難發現之所以會導致發音錯誤，其實與使用者本身的母語有一定的關係存在。因此，若能分析出中文與英文在子母音上的異同，便可更有效的從中文的發音習慣上來給予使用者較好的發音建議。(3) 錯誤發音的糾正：目前的系統在這個部份的一直沒有較好的成效。若能從聲學的角度或母語發音上的習慣來糾正錯誤發音，藉此建立一套更好的發音錯誤糾正的機制。

參考文獻

[1]. T. M. J. Munro and M. Carbonaro, “Does Popular Speech Recognition Software Work with ESL Speech?”, TESQL Quarterly 34, pp.592-603, 2000

[2]. D. Coniam, “Voice Recognition Software Accuracy with Second Language Speakers of English”, System 27, pp.49-64, 1999

[3]. Witt,S.M. and Young, S.J. “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning”, Speech Communication 30, 95-108. 2000

[4]. Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., and Butzberger, J. “The SRI EduSpeak(TM) System: Recognition and Pronunciation Scoring for Language Learning", Proceedings of InSTILL 2000, Dundee, Scotland, 123-128. , 2000

[5]. Seiichi Nakagawa, Kazumasa Mori, Naoki Nakamura “A Statistical Method of Evaluation Pronunciation Proficiency for English Words Spoken by Japanese”, Eurospeech 2003

[6]. Yasushi Tsubota, Tatsuya Kawahara, Masatake Dantsuji “CALL System for Japanese Students of English Using Pronunciation Error Prediction and Formant Structure Estimation” InSTILL 2002

[7]. Jong-mi Kim, Chao Wang, Mitchell Peabody, Stephanie Seneff “An Interactive English Pronunciation Dictionary for Korean Learners” ICSLP 2004

[8]. Menzel, W., Herron, D., Bonaventura, P., and Morton, R. (2000). “Automatic detection and correction of non-native English pronunciations”, Proceedings of InSTILL 2000, Dundee, Scotland,

[9]. Mak, B., Siu, M.H., Ng, M., Tam, Y.C., Chan, Y.C., Chan, K.W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., Lo, J. (2003). “PLASER: Pronunciation Learning via Automatic Speech Recognition”, Proceedings of HLT-NAACL 2003, Edmonton, Canada, 23-29

[10]. Steve Young, *The HTK Book version 3*, Microsoft Corporation, 2000

[11]. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models C. J. Leggetter and P. C. Woodland, Computer Speech and Language (1995) 9, 171–185

[12]. Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. (1999). “Automatic Detection of Phone-Level Mispronunciation for Language Learning”, Proceedings Eurospeech '99, Budapest, Hungary, 851-854.

[13]. H. Franco, L. Neumeyer, and Y. Kim, “Automatic Pronunciation Scoring for Language Instruction”, Proc. ICASSP , pp.1471-1474, 1997

[14]. Transcriber: Development and use of a tool for assisting speech corpora production Claude Barras , Edouard Geoffrois , Zhibiao Wu , Mark Liberman , speech communication 2001

[15]. Horn, B.K.P and Schunck, B.G., “Determining Optical Flow”, Artificial Intelligence, vol.17, nos.1-3, pp.185-203 (1981-8).

[16]. http://www.uiowa.edu/~acadtech/phonetics/english/frameset.html