

以語音辨識與評分輔助口說英文學習

¹陳江村 ¹羅瑞麟 ¹張智星 ^{1,2}李俊仁

¹國立清華大學 資訊工程系

新竹市光復路二段 101 號

E-mail : { jtchen, roro, jang }@wayne.cs.nthu.edu.tw, cjlee@cht.com.tw

²中華電信研究所

桃園縣楊梅鎮民族路 5 段 551 巷 12 號

摘要：

本論文提出利用音訊處理和語音辨識的技術，進行英文語音評分。依英文發音的特有性，進行評分系統的各個模組之設計、製作及實驗，期許建立一套合理的英文語音評分系統。

本論文結合三項技術——「說話驗證」、「語音訊號切割」和「英文語音評分」達成輔助口說英文之學習。「說話驗證」利用說話驗證的可信度評估，依此拒絕文句內容(context)不正確的評分語句。「語音訊號切割」提供一個將語音訊號切割出每個音素時間區段的方法，以預先訓練好的英文發音聲學模型當作切割依據，爾後經由語音辨認技術，以合適的聲學模型切割出正確的發音區段。

「英文語音評分」為評分系統的核心，使用的評分方式是比較標準語音和評分語音的相似度。本文採用四個評分參數——音量強度曲線、基頻軌跡曲線、發聲急緩變化及 HMM 對數機率差異進行語音相似度評分。經由實驗，對於一個合理的評分系統，我們得到音量強度曲線的權重為 7.45%，基頻軌跡曲線的權重為 22.40%，發聲急緩變化的權重為 17.24%，HMM 對數機率差異的權重為 52.91%，經由實驗證實本系統之語音評分與人工評分具有約 60%的正相關性。

關鍵詞：

語音辨識、語音評分、HMM、聲調辨識、Viterbi Search、音量強度、音高向量、梅爾式刻度倒頻譜、Forced Alignment、Downhill Simplex Search

1 前言

由於近年來電腦計算能力的提昇以及語音辨識技術的進步，語音處理在我們日常生活上的應用與日俱增，如語音辨識、語音合成、語者識別等等。其中，在跨國界的語言學習中，以電腦輔助使用者進行非母語學習(CALL, Computer-Assisted Language Learning)已受到相當重視，各方也紛紛投入相關的研究[10][11][18][15][20]。

電腦輔助發音訓練(CAPT, Computer-Assisted Pronunciation Training)可視為是語音辨識和圖形比對(Pattern Matching)兩項技術的結合。本論文研究主題，包含「說話驗證」、「語音訊號切割」以及「英文語音評分」三個部份，希望融合目前語音辨識和圖形比對的技術，對使用者進行公正的語音評分。

在語音評分系統中，如果能先濾除內容和標準語音完全不同的評分語音，可以使整個語音評分系統更具公信力。本論文運用了可信度評估的技術來達成說話驗證(Utterance Verification)。確保了評分語音內容的正確性後，對於評分語音我們使用 HMM(Hidden Markov Model)切割出每個音素(phoneme)的時間區段，使用高辨識率的 HMM 聲學模型可確保切割出來的音素區段有一定的可信度及正確率。在英文語音評分部份，我們利用標準語音資料來進行一種較為主觀的評分方式，主要使用圖樣比對(Pattern Matching)的方法，根據四個評分參數：音量強度曲線(Magnitude)、基頻軌跡曲線(Pitch Contour)、發聲急緩變化(Rhythm)以及 HMM 對數機率差異(HMM Log-Likelihood)，將評分語音和標準語音的資料逐音素地來做比較，以期找出評分語音和標準語音的差異程度。

2 相關研究

1997 年時，C. Cucchiariini、H. Strik 及 L. Boves 以荷蘭語為主，定義了 Total Duration of Speech no/plus Pause、Mean Segment Duration、Rate of Speech 以及 Global Log-Likelihood，經由類似的實驗後得出 Global Log-Likelihood 對於人類主觀評分占較重的比重[19]。1999 年 L. Neumeyer、H. Franco、V. Digalakis 和 M. Weintraub 以法語語料庫進行實驗，採用 HMM Log-Likelihood、Normalized Acoustic、Segment classification、Segment Duration、Timing 當作其實驗的評分參數，經由實驗後得出了 Normalized Acoustic 在評分系統和語言專家給予的分數中，其相關性高於 Segment Duration[10]。

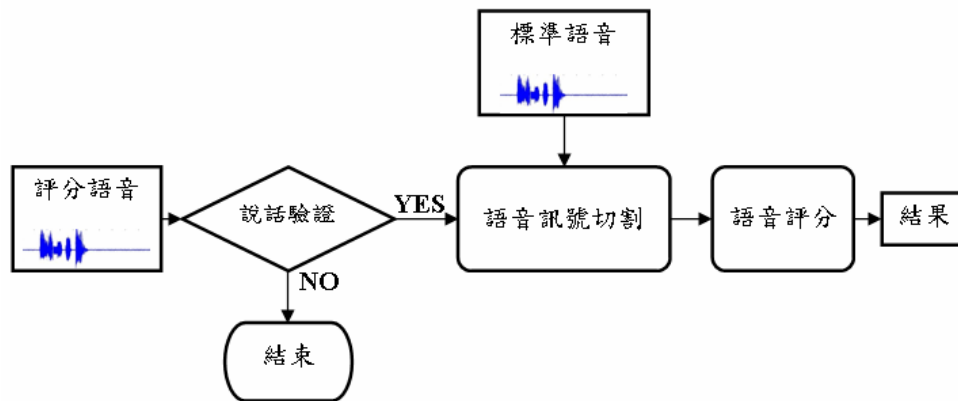
至於英文的語音評分，2002 年清華大學的李俊毅以梅爾倒頻譜、Magnitude 及 Pitch 三種評分參數觀察對英文語音評分的影響，其實驗發現梅爾倒頻譜參數對英文語音評分的重要性最大，另外他也將各個特徵的差異程度轉換成分數，以回饋給使用者參考[15]。2004 年陳江村和張智星等人利用了 HMM 和 GMM 分別對中文的發音和聲調進行評分，並以 Downhill Simplex Search 進行了評分系統參數的最佳化，以求達到和中文專家一致

的評分標準[20]。

接下來的論述中，首先我們提出實作「說話驗證」的方法，包含聲學模型相似度排名、驗證系統的建立及驗證系統的可靠性等。接著是「語音訊號切割」。這部份包含隱藏式馬可夫模型(Hidden Markov Model)的訓練和以維特比演算法(Viterbi algorithm)為基礎的語音訊號切割技巧。再來是「英文語音評分」，其中提到了關於評分參數的擷取、評分參數正規化、圖樣比對流程、評分機制的建立等，並設計實驗以求出各評分參數在英文語音評分中的權重，以符合人類專家對英文語句好壞的看法。最後是總結及今後研究工作的展望。

3 英語評分系統架構

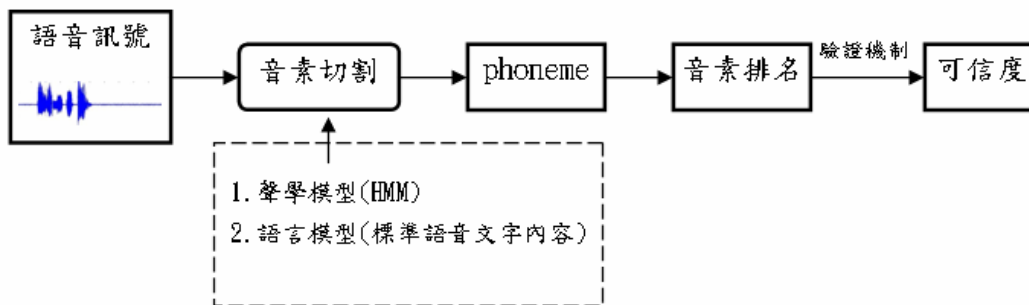
在此英文語音評分系統中，首先以說話驗證做為第一道檢視關卡，爾後以聲學模型來對標準語音及評分語音切割音素的時間區段，再將這些資訊送至英文語音評分系統的核心，利用各種評分參數，逐音素地比較評分語音和標準語音的差異程度。本文所提之英文語音評分系統架構流程，如圖表 1所示。



圖表 1 英文語音評分系統流程圖

3.1 說話驗證

所謂的說話驗證(Utterance Verification)，就是我們可以針對不同的評分語音產生判斷數值，並依此而對該評分語音內容的正確性做出判斷[1]。此說話驗證流程如圖表 2 所示，當驗證系統接收到語音訊號後，分別對每個音素進行語音辨識，之後再依辨識結果的機率值排名並配合驗證機制給予最後的可信度值。



圖表 2 說話驗證系統流程圖

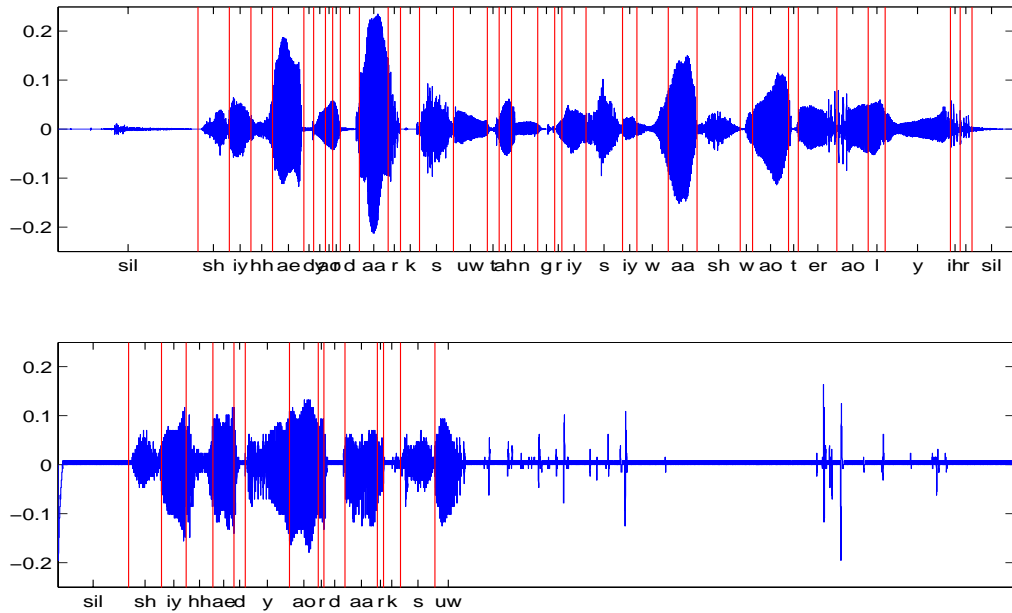
3.1.1 音素切割

這裡切割用的技術，並不是用 Viterbi Decoding 中常見的 Forced Alignment，而是使用 beam search 中 pruning 的方式，將語音盡可能地依序切割出每一個音素。在這種情況下，評分語音切割後，如果原來的內容和標準語音相當類似，則經由切割後產生音素的數量將接近甚至等同於標準語音音素的數量。相反地，若亂講的評分語音中只有前 n 個音素和標準語音相同(後幾個音素完全不同)，則經由 pruning 後的音素也大約等於 n。舉例來說，如果標準語音為「she has your dark suit in greasy wash water all year」、評分語音為「she has your dark suit」，則經由語音辨識後，在評分語音中所能切割出來的音素數量是 15，如圖表 3。

對於沒有切割出來的音素，我們則將其可信度值設為 0，如此一來可以增加驗證系統的區別性，使得和標準語音內容完全不同的評分語音，其可信度值變得相當低。

圖表 3 為兩個語音經由語音訊號切割後產生的不同結果。上半部的語音內容等同於標準語音內容，因此

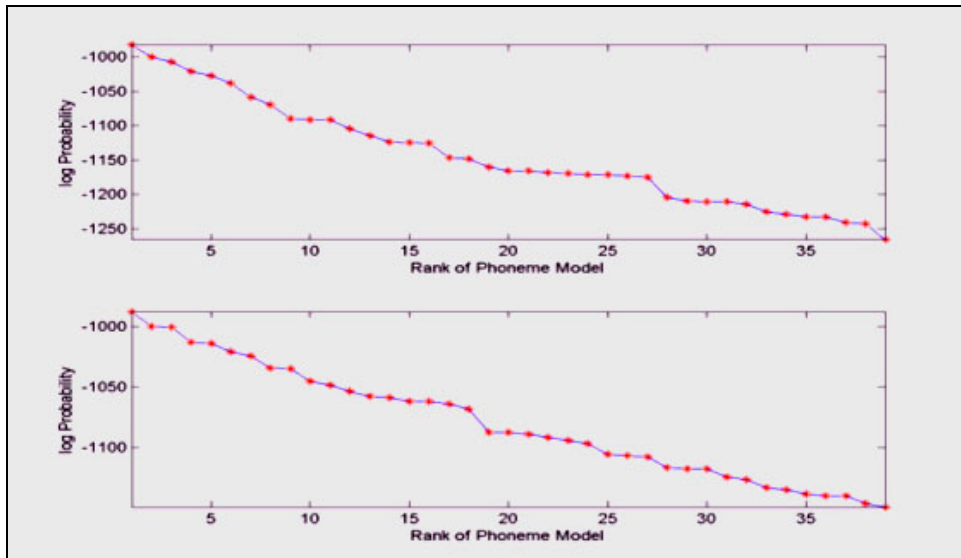
切割出來的音素很完整，而下方的語音內容和標準語音內容不盡相同，因而辨識程式將樹狀網路展開至節點 uw，就無法再繼續。圖表 3 之音素符號是採用 CMU Phone Set 表示法[21]。



圖表 3 說話驗證的語音訊號切割比較圖

3.1.2 音素排名

切割語音訊號得到音素時間區段後，首先以每個音素對 39 個 phone models 計算對數機率[21]，並以排名的順序得到相對應的可信度值。機率排名的示意圖如圖表 4：



圖表 4 音素機率排名

上下兩個機率分佈表示不同的音素經由辨識程式求得 39 個對數機率的結果，由圖表 4 可以看出，對於不同的音素，即使排名同樣是第二名，可是和第一名的對數機率差距卻不相同，會造成這樣的原因在於有些音素的發音相似，而有些音素的發音差異則相當大[16]，因此我們對於上方圖中的音素，可解釋成其第一名和第二名 phone model 的發音很接近，造成對數機率的差距相當小。而在下方圖中的音素，也許在我們 39 個 models 中，只有一個 model 的發音和該音素接近，因此更加突顯了其第一和第二名的對數機率差距。

3.1.3 驗證機制

經由語音訊號切割之後，產生的結果可能有兩種情況：一種是部份的語音訊號已經成功切割出時間區段的音

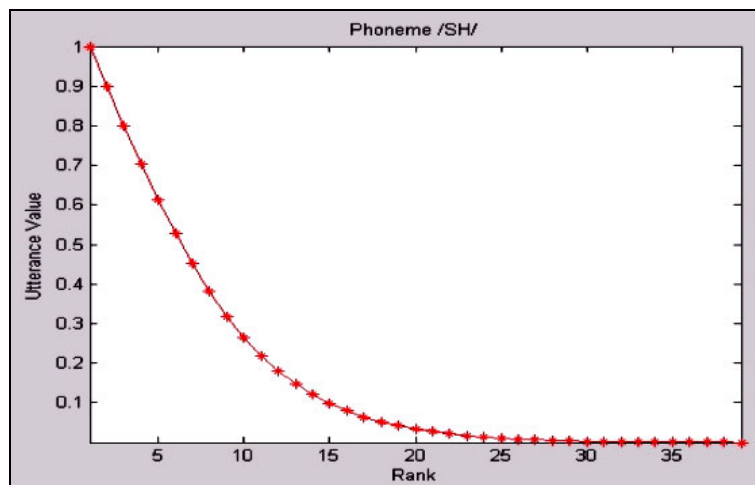
素，另一種則是語音訊號的後半部可能沒有辦法切割出音素。而在這一節討論的驗證機制，主要是針對前者的情況，也就是如何將音素的排名正規化，得到一個合理的數值。

在 Sukkar 和 Lee 於 1996 年發表的論文[17]中提到，音素的對數機率以及對所有音素的對數機率排名，和驗證系統的可信度值是有很大影響的。基於以上的前提，我們將 Sukkar 和 Lee 所提出求取可信度值的式子改寫並以下列公式表示：

$$value_{pho} = \frac{2}{1 + \exp\left(\alpha \cdot (Rank_{pho} - 1) \cdot \frac{\log P_{Rank_{pho}}}{\log P_{Rank_1}}\right)}$$

$\exp(x)$ 表示 e^x ，即自然對數的 e 的 x 次方。 $Rank_{pho}$ 和 $\log P_{Rank_{pho}}$ 分別表示該音素在 39 個 models 中的排名及對數機率值，1 表示第一名， α 為我們調整的參數值。由此公式可得知，當某音素相對於 39 個 models 的排名為第一名時，該音素的可信度值為 1。

圖表 5 表示對於「SH」這個音素之語音區段藉由上述的公式可將其對應於 39 個 models 所產生的對數機率及名次換算成可信度值。從圖中可以看出，當名次在第 10 名左右時，可信度值已經降至 0.2 了。



圖表 5 音素 SH 的排名與可信度值的關係

另外由於音素間發音的差異性，因此我們在評斷可信度值時，不能單純地以排名來做比較。舉例來說，音素「OW」〔o〕和「S」〔s〕比對完 39 個 models 後同樣都得到第二名的結果，但是對於「OW」而言，其第一名是「AO」〔ɔ〕，而「S」音素的第一名是「T」〔t〕，則我們可以很明顯地看出「OW」和第一名的對數機率差距較小，也因此可信度值應該要比較高才合理。因此在上述公式中，我們將排名的差異再乘上對數機率的比較差異，如此一來就會使得每個音素的可信度值受到排名及對數機率的影響。最後經由計算得到的可信度值介於 0 和 1 之間。

當計算出句子所有成功切割的音素可信度值之後，利用每個音素的時間長度占句子時間長度的百分比作為權重，即可推導得出一句語音訊號的可信度值。以下是設定的公式：

$$value_{sen} = 100 \cdot \sum_{n=1}^N \frac{len(pho_n)}{len(sentence)} \cdot value_{pho_n}, N \text{ 為一單字中評分音素的數量, } len(x) \text{ 表示 } x \text{ 的時間長度。}$$

至於有些單字可能其中的一些音素沒有辦法經由語音訊號切割產生，對於這些音素，我們就直接將其 $value_{pho}$ 設為 0。最後乘上常數 100 代表我們將說話驗證系統的結果定義在 0 至 100 之間。

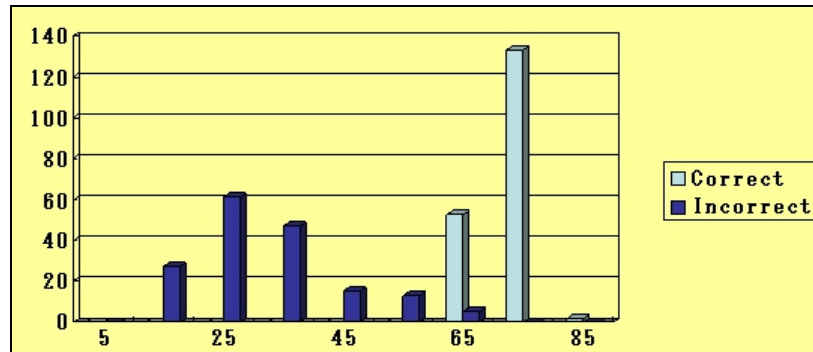
3.1.4 說話驗證實驗結果

對於在實驗中求出的門檻值(threshold)而言，如果語音訊號得到的可信度值高於門檻值，則我們稱「此句語音訊號的內容和標準語音訊號的內容相同」這句話是相當可靠的，也就表示我們可以放心地針對這句語音訊號進行評分。相反的，則表示這句話和標準語音的內容不相同，因此我們也就停止讓兩句不相同的語音進行後續的評分動作。對此我們蒐集兩部份的實驗語料：

1. Correct: 取 168 句說話內容相同的語音訊號當作標準語音內容，這部份語音檔案長度總和約為 9 分 10 秒。
2. Incorrect: 取 168 句內容不等於標準語音內容的語料，這部份語音檔案長度總和約為 7 分 31 秒。其中一部份內容和標準語音完全不相同。另一部份則是語音訊號內容「部份相同」於標準語音內容。在此我們定義一句

話中若存在連續 3 個單字以上和標準語音內容相同，但並不是完全相同，即為「部份相同」。

實驗用的語料其音訊格式皆為 PCM，音訊取樣頻率為 16 kHz，位元解析度為 16 bits，所有的實驗語料皆為單聲道。接著將上述兩部份各 168 句的實驗語料經由說話驗證系統得到對應的可信度值，而後再統計、分析這些可信度值即求得驗證系統的門檻值。圖表 6 為求取門檻值的實驗結果分佈圖，橫軸為可信度值的範圍，縱軸為可信度值處於該範圍內的語音訊號個數。



圖表 6 說話驗證求取門檻值實驗結果分佈情況

我們以「型別 I 錯誤率(Type I error, False Reject)加上型別 II 錯誤率(Type II error, False Accept)為最小」作為尋找門檻值的前提。根據實驗結果，我們發現 Correct 中的語料其最小可信度值為 63.21，而在 Incorrect 可信度值大於 60 的語料中最接近 63.21 的可信度值為 61.59，因此我們將說話驗證系統的門檻值設定成 62.40(即兩者的平均)，如此可達到型別 I 錯誤率為 0%，型別 II 錯誤率為 1.19%。

經由上述實驗計算求出門檻值後，我們另外準備一組內含 Correct 及 Incorrect 各為 168 句的測試語料，其中 Correct 語料的語料長度總和約為 7 分 27 秒，Incorrect 語料的語料長度總和約為 8 分 57 秒。將這些語料以門檻值為 62.40 的實驗結果，其型別 I 錯誤率為 7.14%，型別 II 錯誤率為 0.60%。

3.2 語音訊號切割

「語音訊號切割」模組的功能乃是將標準語料及評分語料切割出音素發音的區段。其作法是以預先訓練好的英文發音聲學模型，切割出語料中之正確的音素發音區段。以下章節將分成「聲學模型的訓練」和「利用語音辨識來進行語音訊號切割」這兩部份來介紹。

3.2.1 聲學模型 HMM 的語料

實作語音訊號切割之前，我們必須先產生聲學模型，才能針對各種不同的語音進行切割動作。本論文中我們設計了兩種不同的聲學模型：一個是臺灣人口音的聲學模型，一個是外國人標準語音的聲學模型。

首先針對母語為英文的聲學模型，我們使用 TIMIT 語料來加以訓練。語料內容為 2,342 句平衡語料，由 438 位男性、192 位女性，共 630 人錄製，每人分配錄製 10 句，故共有 6,300 句語音。依 TIMIT 的建議取其中 4,620 句、語料長度總和約為 3 小時 49 分 10 秒的語音訊號作為母語為英文的聲學模型訓練，另外 1,680 句、語料長度總和約為 1 小時 23 分 51 秒的語音，則作為外在測試檔(Outside Test)。

另一方面針對母語為國語的聲學模型，我們請 33 位學生，其中包含了 23 位男性、10 位女性，依 TIMIT 的資料錄製 7,026 句平衡語料，我們取其中的 4,684 句、語料長度總和約為 4 小時 11 分 3 秒的語音作為母語為中文的聲學模型訓練，而另外的 2,342 句、語料長度總和約為 1 小時 57 分 43 秒的語音作為外在測試檔。上述語料的音訊格式皆為 PCM，取樣頻率為 16 kHz，位元解析度為 16 bits。

3.2.2 聲學模型設計

英文中每一個音節可能由一個或數個音標所組成，而每一個音標都會對應到一個音素，而聲調、重音和破音(multiple pronunciation)的問題，在目前的聲學模型設計中則暫時忽略。TIMIT 的字典有 62 個音素，由於華人對於一些音素不像外國人念得那麼準確，再加上訓練語料不足下，如果我們減少訓練 model 的個數，則可使每個 model 的訓練語料取樣數目增多。鑑於上述兩個原因，我們將原先 TIMIT 設計的 62 個音素刪減成 40 個音素(含靜音 SIL 音素)。在本章中我們使用的聲學模型和音素是一對一對應的。舉例來說，“school” 這個單字，其 KK 音標為 [skul]，以我們設計的聲學模型來說，就是「S」+「K」+「UW」+「L」。表格 1 是我們所設計的 40 個聲學模型與 KK 音標對照表：

表格 1 40 個聲學模型與 KK 音標對照表

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

我們使用以下三種原則來對 TIMIT 的 62 個 model 做刪減的動作，各 model 後面刮弧內的英文字其底線部份即表示該 model 的發音。

- ◆ 替換：將發音相似的音素使用一個 model 代替。例如：AXR (butter [*ʌ*]) → ER (bird [*ɜ*]), NX (winner [*n*]) → N (noon [*n*])
- ◆ 分解：將一個 model 拆開成兩個以上的 model 來組成。例如：EN (button [*n*]) → AH + N ([*ʌ n*]), ENG (Washington [*ɪ ŋ*]) → IH + NG ([*ɪ ŋ*])
- ◆ 刪除：將許多設定細微的暫停音素刪除。例如：PAU、EPI

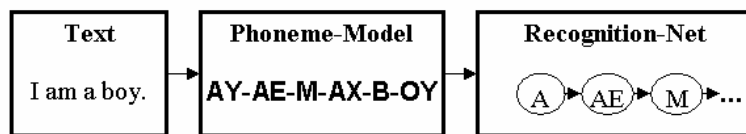
3.2.3 聲學模型訓練

在訓練語料的特徵擷取中，我們以 20ms 為單位取音框，每隔 10ms 重取音框。每一音框採用 39 維特徵向量當作聲學模型之特徵參數[7]，該特徵向量包含了 12 維的 MFCC 和一維能量參數，以及其一階微分、二階微分。

聲學模型訓練部份，我們採用隱藏式馬可夫模型(Hidden Markov Model, HMM)[6]作為聲學模型。在 monophone 的 model 下，每個 model 有 3 個 state，每個 state 則使用了 18 個 mixtures，以 HTK(Hidden Markov Model Toolkit)進行訓練。

3.2.4 訓練結果

語音訊號切割的主要目標即是希望能夠將連續的英文語音句子，其中包含了標準語音和評分的語音，切割成獨立的音素，如此一來我們才可以針對每一段句子中的音素和標準語音中的每一個音素做比較。在此我們使用強迫對應(Forced Alignment)[6]的方式將語音訊號切割成各個音素的時間區段，以利評分機制的運作。在前處理的過程中，我們利用內含 127,102 個英文單字的 CMU 字典(Dictionary from Carnegie Mellon University)對各單字標音並建立各自獨立的辨識網路[21]。如下圖：



圖表 7 語音訊號切割前處理流程示意圖

完成前處理動作後，我們可繼續進行語音訊號切割的流程，首先將一語音訊號經過端點偵測後再經由特徵擷取，取出語音中的特徵，然後將這些特徵參數透過聲學模型(隱藏式馬可夫模型)及語言模型(辨識網路)，利用維特比演算法(Viterbi algorithm)即可找出最相似的音素，並得知各音素的時間區段。

關於實驗測試語料的部份，我們使用了 1,680 句母語為英文的語音檔案，其語料的長度總和約為 1 小時 23 分 51 秒，以下我們簡稱為 N-Wave (Waves from Native-Speaker)。另外使用了 2,342 句母語為國語的語音檔案，語料的長度總和約為 1 小時 57 分 43 秒，以下簡稱為 T-Wave(Waves from Taiwanese)，來做 Outside Test。實驗用的語料其音訊格式皆為 PCM，音訊取樣頻率為 16 kHz，位元解析度為 16 bits。

在聲學模型這個部份，我們訓練出了兩個聲學模型：一個是由以英文作為母語的使用者所錄製的訓練語料產生的聲學模型，以下我們簡稱為 N-HMM(HMM trained from Native-Speaker)，另一個則是由臺灣人所錄製的訓練語料所產生的，以下我們簡稱為 T-HMM(HMM trained from Taiwanese)。

關於實驗的方式，我們分別對每一句語音訊號和已知的語音內容文字作 Forced Alignment，再由產生的結果對每個單字及音素判斷其時間區段的切割是否正確。

為了比較兩個聲學模型所產生的影響，我們對語料(N-Wave, T-Wave) 和聲學模型(N-HMM, T-HMM)作交叉實驗。表格 2 列出音素切割正確率的實驗結果：

表格 2 語音訊號切割實驗結果

項目 \ 實驗方式	N-Wave /N-HMM	N-Wave /T-HMM	T-Wave /N-HMM	T-Wave /T-HMM
實驗語料音素總數	58,282	58,282	81,229	81,229
切割後正確音素總數	58,253	57,142	77,293	80,230
音素時間正確率	99.95%	98.04%	95.15%	98.77%

在判斷音素時間正確率的部份，對於 N-Wave 而言，由於所有的語料 TIMIT 都有提供標音檔，因此我們可比對切割出來的時間點和標音檔，若相差在 0.1 秒以內(5 個音框)，則我們稱此音素的時間為正確。而對於 T-Wave 而言，由於並沒有經過人工標音，因此我們只在龐大的語料中取樣 10% 進行人工判斷，只要該區段人耳聽起來相差不大，則我們稱該音素的時間為正確。

由表格 2 的實驗結果可知，在不同的聲學模型下，Forced Alignment 的音素時間區段都非常準確。表格 3 則是 N-Wave、T-Wave 透過大詞彙辨識的方式，經由 N-HMM、T-HMM 所得出的辨識率，其中詞彙內容為 2,342 句英文句子。

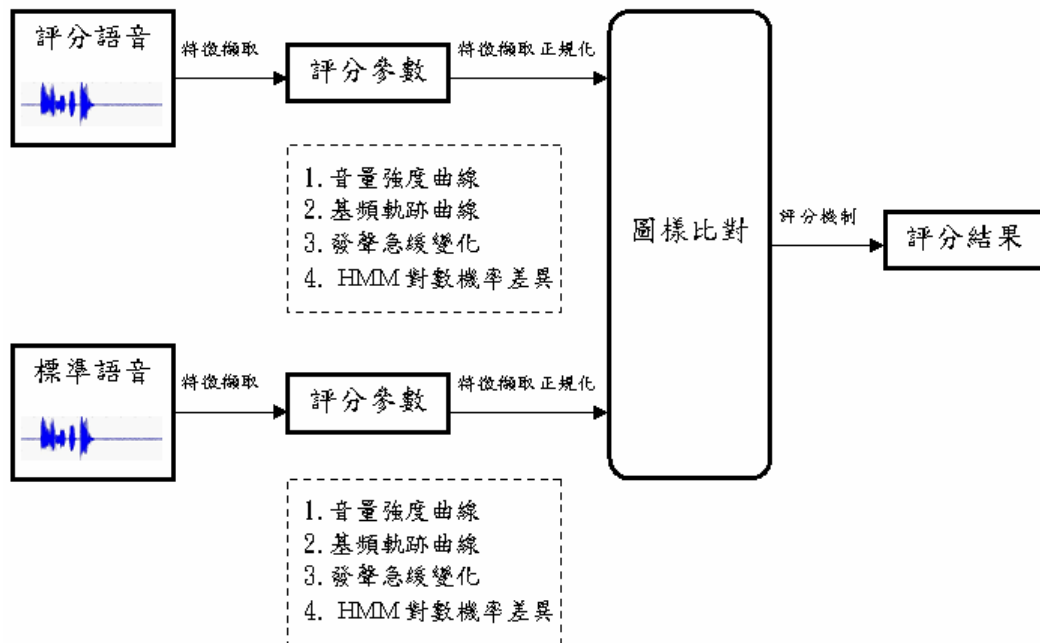
表格 3 英文語音辨識率

項目 \ 實驗方式	N-Wave /N-HMM	N-Wave /T-HMM	T-Wave /N-HMM	T-Wave /T-HMM
實驗語料句子總數	1,680	1,680	2,342	2,342
辨識正確句子總數	1,650	622	1,997	1,425
句子辨識率	98.21%	37.02%	85.26%	60.85%

由表中的結果我們可以發現，對於相同語料，N-HMM 的辨識率皆高於 T-HMM，這就表示當我們以 N-HMM 為聲學模型來對語音訊號求取對數機率時，所得到的對數機率值其可信度會高於 T-HMM。根據此實驗結果，在接下來的章節中，我們將會以 N-HMM 當作我們評分比對的聲學模型。

3.3 英文語音評分

圖表 8 為評分系統流程圖，我們將就評分參數擷取、圖樣比對方式和評分機制建立分別作介紹。



圖表 8 評分系統流程圖

3.3.1 評分參數擷取

除了音量強度曲線、基頻軌跡曲線為評分參數外[15]，我們也採用了 HMM 對數機率差異和發聲急緩變化這兩項評分參數。在 Forced Alignment 的同時，我們可以得到每個音素對應於聲學模型的對數機率(HMM log-Probability)[10][11]和各音素的時間區段，這就是所謂的 HMM 對數機率差異和發聲急緩變化這兩項評分參數。

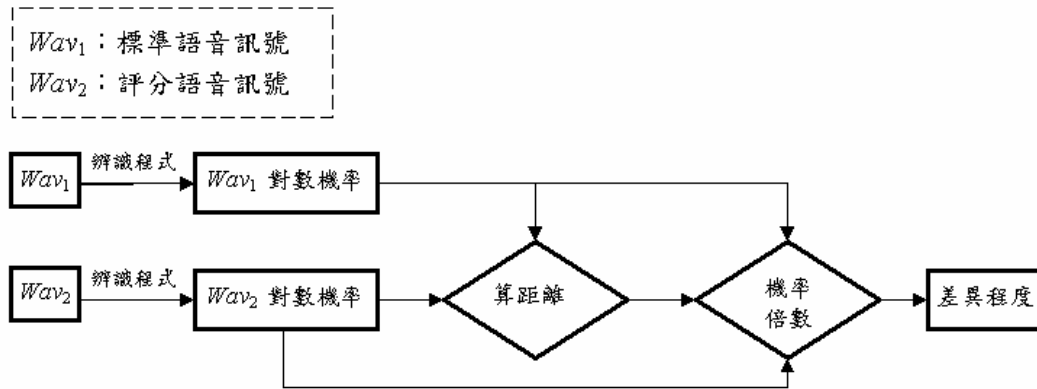
3.3.2 圖樣比對方法

在前三個評分參數中，我們使用不同的正規化方法如內插法、線性平移和線性縮放[15]，如表格 4。而 HMM 對數機率差異則採用較為不同的比對方法，在以下說明。

表格 4 各評分參數採用的正規化及距離算法

評分參數	正規化方法	距離算法
音量強度曲線	內插法、線性縮放	Euclidean Distance
基頻軌跡曲線	內插法、線性平移	Euclidean Distance
發聲急緩變化	無	Euclidean Distance

為了計算 HMM 對數機率的差異，我們先以 N-HMM(HMM trained from Native Speaker)求出標準語音訊號及評分語音訊號中每個音素的對數機率，若對數機率值愈大，表示該音素的發音愈接近聲學模型。圖表 9 為 HMM 對數機率差異比對的流程圖：



圖表 9 HMM 對數機率差異比對流程圖

由於機率值是絕對的，不容易從數值直接作比較，因此我們設計了機率倍數來修正對數機率的差異值，當兩語音的對數機率絕對值皆小於 1050 時，機率倍數的變化趨勢較小。當兩語音的對數機率絕對值皆大於 1050 時，機率倍數的變化趨勢較大。關於機率倍數我們定義以下的公式：

$$Const = \begin{cases} \left\lceil \frac{|\log - probability|}{350} \right\rceil, & 0 \leq abs_{\log} \leq 1050 \\ 3 + \min \left(1, \left\lceil \frac{|\log - probability|}{1400} \right\rceil \right), & abs_{\log} > 1050 \end{cases}$$

$$factor_p = (Const_{stand})^2 + (Const_{Evaul})^2$$

當算出標準語音和評分語音的 Const 值後，再經由平方相加即可得到機率倍數 $factor_p$ ，將此機率倍數乘上兩語音訊號對數機率的差距就是我們發音特徵的差異程度。

3.3.3 評分機制建立

在音素層次，我們由四種評分參數得到不同的分數，再往上由單字(word)和句子(sentence)層次作評分，就可以得到最後評分的結果，以下則分四個層次作介紹。

評分參數層次：對於每個音素中評分參數的分數，我們設定以下的公式[15]：

$$score_{fea} = \frac{100}{1 + a \cdot (dist)^b}$$

由這個公式我們就可以將兩音素間某個特徵的差異程度轉成 0 到 100 之間的分數，只要設定好兩組的 $dist$ 及對應的 $score_{fea}$ ，即可從中求出 a 和 b ，接著所有的距離也將可以計算出對應的分數。

音素層次：當計算出每個音素中四項評分參數的分數後，利用四項特徵對於英文語音評分系統所占的權重加總後即可得到每個音素的分數。以下是設定的公式：

$$score_{pho} = w_1 \cdot score_{fea_1} + w_2 \cdot score_{fea_2} + w_3 \cdot score_{fea_3} + w_4 \cdot score_{fea_4},$$

w_1, w_2, w_3, w_4 分別代表四個評分參數的權重。經由下一節的實驗，我們可以求出這四項權重，也可以由權重的比例得知四項評分參數對於英文評分的重要性。

單字層次：得知每個音素的得分後，以每個音素占單字的時間為權重，即可求出句子中每一個單字的分數，以下為設定的公式：

$$score_{word} = \sum_{n=1}^N \frac{\text{len}(pho_n)}{\text{len}(word)} \cdot score_{pho_n}, \text{ 其中 } N \text{ 為一單字中評分音素的數量, } \text{len}(x) \text{ 表示 } x \text{ 的時間長度。}$$

句子層次：由於單字的時間長短會影響人耳對於一句話的關注點，因此我們也是以單字的時間為權重來計算出一句語音訊號最後得到的分數。以下為定義的公式：

$$score_{sen} = \sum_{n=1}^N \frac{\text{len}(word_n)}{\text{len}(sentence)} \cdot score_{word_n}, \text{ 其中 } N \text{ 表示句子中單字的總數, } \text{len}(x) \text{ 表示 } x \text{ 的時間長度。}$$

4 實驗結果

得到四個評分參數中各音素的差異程度後，我們依所佔的比例求出一個句子的平均差異程度，即可代入以下的公式：

$$score = w_1 \cdot \frac{100}{1 + a_1 \cdot (dist_1)^{b_1}} + w_2 \cdot \frac{100}{1 + a_2 \cdot (dist_2)^{b_2}} + w_3 \cdot \frac{100}{1 + a_3 \cdot (dist_3)^{b_3}} + w_4 \cdot \frac{100}{1 + a_4 \cdot (dist_4)^{b_4}}$$

其中 $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 為差異程度轉成分數的參數， w_1, w_2, w_3, w_4 為四個評分參數的權重，而 $dist_1, dist_2, dist_3, dist_4$ 表示標準語音和評分語音訊號在比對後其四項評分參數的距離，再經由以下的實驗，即可求得各參數值。

在語料訓練部份我們收集 200 組語料，每一組的語料分別包括一句標準語音和一句評分語音，每句語音長度為 5 秒、音訊格式為 PCM、音訊取樣頻率為 16 kHz、位元解析度為 16 bits。其中標準語音的語料長度總和約為 12 分 51 秒，評分語音的語料長度總和約為 18 分 39 秒。接著請外語所老師協助我們對每一句評分語音作主觀的評分，之後再統計實驗中每一句語音人為評分的平均分數。同樣的，按照訓練語句的作法，我們也收集了 200 組語句作為測試用。

將這 200 組訓練語料透過評分系統評分，則每組評分語音都會得到四個特徵對應的差異程度 $dist_1, dist_2, dist_3, dist_4$ 。收集了這些差異程度和對應的分數後，使用 Simplex Downhill Search，就可以找出 $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 和四個評分參數的權重 w_1, w_2, w_3, w_4 。

經由上述的實驗，我們得到音量強度曲線的權重為 7.45%，基頻軌跡曲線的權重為 22.40%，發聲急緩變化的權重為 17.24%，HMM 對數機率差異的權重為 52.91%。

接著我們將 200 句測試語句的人工評分結果分成三個等級：Bad(0~59)、Average(60~79)、Good(80~100)，另外也將 200 句測試語句的系統評分結果依此分成三個等級。最後再統計每個句子的人工評分和系統評分後，就可以得到表格 5 的結果：

表格 5 人工評分和系統語音評分的關係對照表

人工評分 \ 系統評分	Bad	Average	Good
Bad	28	17	7
Average	20	27	20
Good	10	11	63

其中橫軸表示人工評分的等級項目，縱軸表示系統評分的等級項目，表格中的數字則表示相對的語句數目。從表中我們可以明顯地看出來，對角線的數目都比同一列、同一欄的數目高，這就表示在經由 Simplex Downhill Search 調整各參數之後，我們的評分系統和人工評分已有一定的正相關性，約 $(28+27+63) / 200 = 59\%$ 。

5 結論

「說話驗證」對評分語音進行初步的評估，若可信度夠高，接下來的評分才具有可信度。「語音訊號切割」則是以 Forced Alignment 得到每個音素的時間區段。經由實驗結果我們可以知道，使用辨識率較高的聲學模型，其 Forced Alignment 的音素切割時間將更為準確。「英文語音評分」包括評分參數的擷取、圖樣比對方法的設計和評分機制的建立等三個部份。藉由實驗我們可以知道，「HMM 對數機率差異」在英文語音評分中所代表的重要性最高，而「音量強度曲線」則是最低。

語音評分的運用相當廣泛且實用，配合未來技術的成熟，不只可作為英語學習的工具，之後的台語、客語評分學習也將是台灣地區重要的研究之一。

參考資料

- [1] 鐘林，“漢語語音辨別說話驗證”，北京清華大學碩士論文，民國 91 年
- [2] 楊永泰，“隱藏式馬可夫模型應用於中文語音辨識之研究”，中原大學碩士論文，民國 89 年
- [3] 陳柏琳，“中文語音資訊檢索—以音節為基礎之索引特徵、統計式檢索模型及進一步技術”，台灣大學博士論文，民國 90 年
- [4] 呂道誠，“不特定語者、國台雙語大詞彙語音辨識之聲學模型研究”，長庚大學碩士論文，民國 90 年
- [5] G.S. Ying, L.H. Jamieson and C.D. Michell, A probabilistic approach to AMDF pitch detection, Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Volume: 2, 1996, Page(s): 1201-1204 vol.2
- [6] Steve Young, The HTK Book version 3, Microsoft Corporation, 2000
- [7] Lawrence Rabiner, B.H Juang, Fundamentals of speech recognition, Prentice Hall, 1993
- [8] J.D., J.G., J.H. and L.H., Discrete-Time Processing of Speech Signals, Prentice Hall, 1993
- [9] Giuliano Monti, Mark Sandler, Mnophonic transcription with autocorrelation, Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, December 7-9, 2000
- [10] L. Neumeyer, H. Franco, V. Digalakis and M. Weintraub, Automatic scoring of pronunciation quality, 1999
- [11] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, Automatic pronunciation scoring for language instruction, Proc. Int. Congress on Acoustics, Speech and Signal Processing(ICASSP), 1997
- [12] J.-S. Roger. Jang, C.-T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing, Prentice Hall, 1996
- [13] 高名揚，“以聲音內容為主的音樂資料庫檢索系統的加速方法”，清華大學碩士論文，民國 90 年
- [14] J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley Publishing Company, 1974
- [15] 李俊毅，“語音評分”，清華大學碩士論文，民國 91 年
- [16] Gies Bouwman and Lou Boves, Utterance Verification based on the Likelihood Distance to Alternative Paths, Department of Speech, University of Nijmegen, The Netherlands, 2002
- [17] Rafid A. Sukkar and Chin-Hui Lee, Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition, IEEE Transactions on Speech and Audio Processing, VOL. 4, No. 6, November 1996
- [18] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price, Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech, 1996
- [19] C. Cucchiari, H. Strik and L. Boves, Automatic Evaluation of Dutch Pronunciation by Using Speech Recognition Technology, Department of Speech, University of Nijmegen, The Netherlands, 1997
- [20] Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li and Ming-Chun Wu, “Automatic Pronunciation Assessment for Mandarin Chinese”, Proc. Int. Conf. on Multimedia And Expo (ICME), 2004
- [21] http://www.speech.cs.cmu.edu/sphinx/doc/phoneset_s2.html