台灣閩南語聲調評分系統評估與研究

蔡岳廷[1]，廖嘉新[1]，呂道誠[2]，呂仁園[2]

1. 資訊工業策進會電子商務研究所
2. 長庚大學資訊工程研究所

E-mail: atsai@iii.org.tw　　TEL:886-2-87326222 ext 298

摘要

聲調的學習與掌握，對學習任一種漢語而言，都是極其重要的。由於西方人的語言大多沒有聲調辨義的功能，故聲調的掌握對西方人學習漢語的經驗，常是最難的部分。本論文提出一個聲調辨認及評分之系統，有助於非華語人士對漢語的學習。本系統可對已知文字或注音的中文語音訊號正確切出音節邊界，並針對聲調的變化擷取出相關的特徵，以作爲機器評分的依據。在實驗方面，我們實際的錄製學習者語音來進行人工評分與機器評分的比對，也說明了本實驗是如何進行語音的錄製及人工的評分。最後，針對實驗結果提出探討，比較兩者評分之差異，提出改善的方式，以提高系統評分之可靠度，使其評分機制能與人類的聽覺相近。

一、前言

漢語是由數個互不相通的語言所組成的一個語族，包含華語（在台灣以前稱作國語、在中國向來稱爲普通話）、閩南語、客語、粵語、吳語、湘語、贛語…等。在這個語族內，最重要的特徵之一就是音節結構明顯，每種語言都有數百到數千不等的相異單音節；另一個重要特徵是聲調，每個單音節都用聲調來區分彼此的意義，說話人若把聲調混淆了，則他人聽起來不但覺得不順暢，甚至有時候意思也會誤解了。故聲調的學習與掌握，對學習任一種漢語而言，都是極其重要的。由於西方人的語言大多沒有聲調辨義的功能，故聲調的掌握對西方人學習漢語的經驗，常是最難的部分。在台灣，最普遍的漢語包含台灣華語、台灣閩南語以及台灣客語，三種漢語各有不同的音節結構及聲調結構，其中若以相異音節總數或相異聲調類別總數來計，則台灣華語以約 400 相異音節數及 5 種聲調類別最爲簡單，而台灣閩南語以約 900 相異音節數及 9 種聲調類別（含變調衍生類別）最爲複雜，至於台灣客語，由於在台灣有兩大腔調（海陸及四縣）之變異，其複雜度恰介於兩者之間。經過比對參照三語聲調之聲學特徵，一般認爲用 9 種聲調類型將可以涵蓋全部的聲調類型。所以，本論文希望針對閩南語聲調設計一個教學評分系統，未來亦可以應用在華語及客語的教學上面。

語言的學習大致上可以分爲聽、說、讀、寫、這四大部分，其中利用影音媒體，以及書籍的方式可以幫助我們練習聽、讀和寫，例如看外國影集來訓練聽力，讀報章新聞來增進閱讀以及寫作敘事的能力；但是對於『說』，卻沒有一個很好的解決方案來幫助我們自我練習與評量。然而，利用本論文的教學評分系統，學習者將可以了解自己哪一個字、哪一個音節的發聲方式或是

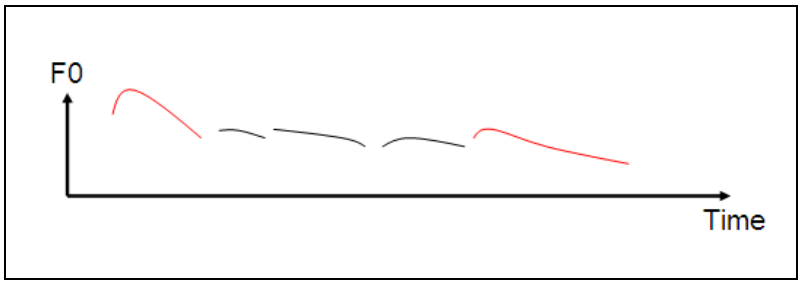
語調不正確，以及如何改進這些缺失。學習者自行反覆的練習，不需因爲發音不正確而羞於請教，利用系統中的語音評量機制，可以量化使用者發音的正確性，幫助使用者自我評量。本論文設計並錄製閩南語教學語句 200 句，作爲系統發展測試的語料。

以下爲本論文的章節安排，第二章介紹目前國內外相關的研究發表與產品系統，第三章爲說明中文正音評分系統的技術及架構，第四章則進行實驗分析及探討，最後第五章提出結論。

二、相關研究

目前國內外已經有許多類似的研究發表及產品系統。在國內的研究發表上，有清華大學的張智星教授[8]，主要是根據聲音音量強弱、聲音音高的起伏及聲紋來作語音的評分。另外，在音高擷取上的技術也是值得參考[10]。

在國外的研究中，H. Fujisaki[9]的音調研究理論是被廣泛的利用與討論，Fujisaki Model 簡單的來說是一種模擬語音聲調走勢的模型。比如在朗讀比賽，爲了達到說話的韻律美感，通常會強調一個句子裡的某些重要字來加強語氣，使得聽者能夠清楚瞭解到演講者的重點。因此，Fujisaki Model 就是爲了模擬這個整體聲調韻律而建立的一種模型之一。此模型通常是以句子爲單位來模擬韻律走勢。而另一個例子就是英語的疑問句和肯定句，通常在疑問句中，結尾音調會上揚，而肯定句中句末的音調會下降。用一個簡單的示意圖來說明用 Fujisaki Model 所建立的韻律走勢來模擬平常人說話的韻律習慣，如圖(一)所示。



圖(一) 以句子爲單位的音調走勢圖

橫軸代表著一句話的時間，而縱軸用頻率來表示韻律單位。而其中的線條就是一般人類說話的韻律習慣。通常在一開始的時候，音調會比較高，慢慢地中間的語調會漸漸下降或趨於平坦，而到了一句話快要結束的時候音調又會稍微的高起來，然後又降下去，而一句話的末端通常是頻率最低的地方。另外一方面， Fujisaki Model 也將詞句的整體節奏也考慮在內，如圖中，一句話或一個詞，最前面和最後面字的時間會拉的比中間部分字的時間來的長。有了類似 Fujisaki Model 的模型之後，就能掌握整體語音聲調走勢和協調性。

另外，目前也有相關的產品系統，相關的說明如表(一)所示。

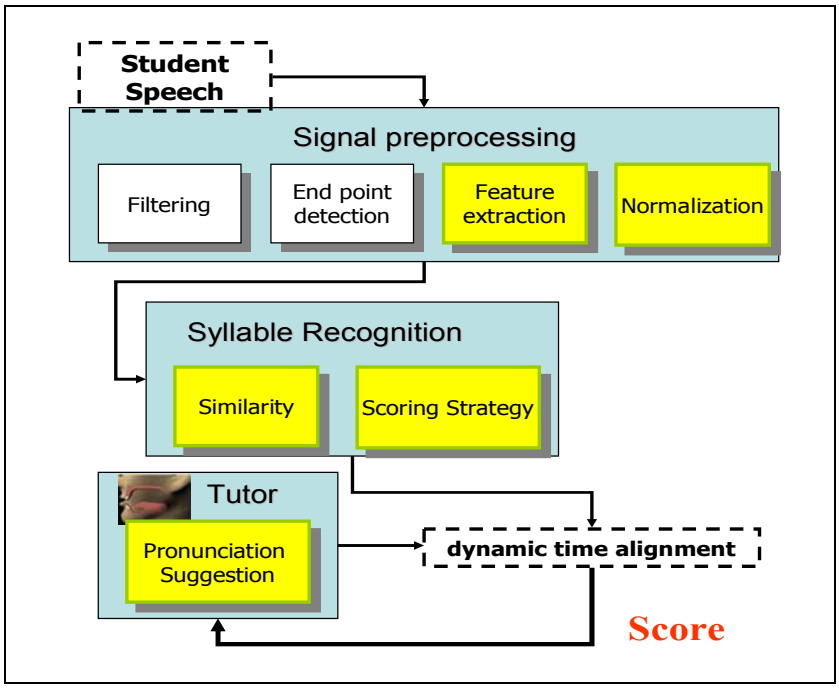
表(一) 發音教學相關產品及系統之說明

系統名稱	發表單位	功能	主要技術	資料庫	限制
MyET[11]	艾爾科技	使用者根據課程安排，練習音文發音，系統根據練習結果，給予使用者評分。	Automatic Speech Analysis System	共有 8 種課程包括:生活英語，辦公室英語…等	英文， 使用者只能根據課程安排來練習，無法即查即學即用。
American English Pronunciation Practice[12]	manythings.org	英文發音課程，提供聽力測驗。	利用 Flash 製作教材，免費提供學習者，線上學習。	24 種教學課程	英文， 只有單字，沒有提供使用者語音評量。
The CMU Pronouncing Dictionary[13]	Carnegie Mellon University	提供 machine-readable pronunciation dictionary for North American English.	隱藏式馬可夫模型(Hidden Markov Model, HMM)	that contains over 125,000 words and their transcriptions	英文。
Pronunciation power[14]	English Computerized Learning	100 小時的課程訓與聽力測驗。	Native Language Model	1040 個不同的句子	英文， 課程無法擴充，無法即查即學即用。

三、中文正音技術評分系統

3.1 系統技術架構

圖(二)爲本系統技術流程圖，系統的目的爲讓學習者可以進行發音，並且系統給予評分及矯正。以下針對系統的主要方法進行說明。



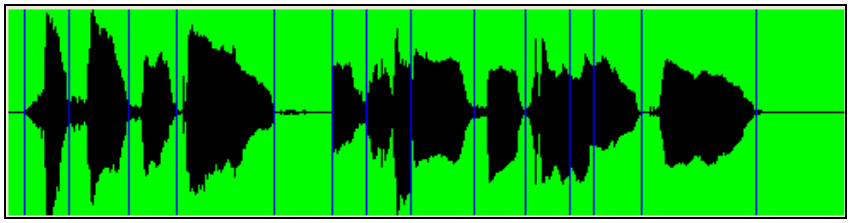
圖(二) 評分系統技術流程圖

3.2 聲調特徵擷取(Feature Extraction)

本論文採取四個特徵值作爲聲調評分的依據，其特徵值如下所示。

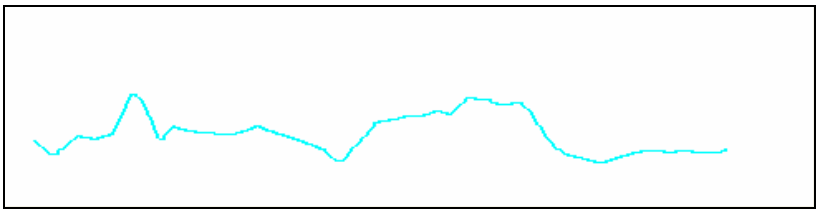
a. pitch curve
b. Spline curve
c. Fuzisaki phrase command curve
d. Fuzisaki accent command curve

Pitch curve 是求取音節部份在短時間 pitch 的變化，如圖(三)是一個聲音波形，圖(四)爲利用 Autocorrelation[4]的方法求出圖(三)的 pitch value。圖(五)是利用 SPline[5]的內插方法，將沒有 pitch value 的部分連接起來，求取比較中長時間的 pitch 變化，所求得的是一個 pitch contour。圖(六)的 long term intonation 是代表一個長時間的 pitch 變化，是將 pitch contour 經過 1.5Hz 的低通濾波器求得，也是仿照[6][9]中的方法求出 Fujisaki model 中的 phrase command value。另外，最後一個特徵就是將 Spline curve 及 Fuzisaki phrase command curve 相減求得 Fuzisaki accent command curve。

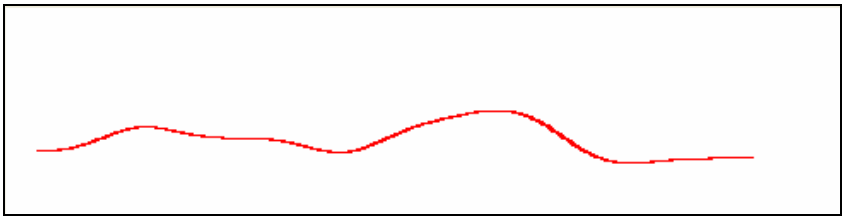


圖(三) 聲音波形

圖(四) pitch value



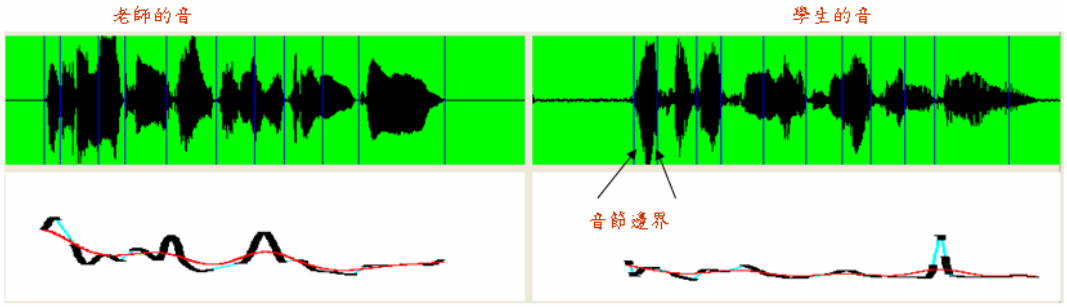
圖(五) pitch contour



圖(六) long term intonation

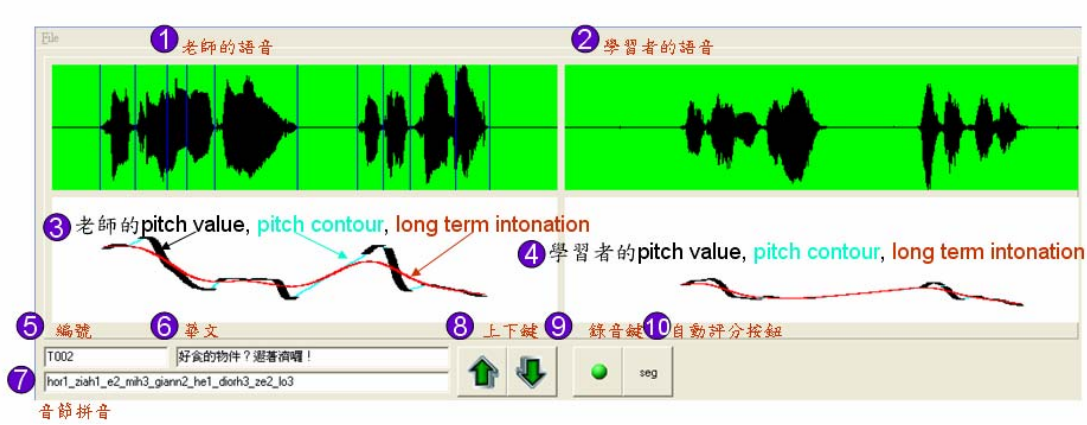
3.3 相似度比對演算法(Similarity)

隱藏式馬可夫模型可用來切割學習者語音音節的邊界，並且計算每個音節的分數，如圖(七)所示。



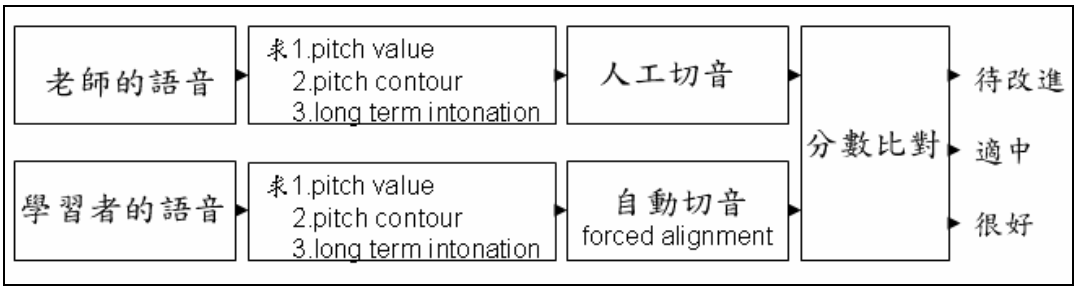
圖(七) 利用隱藏式馬可夫模型切出學習者的音節邊界

再將每個音節的分數做相加，就可得到一個總分，這個分數就作爲學習者的分數。整個評分系統的界面如圖(八)所示。



圖(八) 評分系統界面

爲了求得精確的評分，在標準語音的部份(老師的語音)，是先用人工切音完成。而學習者的語音是學習者輸入至系統後，透過 HMM 的方式來自動切割。切割完之後，會先求得特徵值的曲線，如章節 3.2 所描述。再來利用動態時間扭曲(Dynamic Time Warping, DTW)的方式計算每個對應音節之距離來進行評分，在章節 3.4 會有詳細的描述。整個評分的流程如圖(九)所示。



圖(九) 評分流程圖

目前已有幾種適用於連續語音比對的方法被提出，而常被使用的爲隱藏式馬可夫模型(HMM)和 Viterbi Algorithm。隱藏式馬可夫模型使用一個隨機的機率模型來描述聲紋的變化情形。由於短時間內，聲音隨著時間的改變量不大，隱藏式馬可夫模型使用連續的狀態改變，描述聲音的特徵值在短時間內的變化情形，其辨識效能優於動態時軸校正。

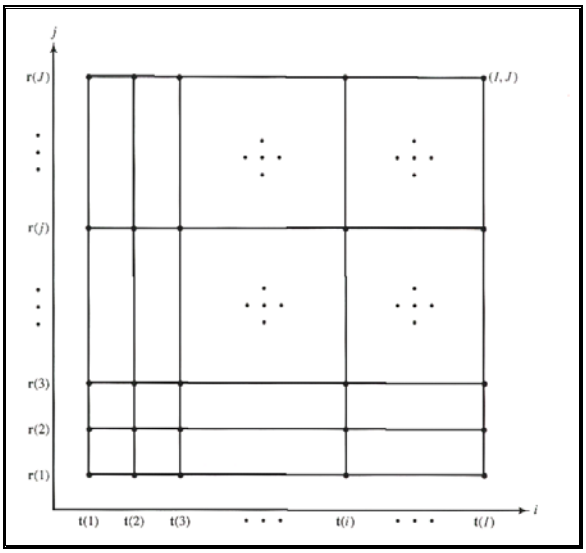
3.4 評分方法

系統的評分主要是計算標準語音系學習者語音在特徵值部分的差異，特徵值的部份如之前所描述的。將學習者的第 i 句語音信號透過自動相關函數取出基頻(F0 或 Pitch)的信號，稱爲 $StuPitch_i$，如圖(四)所示。再利用 Spline 的內插法將沒有基頻信號的部分(如無聲子音)填滿，此部份稱爲 $StuSpline_i$，如圖(五)所示。之後依據 Fujizski 模型的方式，取出 phrase command 和 accent command 所求出的曲線，稱爲 $StuFujiPC_i$ 及 $StuFujiAC_i$，如圖(六)所示。同樣的，也從標準語音取出相同的特徵值 $TeaPitch_i$、$TeaSpline_i$、$TeaFujziPC_i$、$TeaFujiAC_i$。

爲了方便機器自動評分，本論文將採 DTW 的方式將學習者的語音和老師的語音求出一個距離，包括 DTW[(pitch curve), (Spline curve), (Fuzisaki accent command curve), (Fuzisaki phrase

command curve)]。由於在作 DTW 計算之前，有先將音節作自動或人工的切割，所以在 Spline 及 long-term 的比對計算上，不會把靜音或無聲的部份考慮進來，只會計算音節部份的距離。然而，每個音節的距離可視爲一個 local-distance，最後再將全部相加求得 global-distance，以下說明 DTW 之計算方法。

DTW 的方法在語音訊號處理中是一種很常用來做相似度比對的方法，其主要的精神在於提供一個具有更大彈性的相似度比對法，使測試資料能透過伸展或壓縮，找到與參考資料間最小誤差的非線性對應。舉一例子，假設我們的測試資料爲 *t*，長度爲 *I*，參考資料爲 *r*，長度爲 *J*，圖(十)是常見的動態時間扭曲比對示意圖：



圖(十) 動態時間扭曲比對示意圖

DTW 的主要目的便是在 *t*、*r* 構成的平面上找出一條最佳的對應路徑 $path(i_k, j_k)$，即是使得測試資料與參考資料間的距離 *D* 爲最小，並且使得 $t(i_k)$ 對應到 $r(j_k)$，其中，k = 1, 2, …, K，i_k 與 j_k 都必須遞增，以數學式子表示如下：

$$D = \sum_{k=1}^{K} d(i_k, j_k)$$

$$d(i_k, j_k) = dist(t(i_k), r(j_k))$$

其中 $d(i_k, j_k)$ 可以爲任意一種距離測量方式，最常見的就是歐幾里得距離，在此我們是計算測試語音及標準語音之兩組梅爾倒頻譜參數以音框爲單位的歐幾里得距離。

最後所求出來的值爲

$DisPitch_i = DTW(TeaPitch_i，StuPitch_i)$

$DisSpline_i = DTW(TeaSpline_i，StuSpline_i)$

$DisFujiPC_i = DTW(TeaFujiPC_i，StuFujiPC_i)$

$DisFujiAC_i = DTW(TeaFujiAC_i，StuFujiAC_i)$

因此，針對每一句學習者的語音，我們就會有四組特徵值 $DisPitch_i$、$DisSpline_i$、$DisFujiPC_i$ 及 $DisFujiAC_i$。而在建立模型方面，本論文採用了 k-means 的演算法來分類，利用疊代 20 次，來將這些特徵值分成三類(待改進、普通、很好)。

四、實驗分析

4.1 語料

本實驗是針對不會說閩南語的國外人士(Non-speakers)錄製語音來評估。錄製的語料來源爲二百句的標準閩南語句。表(二)爲標準語句其中的十句，包含編號、語句及拼音。

表(二) 標準閩南語語句十句。

編號	語句	拼音
S001	學校附近有啥物好食的物件？	hak3_hau2_hu3_gin2_u3_siann1_mih1_hor1_ziah1_e2/e3_mih3_giann2
S002	好食的物件？遐著濟囉！	hor1_ziah1_e2_mih3_giann2_he1_diorh3_ze2_lo3
S003	我上愛去夜市仔食物件	ghua1_siong3_ai4_ki4_ia3_ci2_a4_ziah3_mih3_giann2
S004	暗頓攢好囉！來食飯。	am4_dng3_cuan2_hor4_lo3_lai2_ziah3_bng2
S005	趁燒，緊食！	tan4_sior1_gin1_ziah1
S006	暗頓誠腥臊	am4_dng3_ziann2_cenn2/cīnn2_cau1
S007	這項菜真好食	zit1_hang3_cai3_zin2_hor1_ziah1
S008	甘欲食一寡果子？好矣！我欲食芎蕉	gam1_bheh1_ziah3_zit3_gua1_gue1_zi4_hor4_a1_ghua1_bheh1_ziah3_ging2/gin2_zior1
S009	我會使閣[口林]一杯茶無?	ghua1_e3_sai1_gorh1_lim2_zit3_bue2_de5_bhor0
S010	汝欲食一寡物件無？	li1_bheh1_ziah3_zit3_gua1_mih3_giann2_bhor0

4.2 實驗方法

本實驗從 5 位 Non-speakers 學習者進行語句的錄製，這些習學者的國籍包含了日本、韓國、印度及法國四個國家。我們挑選了部份的長句與短句來作錄製，總共錄製了 350 句，其中包括了 250 個短句與 100 個長句。短句平均音節爲 7 個以下，例如表(二)中編號 S007 這個語句。長句平均爲 9 個音節以上，如表(二)中編號 S008 的語句。由於國外人士對於閩南語的發音較爲陌生，爲了能夠成功的錄製，錄製的方式除了提供拼音給學習者參考外，也讓學習者試聽標準語句，讓他們熟悉發音的部份，然後再依照他們所聽到發音重覆念過一次，並錄製起來。在短句的部份是讓學習者試聽 1 至 2 遍，長句的部份爲 3-4 遍。未來此系統也是會朝向讓使用者透過不斷反複的聽學及矯正來學習，並評估是否改善。

錄製完的語句將交由三位精通閩南語的人士進行評分。評分標準分爲"待改進"、"普通"及"很好"三種等級。評分項目分爲聲調及發音兩部份，若評分結果爲「待改進」，則圈出句子上需

改進的部份，若爲「適中」，則可自由選擇是否需要圈取。聲調部份是以詞爲圈取單位，發音部份則以字爲單位。其評分的格式，如表(三)所示。

表(三) 人工評分格式。

評分語句	T003 我上愛●去夜市仔●食物件		
聲調	☑待改進 □適中 □很好	改正部份	我上愛去夜市仔食物件
發音	□待改進 □適中 ☑很好	改正部份	我上愛去夜市仔食物件

4.3 結果與討論

本實驗所探討的爲聲調的評分結果，本實驗總共錄製了 350 句學習者的語句，包括 200 短句與 100 句長句。錄製的學習者，包含了五個人（兩男三女），其資料如表(四)所示。

表(四) 學習者及語音錄製資料

國籍	性別	名字	錄製句數	長短句
印度	男	Guru	50	短句
印度	男	Mohan	50	短句
韓國	女	金寶林	100	50 短句+50 長句
日本	女	西田	100	50 短句+50 長句
法國	女	Grace	50	短句

以下列出人工評分的結果，將結果區分爲三個部份 A、B 及 C，可以比較出三位評分者的評分的差異。

A. 三位都評一樣的共有 141 句，其中(待改進：44 句，普通：16 句，很好：81 句)

B. 至少兩個評爲一樣有 329 句，其中(待改進：73 句，普通：61 句，很好：195 句)

C. 都不相同的有 21 句

由上述的結果可以得知，三位評分者對於"很好"的句子，有比較大的共同認知。

以下針對 A 及 B 與個結果，分別與機器評分的結果作比對，分別以表(A)及表(B)兩表表示。

表(A) 機器與人工評分比對表(141 句)

		人工評分		
		待改進	普通	很好
機器評分	待改進	19	2	0
	普通	19	3	3
	很好	6	11	78

由表(A)的部份可得知人工與機器評分結果的分析，如下

兩者評分相同的部份爲 100 句(19+3+78，表格對角線的部份)，佔了 70.9%。

機器評分較人工評分高的部份爲 36 句(19+11+6，表格下三角型的部份)，佔了 25.6%。

人工評分較機器評分高的部份爲 5 句(2+3+0，表格上三角型的部份)，佔了 3.5%。

表(B) 機器與人工評分比對表(329 句)

		人工評分		
		待改進	普通	很好
機器評分	待改進	31	9	1
	普通	30	12	19
	很好	12	40	175

由表(b)的部份可得知人工與機器評分結果的分析，如下

兩者評分相同的部份爲 218 句(31+12+175，表格對角線的部份)，佔了 66.3%。

機器評分較人工評分高的部份爲 82 句(30+40+12，表格下三角型的部份)，佔了 24.9%。

人工評分較機器評分高的部份爲 29 句(9+19+1，表格上三角型的部份)，佔了 8.8%。

由上述的結果得知，機器的評分與人工的評分結果大部份是相同的，但機器的評分有略爲偏高的情形。由於本實驗只是利用 k-means 的方式作聚類來得到機器的評分結果，所以這方面在未來可用機器訓練學習的方式來改善評分機制。

五、結論

本論文提出一個閩南語聲調辨認及評分之系統，在聲調的特徵擷取上，採用了四個特徵，爲聲調短時間至長時間的變化，包括 pitch curve、Spline curve、Fuzisaki phrase command curve 及 Fuzisaki accent command curve。其特徵值的計算是利用 DTW 的方式來求出學習者與標準語音特徵的差異。另外，也利用隱藏式馬可夫模型來切割學習者語音音節的邊界，以求得每個音節對應的分數。在實驗的部份，提出了實驗語音錄製的方法，並製訂一個人工評分的機制。最後，最後將機器的評分與人工的評分作比較，得到七成左右的相似度。然而，在機器的評分上，有略爲偏高的情形，提出可用機器學習的方式來改善此機制。在未來，系統會朝向以互動回饋的機制來發展，讓習學者知道何處需矯正及矯正的方式，以提昇學習的效能。

誌謝

本研究由經濟部委託財團法人資訊工業策進會創新前瞻技術計畫辦理。

This research was supported by the III Innovative and Prospective Technologies Project of Institute for Information Industry and sponsored by MOEA, ROC.

參考文獻

【1】 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

【2】 羅瑞麟，” 以語音辨識與評分輔助口說英文學習”，清華大學碩士論文，民國 93 年。

【3】 Lawrence Rabiner, B.H Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

【4】 Paul Boersma "Accurate Short-Term analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratip of A sampled Sound", 1993.

【5】 Patavee Charnvivit, et al., “Recognition of Intonation Patterns in Thai Utterance,” EuroSpeech, 2003.

【6】 Hansjorg Mixdorff, et al., “Towards the Automatic Extraction of Fujisaki model Parameters for Mandarin,” EuroSpeech, 2003.

【7】 Steve Young "HTK Book.3.1".

【8】 陳江村，羅瑞麟，張智星，李俊仁，"以語音辨識與評分輔助口說英文學習"，第十六屆自然語言與語言處理研討會, Taipei, Taiwan, Aug 2004.

【9】 H. Fujisaki and K. Hirose. “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” Journal of the Acoustical Society of Japan (E), 5(4):233--241, 1984.

【10】 http://neural.cs.nthu.edu.tw/jang/, CS Dept., Tsing Hua University, Taiwan.

【11】 http://www.myet.com/TW/Index.htm

【12】 http://www.manythings.org/pp/

【13】 http://www.speech.cs.cmu.edu/cgi-bin/cmudict

【14】 http://www.englishelearning.com/

【15】 Steve Young, *The HTK Book version 3*, Microsoft Corporation, 2000.

【16】 Lawrence Rabiner, B.H Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.