自動口說評估於英語作為第二語言學習者的初步研究

A Preliminary Study on Automated Speaking Assessment of English as Second Language (ESL) Student

Anonymous ACL submission

摘要

為順應國際化潮流，大學生對於國際交流以及英語授課所需之英文口說有越來越迫切的需求。本論文旨在發展自動化英語評測系統，並初探臺灣大學生之英語精熟程度。基於近期在臺灣所蒐集的口說語料，我們藉由一套自動語音辨識(Automatic Speech Recognition, ASR)系統，將語音轉寫成文字並擷取其中聲學特徵，最後使用機器學習模型來挑選適用的特徵以預測學生英語口說精熟度(English Speaking Proficiency)。經一系列在所蒐集的臺灣大學生口說測驗語料的實驗和分析顯示，使用機器學習方法來進行自動英語口說能力分級，能較專家人工分級有更高的穩定性。

Abstract

Due to the surge in global demand for English as a second language (ESL), Developments of automated methods for grading speaking proficiency have gained considerable attention. This paper aims to present a computerized regime of grading the spontaneous spoken language for ESL learners. Based on the speech corpus of ESL learners recently collected in Taiwan, we first adopt an automatic speech recognition (ASR) system to transcribe the speech into text, from which we subsequently extract pronunciation, fluency, and prosody features from the utterances, respectively. These extracted features are, in turn, fed into a tree-based classifier to produce a new set of indicative features as the input of the automated assessment system, viz. the grader. Finally, we use different machine learning models to predict ESL learners' respective speaking proficiency and map the result into the corresponding CEFR level. The experimental results and analysis conducted on the speech corpus of ESL learners in Taiwan show that our approach holds great potential for use in automated speaking assessment, meanwhile offering more reliable predictive results than the human experts.

關鍵字：自動發音檢測、英語能力分級

Keywords: automated speaking assessment, grader, CEFR

1 緒論 (Introduction)

為增加國際競爭力，臺灣的大學校園裡需要用到英語口說的情境大幅增加，舉凡國際交流以及全英語授課(English as a Medium of Instruction, EMI)，其中對於英文的口說能力的教學與測驗也有迫切的需求。根據過往經驗，學習者會使用線上的英語口說教學資源練習英語，而近年主流的應用程式[1]，在其口說練習中，題目以朗讀的發音練習為主，並使用自動語音辨識(Automatic Speech Recognition, ASR)檢視語者的音素(Phoneme)發音是否與系統中的母語語者的發音資料相同，來提供語者發音正確與否的回饋。

而劍橋自動化語言教學與評估中心(Institute for Automated Language Teaching and

[1] ELSA SPEAK: https://elsaspeak.com/en/ EF Hello: https://tw.hello.ef.com/

已刪除：i.e.,
已刪除：students'
已刪除：English
已刪除：此
已刪除：Taiwanese
已刪除：has
已刪除：on
已刪除：and can
已刪除：
已刪除：並
已刪除：of
已刪除：spoken
已刪除：.
已刪除：Subsequently,
已刪除：the ASR system and audio signal
已刪除：;
已刪除：t

Assessment, ALTA)所發展的 Speak & Improve 網站[2]，則提供更進階的口說練習，使用者會需要回答一連串的問題，經過系統分析，並得到以歐洲共同語言標準 (Common European Framework of Reference for Language, CEFR) 為分級的英語程度級數。相較於只考慮發音正確性的朗讀測驗，此系統需考量更全面的面向(如單詞使用、文法等)來取得整體式評分 (Holistic Scoring)。

受到上述多元的英語口說工具啟發，本研究想發展英語評測平臺以初探臺灣大學生的英語音韻程度。我們希望藉由多樣的特徵以及迴歸與分類模型來客觀地預測受試的臺灣大學生的英語音韻精熟度，發展適合臺灣本地大學生的英語口說評量平臺，以歐洲通用語言參考框架 (CEFR) 作為標準，有效地為其分析英語口說能力表現且分級，並期許大學生能透過平臺的回饋，使之成為自我英語口說精進之依據。

2 相關研究 (Related Work)

目前國外已有很多針對非母語語者之英語口說自動化測驗的研究，做為自動化評分依據來研究口說精熟度的特徵以及題型都相當多元。

Zechner et al. (2011) 於口說測驗中「朗讀」的任務上，藉由自動產生音韻特徵 (Prosodic Features) 來預測非母語人士英語精熟度分數，因為是朗讀文本，所以相較回答問題的題型，在語音辨識上的複雜程度相對較低。而 Knill et al. (2018) 使用 ASR 轉錄的文字資訊，探討了 ASR 的表現對於口說測驗中「回答問題」這類題型所造成的影響。由於 ASR 錯誤率會影響到自動化口說評分系統的結果，因此他們嘗試改進 ASR 轉錄文本上的單詞錯誤率(Word Error Rate, WER)，並納入其他語音相關特徵，以豐富其自動化口說評分系統。Craighead et al. (2020) 則僅基於來自 ASR 獲得的轉錄文本，使用多目標訓練的預訓練語言模型(Pretrained Language Model)，來為學習者評分。

除了使用文本作為英文精熟度分析基礎之外，也有一些文獻加入聲音或視覺特徵來預測不同任務上的英語口說精熟度。在人與機器的對話測驗上，Litman et al. (2018) 使用如 F0, 能量 (Power) 的聲音特徵來為非母語的英文口說精熟度評分。而在口說問答自動化評估的研究，Wang et al. (2018) 使用能量 (Energy) 作為其口說測驗訓練模型基準的輸入特徵之一。Saeki et al.(2021) 在為面試任務的口說任務上，使用詞彙、聲學以及視覺特徵，由類神經網路訓練並預測非母語口說語者的 CEFR 分級，其中的聲學特徵是使用音高 (Pitch) 和能量 (Power) ，而其實驗結果顯示結合詞彙及聲學特徵就能取得很好的正確率。

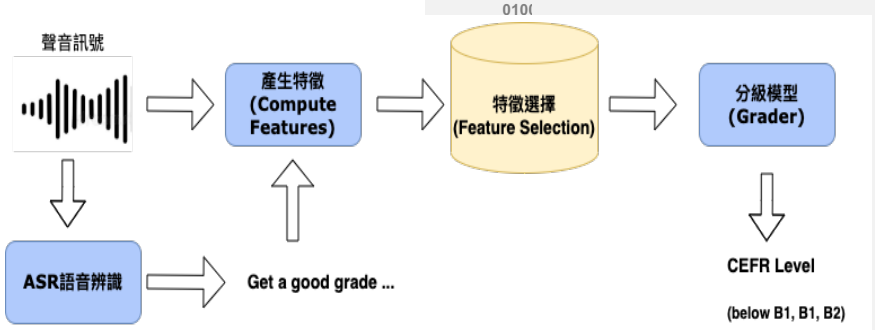

圖 1，自動化英語精熟度系統

[2] Speak&Improve：https://speakandimprove.com/

從前述的研究都顯示出從 ASR 與這些聲學特徵都能提升自動化為非母語學習者評價口說精熟度的有效性。而我們的自動化評測系統主要建立於聲音特徵上，會使用從 ASR 聲學模型獲得的以音段為級別的 (Segmental level) 特徵，與跨音段的超音段 (Suprasegmental level) 特徵，來處理從語言角度所分類之各面向特徵，來作為預測受試者英語程度的輸入特徵。

3 方法 (Method)

本研究中，方法的架構如圖 1。一共分成三個階段：第一階段，是使用預先訓練好的 ASR 模型與原始聲音訊號抽取聲音特徵；第二階段，使用極限樹 (Extra Tree) 來挑選適用於本任務之聲音特徵；第三階段，使用機器學習模型來預測受試臺灣大學生的音韻精熟度。在後續章節，我們將本論文所使用的方法拆分為特徵、特徵選擇，以及分級模型個別描述。

3.1 特徵 (Features)

綜合前述研究的成果，我們將此次任務的音韻特徵分成發音 (Pronunciation) 、流暢度 (Fluency) 與韻律 (Prosody) 面向，如表 1。在所有三個面向底下的特徵皆屬於聲音特徵，在發音與流暢度面向的特徵都是由 ASR 聲學模型將音訊以及文本對齊來獲得。而韻律面向的聲學特徵是從聲音訊號所抽取，以語音學的定義來看，韻律面向需包含發音長短、音量以及音高三種要素，而我們分別能藉由持續時間、能量以及基本頻率來獲得此資訊，而值得注意的是，本次任務尚未考量持續時間與韻律面向之關係，因此未納入表 1 之韻律分類中。在研究不同面向的特徵時，不同面向之間採用的特徵向量可能互有重疊，像是音素/字詞的信心分數 (Phone/Word Confidence) ，能同時成為探討發音和流暢度的要素之一。

表 1， 本研究音韻特徵

面向	特徵
發音 (Pronunciation)	Word/Phone Confidence
流暢度 (Fluency)	Silence
	Long Silence
	Disfluency
	Word/Phone Duration
	Word/ Phone Confidence
韻律 (Prosody)	F0
	Energy

3.1.1 發音 (Pronunciation)

英文非母語的學習者往往會將其母語的發音法連帶應用到英文上，進而產生發音誤差。若以音素來說明，其通常可被分為三類，分別是替代 (Substitutions) 、增加 (Insertion) 、刪除 (Deletion) 。母語的發音限制往往會觸發替代跟刪除這兩類錯誤，其中刪除對聽者的理解影響最大；替代則是使用類母語的發音去說其他語言 (Chen, 2016) 。

為了檢視發音誤差，我們使用音素與單詞級別的發音良好度 (Goodness of Pronunciation, GOP) (Witt and Young, 2000) 作為音段級別 (Segmental level) 特徵以計算信心分數，藉由取事後機率對數之持續時間標準化的方法 ，比對 ASR 所識別的文字和英文為母語者的發音模型。以 GOP 在音素級別的發音良好度公式為例：

$$GOP(r,n) \equiv \frac{\log P(Xr,n|Yr,n)}{Tr,n} \quad (1)$$

在 GOP 中，Yr,n為語者所產生的音素；Xr,n為相應的目標聲學段落；Tr,n 為聲學段落所經歷的時間範圍數量，其中r與n表示第r個語句中的第n個音素。

而受試者的總體發音與 ASR 模型的差異越大，獲得信心分數就越低。其中原因可能是發音不清楚或不正確，或是不流暢和語法錯誤。因此，信心分數可以反應非母語人士的

英語熟練程度，發音較好的受試者理應能獲得較高的信心分數 (Wang, 2018)。

3.1.2 流暢度 (Fluency)

流暢度也是音韻的其一面向，關乎受試者的講話語速、遲疑程度等。我們在單字級別的流暢度分析，會收集字數、語速、不流利度 (Disfluency)、重複字數等資訊。根據 (Loukina and Yoon, 2019) 的研究，英文程度好的第二外語受試者，往往能在相同時間內講出更多字詞。至於不流利度的衡量，我們會透過 ASR 轉錄之文字，以計算「um」、「uh」、和「hmm」這些遲疑文字的個數。而在停頓 (Silence) 和較長停頓 (Long Silence) 的特徵的細節上，我們使用 ETS 的方式來認定，當停頓超過 0.145 秒時，會做計算，而超過 0.495 秒時，會當作較長停頓。

3.1.3 聲學特徵 (Acoustic Features)

理論上探討超音段級別 (Suprasegmental level) 資訊，是要獲得與韻律相關的特徵，我們需要透過聲音訊號計算持續時間 (Duration)、能量 (Energy)及基本頻率(Fundamental Frequency, F0)。但在本次任務中，持續時間作為探討流暢度面向的特徵之一，而非韻律面向考量之範疇。

持續時間 (Duration)： 持續時間就是一個音素或單詞發聲的長度。根據 (Neumeyer et al., 2000) 的論文，音素的相對持續時間和專家評分的分數高度相關。因為通常英語學習者在說英語時，需要邊思考邊說，此行為會干擾講話的速率使其不流暢。而英語學習者也易於產生前段敘述所提到的三種發音錯誤 (替代、增加、刪除)，而導致其說英語時，會產生持續時間的差異，進而影響流暢度。

在計算持續時間時，我們統計在一次回答的過程中，音素及單詞層級發聲持續時間長度的平均值(Mean)、最大值(Max)、最小值(Min)、標準差(Standard Deviation, STD)、中位數(Median)、平均差(Mean Absolute Deviation, MAD)、總和(Summation, SUM)，作為兩個 7 維的特徵向量輸入 (Chao et al., 2022)

基本頻率(Fundamental Frequency, F0)： 基本頻率為語者聲帶振動的頻率，而反映在聽者的感知上，就會是音高。F0 的高低與重音 (Stress)以及語調有關。然而，根據 Sluijter and van Heuven (1996) 針對荷蘭與美國英語上重音與口音的研究，發現 F0 與重音之間沒有可靠的相關性。但是對於如母語為華語的英語學習者而言，使用 F0 來探討英語發音音高還是有其必要性，因為華語是聲調語言，音高的變化會影響到語意的不同(Tepperman and Narayanan, 2005)，而在英語上，音高可能只是傳達不同語氣。在此次任務上，我們評測的對象為母語為華語的台灣大學生，因此 F0 仍作為韻律面向的評斷特徵。如前述之持續時間特徵，我們也使用相同統計量來表示 F0 以及標準化 F0。

能量 (Energy)： 能量能最直接的反映語者的音量大小，而能量的分佈與語調 (Intonation) 有關。在我們的研究中，並沒有使用能量絕對值這個直覺的算法，因為其他研究顯示發音的品質和能量的絕對值沒有高度的相關性 (Dong et al., 2004)。相反地，我們使用均方根能量 (Root Mean Squared Energy, RMSE) 來計算每個音段的統計量作為韻律特徵 (Chao et al., 2022)，而我們使用與持續時間相同的統計向量來表示能量特徵。

3.2 特徵選擇 (Feature Selection)

在本論文中，我們使用極限樹分類器 (Extra Trees Classifier) 作為特徵選擇的方法。極限樹是隨機森林 (Leo, 2021) 的架構，其演算法在分割隨機樹的節點時，會隨機選擇切點；並且使用所有學習樣本來產生決策樹；在本節，我們使用極限樹分類器計算不純度 (Impurity) 作為特徵重要性，挑選適用之特徵。

3.3 分級模型(Grader)

自動化英文分級系統的優勢，在於透過統一標準來客觀地評論學生的 CEFR 分級，其公式可定義如下：

$$B_i = M(P_i, F_i, D_i, F0_i, En_i) \quad (2)$$

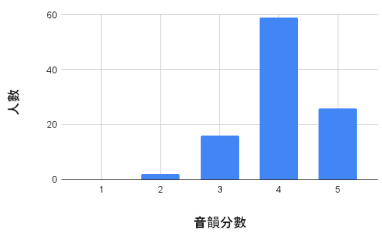

圖 2， 實驗語料之統計資訊

其中i為面向，本論文實驗中表示音韻，根據i面向所選取的特徵分別為：發音特徵P_i、流暢度特徵F_i、持續時間特徵D_i、基頻特徵$F0_i$與能量特徵En_i。M代表分級模型，B_i為最終之 CEFR 分級。經過特徵選取的機制，我們將這些特徵向量作為輸入值，經由迴歸與分類模型的訓練，評測出受試者的英文程度並對應到 CEFR 的級數。

本研究中的迴歸模型，分別有簡單線性迴歸 (Simple Linear Regression, SLR) 、多變項線性迴歸 (Multivariance Linear Regression, MLR) (Friedman et al., 2010; Kim et al., 2007) 、隨機森林迴歸 (Random Forest Regression, RFR) (Breiman, 2001; Geurts, 2006) 、支持向量迴歸 (Support Vector Regression, SVR) (Chang and Lin, 2001; Platt, 2000) 、梯度提升迴歸 (Gradient Boosting Regressor, GBR) (Friedaman, 2001; Hestie, 2009) 模型。根據多個特徵向量，我們使用迴歸模型分析這些向量間的關係來預測精熟度分數的連續數值，並在測試階段將迴歸模型預測的數值做分級，獲得 1 至 3 分的受試者分為 B1 以下; 獲得 4 分為 B1，獲得 5 分則為 B2。

而分類模型，我們使用邏輯迴歸(Logistic Regression, LR)、隨機森林分類器(Random Forest Classifier, RFC) (Breiman, 2001)、支持向量機(Support Vector Machine, SVM) (Chang and Lin, 2001; Platt, 2000)、梯度提升分類器 (Gradient Boosting Classifier, GBC) (Friedman, 2001; Hestie, 2009)、線性分類感知器 (Perceptron) (Freund and Schapire, 1999)。在分類模型中，我們將特徵作為向量輸入，精熟度分數則作為獨立的預測標籤。分類模型在訓練過程中會擬和兩者之間關係，並在測試階段做 CEFR 分級。

4 實驗評估與分析 (Performance Evaluation and Analysis)

4.1 語料 (Data)

口說分級模型若要有好的表現，多半需要有大量的人工標記語料，但目前公開的英語語料集多半是母語語者，僅有少量母語為華語語者的資料集。為了能準確分析學生之音韻表現，本研究收集大學生英語口說測驗語料來測試分級系統的有效性。本論文所使用的語料為英語教學專家設計的口說測驗，測驗內容為三部分：朗讀短文、回答問題與看圖敘述：朗讀短文不限制回答時間。回答問題共 10 題，分為 15 秒簡答與 30 秒詳答。看圖敘述則限制為 90 秒。共 103 位受試者，詳細的統計資料可參考表 2。語料標注流程如下：首先，我們會透過 ASR 自動轉寫語音內容，再透過人工做二階段的校閱。接著，該語料會交由兩位教學經驗豐富的英文專家根據內容、音韻及詞語三面向分別給予 1 到 5 分的精熟度分數，該分數可對應 CEFR 等級，在 CEFR 的框架中，將受試者的程度分成 ABC 三個層級，其中又再細分為 A1/A2 (基礎使用者) 、B1/B2(獨立使用者)、C1/C2 (精熟使用者)[3]。我們的系統分級方式同樣採用劍橋的分級概念：獲得 1 到 3 分表示語者的精熟度未達 B1；得到 4 分表示有 B1 程度；超過 4 分，都視為 B2 程度。本研究是採用第三部分的看圖敘述，測試語料只使用音韻面向的評分。總長為 2.6 小時，學生的平均回答單詞數量為 107 個 (見表 2) ；在看圖敘述的任務中，學生的音韻精熟度分佈如圖 2 所示。

在此次任務中，人工標記的專家之間在看圖敘述部分，其綜合內容、音韻及詞語之關聯性係數 Cohen's Kappa 值為 0.45，而屬於音韻面向的相關係數 0.47，在 0.4 到 0.6 之間的範圍都只屬於一般信度(Moderate)。若兩位專

表 2， 實驗語料之統計資訊

	朗讀短文	回答問題	看圖敘述
時數 (小時)	2.6	6.4	2.6
音檔數	103	103	103
最長回答 (詞數)	-	87	205
最短回答 (詞數)	-	1	6
平均回答 (詞數)	-	29	107

[3] 歐盟 CEFR 語言分級：https://reurl.cc/1mXRlG

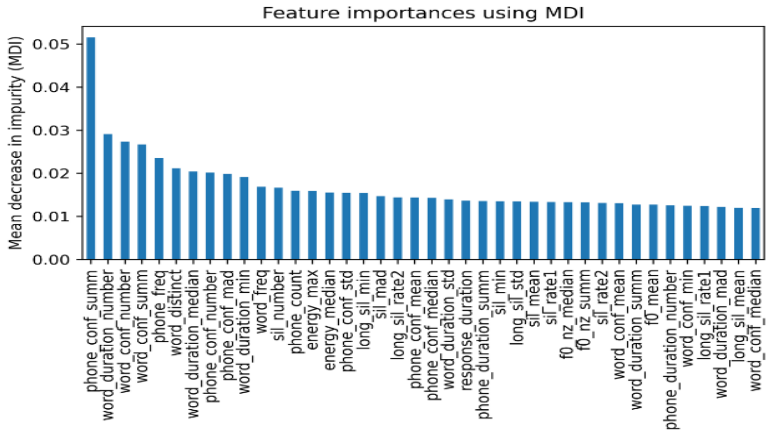

圖 3， 特徵重要性

家給予的評分相差一等級，則會請第三位專家給予實際等級。而這也相當程度反應人工標記語料之困難，因人類標記較易受個人所側重之觀點所影響，傾向於給印象分數而非實際表現分數。部分語料之語者，若是在音韻上表現較佳，而內容及詞語使用表現較一般，會產生評分結果不一致的情形。這種情況凸顯機器評測的重要性，因為機器能夠依照所設定之客觀標準分析語者口說程度。

4.2 實驗設定 (Settings)

本論文中，考量語音資料在不同情境下收音，需要克服噪音以利後續辨識，我們使用經多條件訓練 (Multi-condition Training, MCT) 的 ASR 模型，而聲學模型是使用改良的時間延類神經網路 (Time-Delay Neural Network, TDNN) (Povey et al., 2018)，是在深度神經網路 (Deep Neural Network, DNN) 的架構下，包含多層卷積網路和多層分解過的時間延類神經網路，簡稱為 TDNNF。語言模型則是使用 3-gram 語言模型。

在聲學模型的訓練語料上，我們使用有聲書讀物的英文公開語料 LibriSpeech (Panayotov et al., 2015)，而語言模型則是使用 TED-LIUM 3 (Hernandez et al., 2018) 語料做訓練。在自動化英語分級這個任務上，我們的 ASR 詞錯誤率 (Word Error Rate. WER) 為 30.08%。

在我們的實驗中，流暢度與發音面向底下的特徵是由 ASR 所產生，而韻律面向所需要的聲學特徵是使用 Python 的 Librosa (McFee et al., 2015) 模組所抽取。另外，我們使用 k-fold 交叉驗證 (Cross-Validation)，k 值為 5；所有的實驗採用的特徵一致。

4.3 效能評估 (Evaluation)

我們使用精確率 (Precision)、召回率 (Recall)、F1-score 來做效能評估，精確率指的是正確被辨識的項目，占所有被辨識項目的比例，召回率則是指正確辨識的項目，占需要被辨識項目的比例，F1-score則為精確率與召回率的調和平均數。

4.4 實驗結果 (Results)

4.4.1 特徵重要性 (Feature Importance)

為探究我們所使用特徵選取的效果，在此節我們探討使用的特徵在此任務之重要性，在不同的交叉驗證過程中，其特徵選取所產生的特徵大致相同，我們以圖 3 為例。由圖 3 可以發現，最重要的特徵是音素層級的信心分數總和，為口說清晰度的指標，符合本研究發音面向所需的特徵。我們也發現，此表

前半部分重要特徵，包含持續時間、信心分數、停頓數目等，則反映了我們流暢度面向的特徵。而後半部分的重要特徵有包含 F0 以及能量，能相當程度的考量語者的韻律特徵。總體來看，而對於機器而言，清晰度以及流暢度面向是重要指標，再來則是韻律面向。機器在特徵重要性的選擇上也符合我們從音韻角度思考並且設計特徵的趨勢，而多樣特徵的好處能夠使英語使用者不會因為單就音素發音錯誤，而被否定其流暢度以及韻律面向的表現，因為流暢度以及韻律這些超音段的特徵，也會影響到聽者的理解能力 (Chen et al., 2016)。

4.4.2 分級模型表現 (Grader Performance)

於此節我們分別將方法歸類為迴歸模型及分類模型來探討實驗結果，如表 3 與表 4 所示。其中的精確率、召回率、 F1-score 是根據類

表 3，迴歸模型表現

Regression Model	Precision	Recall	F1	Accuracy
SLR	0.67	0.65	0.63	0.65
MLR	0.71	**0.67**	**0.64**	**0.67**
RFR	0.55	0.59	0.56	0.59
SVR	**0.73**	**0.67**	0.63	**0.67**
GBR	0.62	**0.67**	0.63	**0.67**

表 4，分類模型表現

Classification Model	Precision	Recall	F1	Accuracy
LR	0.71	0.70	0.69	0.70
RFC	**0.74**	**0.73**	**0.71**	**0.73**
SVM	0.50	0.45	0.40	0.45
GBC	0.67	0.67	0.65	0.67
Perceptron	0.44	0.50	0.43	0.50

表 5，召回率混淆矩陣

召回率		RFC 預測結果		
		未達 B1	B1	B2
專家分級	未達 B1	0.52	0.43	0.05
	B1	0.02	0.84	0.14
	B2	0.00	0.36	0.64

表 6，精確率混淆矩陣

精確率		RFC 預測結果		
		未達 B1	B1	B2
專家分級	未達 B1	0.92	0.14	0.04
	B1	0.08	0.73	0.32
	B2	0.00	0.14	0.64

別(未達 B1, B1, B2)的加權平均，再由五次交叉驗證取得平均而得到。

表現最好的模型是隨機森林分類器(RFC)，在這個任務中，正確率為 73%; 精確率為 71%，然而相比之下，隨機森林在迴歸模型的表現較差，我們推論是因為此迴歸模型無法在超出訓練集的範圍做有效預測，而這可能會導致在不同的交叉訓練時，因使用不同特徵進行訓練，使此迴歸模型出現過度擬合的現象，在我們實驗中，的確有一次交叉驗證使用與其他驗證稍微不同的特徵，造成此迴歸模型獲得異常高的準確率。

從迴歸模型來看，簡單的迴歸模型(SLR)就能有 65%的表現，除了隨機森林迴歸模型(RFR)之外，支持向量迴歸(SVR)、多變項線性迴歸(MLR)與梯度提升迴歸(GBR)的準確率都為 67%，而 SVR 的表現又較突出。實驗中的迴歸模型皆沒辦法勝過最好的分類模型，但總體表現比分類模型穩定，其中，在分類模型上表現較差的支持向量機(SVM)和線性分類感知器(Perception)，我們推測原因是易受到資料量不足或標記信度不夠的極端資料之影響。

4.4.3 系統表現 (Performance Overview)

我們使用混淆矩陣來比較隨機森林分類器在預測結果跟人工分級上的分佈。在 103 份資料中，根據專家實際分級的結果：未達 B1 實際人數為 21 人；達 B1 者有 57 人；而 B2 程度者為 25 人。以召回率 (表 5)來看，所有為 B1 程度的學生中分級模型預測的召回率能達到 84%，大多可被模型歸類為 B1 等級，但從精確率 (表 6) 的角度來看，B1 的 73%精確率則稍差於未達 B1 的 92%精確率。

綜合兩張表格，我們發現機器若是沒有正確預測實際的分級，大部分的誤差也能落在一個級距以內。而這也反應和實際專家分級情境，在實際處理資料的過程中，評分專家也會有落差一個級距的情形。

因此，相比人工標記，專家之間音韻面向的相關係數只屬於一般信度(Moderate)，而機器的準確率都有六、七成，反應機器的表現相較於人為判斷可以相對客觀穩定。

5 結論 (Conclusion)

本論文是第一篇針對臺灣大學生發音之研究。使用語音學的觀點切入，探討如何為英語學習者的口說音韻精熟度分級，並且應用機器學習的模型，達到自動化分級英語精熟度的目的。從實驗結果發現，傳統的迴歸模型做法就能有良好的成效，若使用隨機森林分類器則能再提高準確率。未來我們將把持續時間加入至韻律面向，也加入不同的特徵，如：內容及文法特徵，至本英語精熟度評測系統，以期達到更全面的英語能力自動化分析，並給予英語學習者更清楚的回饋，進而幫助學習者提升整體的英語能力。此外，本次實驗因受限於少量資料，未能使用深度學習架構，未來除了會擴增相關語料外，也會探討少資源語料的訓練方向。

參考文獻 (References)

Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, and Yao Qian. 2015. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 338-345.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. In *Proceedings of Interspeech*.

Nancy F. Chen and Haizhou Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pp. 1-7.

Yu Wang, Mark Gales, Katherine M. Knill, Kostas J. Kyriakopoulos, Andrey Malinin, Rogier van Dalen, and M. Rashid. 2018. Towards automatic assessment of spontaneous spoken English. *Speech Communication,* 104, 47-56.

Anastassia Loukina and Su-Youn Yoon. 2019. Scoring and filtering models for automated speech scoring. In Klaus Z. and Keelan E. (Eds), *Automated Speaking Assessment*. pp.75-97.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2022. 3M: An Effective Multi-view, Multi-granularity, and Multi-aspect Modeling Approach to English Pronunciation Assessment, *arXiv preprint arXiv:2208.09110*.

Bin Dong, Qingwei Zhao, Jianping Zhang, and Yonghong Yan. 2004. Automatic assessment of pronunciation quality, in *Proceedings of ISCSLP*, pp. 137–140.

Silke Maren Witt and Steve Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*.

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohamadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proceedings of Interspeech*, pp. 3743-3747.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206-5210.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Proceedings of SPECOM*, pp. 198–208.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1), 5-32.

Brian McFee, Colin Raffel ,Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenbergk, and Oriol Nieto. 2015. Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, vol. 8, pp. 18-25.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech communication*, 30(2-3), 83-93.

Pavel Trofimovich and Wendy Baker. 2006. Learning Second language suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *Studies in Second Language Acquisition*, 28(1), 1-30. doi:10.1017/S0272263106060013

Klaus Zechner, Xiaoming Xi, and Lei Chen. 2011. Evaluating prosodic features for automated scoring of non-native read speech. In *proceedings of 2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 461-466. *DOI:10.1109/ASRU.2011.6163975*

K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines. 2018. Impact of ASR performance on free speaking language assessment. Proceedings of the Annual Conference of the International Speech Communication Association, *Interspeech, 2018-September 1641-1645. https://doi.org/10.21437/Interspeech.2018-1312*

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp. 2258–2269, Online. Association for Computational Linguistics.

Diane Litman, Helmer Strik, and Gad S. Lim. 2018. Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities, in Language Assessment Quarterly. *Vol. 15,* pp. 294–309*, Routledge.*

Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa and Tetsunori Kobayashi. 2021. Analysis of Multimodal Features for Speaking Proficiency Scoring in an Interview Dialogue. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pp. 629-635.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1), 1–22.

Seung-Jean Kim, K. Koh, M. Lustig, Stephen Boyd, and Dimitry Gorinevsky. 2007. An Interior-Point Method for Large-Scale L1-Regularized Least Squares. *In IEEE Journal of Selected Topics in Signal Processing, 2007.*

Pierre Geurts, Damien Ernst., and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning, 63(1), 3-42.*

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines (Software available at \url{http://www.csie.ntu.edu.tw/~cjlin/libsvm}

John Platt. 2000. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers.*

Jerome Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, Vol. 29, No. 5*

Jerome H. Friedman. 2002. Stochastic Gradient Boosting In *proceedings of Computational Statistics & Data Analysis* 38(4):367-378

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Elements of Statistical Learning Ed. *2, Springer*

Yoav Freund and Robert Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning. 37 (3): 277–296. doi:10.1023/A:1007662407062. S2CID 58856 17*

Agaath Sluijter and Vincent Van Heuven. 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *ICSLP 96. Proceedings of the Fourth International Conference on Spoken Language Processing ICSLP96, vol.2,* pp. 630 - 633. DOI:10.1109/ICSLP.1996.607440.

Joseph Tepperman and Shrikanth Narayanan. 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and*

Signal Processing, pp. I/937-I/940 Vol. 1, doi: 10.1109/ICASSP.2005.1415269.