中華民國第 61 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

(鄉土)教材獎

052501

原音重現—自動太魯閣族語語音評分系統

學校名稱：國立花蓮高級中學

作者：	指導老師：
高二 顏鈺峰 高二 王宇森	趙義雄

關鍵詞：機器學習、語音辨識與評分、太魯閣族語

摘要

近年來原住民語使用逐漸減少，面臨傳承的危機。本研究製作太魯閣族語音評分系統，希望幫助人們學習太魯閣族語。利用原委會網站的音檔和田野調查收集錄製的族語語音與族語老師的評分為本研究的研究語料，以 Kaldi 為語音辨識框架來製作語音辨識模型。對模型輸出的音素與正確單詞進行字串相似度計算，將計算數據與模型產出的發音品質分數作為選擇模型的依據，透過機器學習方法選擇最接近族語老師評分的模型。以此模型計算研究語料，再以機器學習製作一個七成正確率的評分分類器，最後實作出評分系統網站。此網站除了可以學習族語之外，也可以幫助未來繼續收集更多語料，提高評分系統的正確率。期望幫助人們學習太魯閣族語，讓原住民族語原音得以重現。

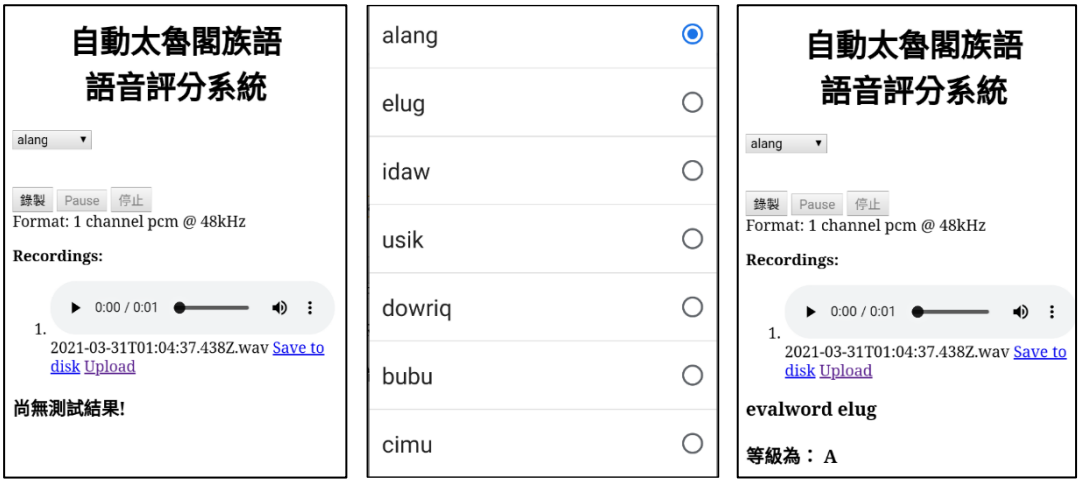
壹、研究動機

許多人會說「語言只是一種溝通的工具。」這樣講乍看之下沒有矛盾，但事實上，語言還是個文化傳承的利器，可以視為一個族群的特色、指標。但近年來，使用母語的人數逐年減少，尤其是原住民語，根據本土語言資源網的統計：「原住民多種語言屬於第 0 級的『滅絕』或第 1 級的『瀕臨滅絕』：其餘也多屬第 2 級的『嚴重危險」及第 3 級的『明確危險』。」相對於閩南語以及客家語，較有傳承的危機。[1]

在聯合國永續發展目標 SDGs 的第 11.4 條有提到「在全球的文化與自然遺產的保護上，進一步努力。」[2]我們身在全國原住民人口數最多的花蓮縣，想為原住民族語的傳承盡一份心力。由於我們的學校位在太魯閣族主要分布地區，學校同學與住家社區有很多是太魯閣族人，本研究將以在地的太魯閣族族語為研究對象，進行機器學習方法的研究。以田野調查收集太魯閣族語料，利用 Kaldi 及各種機器學習方法製作太魯閣族語音評分系統，希望能夠幫助有興趣的人學習太魯閣族語。

貳、研究目的

研究開發可以在行動裝置上，進行操作的自動太魯閣族語語音評分的系統網站，在評分分為 A, B, C(良、普、差)三個等第時，本系統評分的正確率與族語專家老師評分比較，正確率在七成以上，可作為太魯閣族語的學習工具，如圖一。本研究也將以田野調查的經驗與本評分系統，探討太魯閣族語的特性，並討論母語學生與非母語學生學習太魯閣族語的差別。



圖一、可在手機上操作的太魯閣族語語音評分系統

參、研究設備及器材

一、 AMD Ryzen 7-5700X 3.6GHz 八核心 中央處理器

二、 微星 GeForce RTX 2060 SUPER GAMING X 顯示卡

三、 16G 記憶體

四、 Linux 作業系統

五、 Kaldi Speech Recognition Toolkit

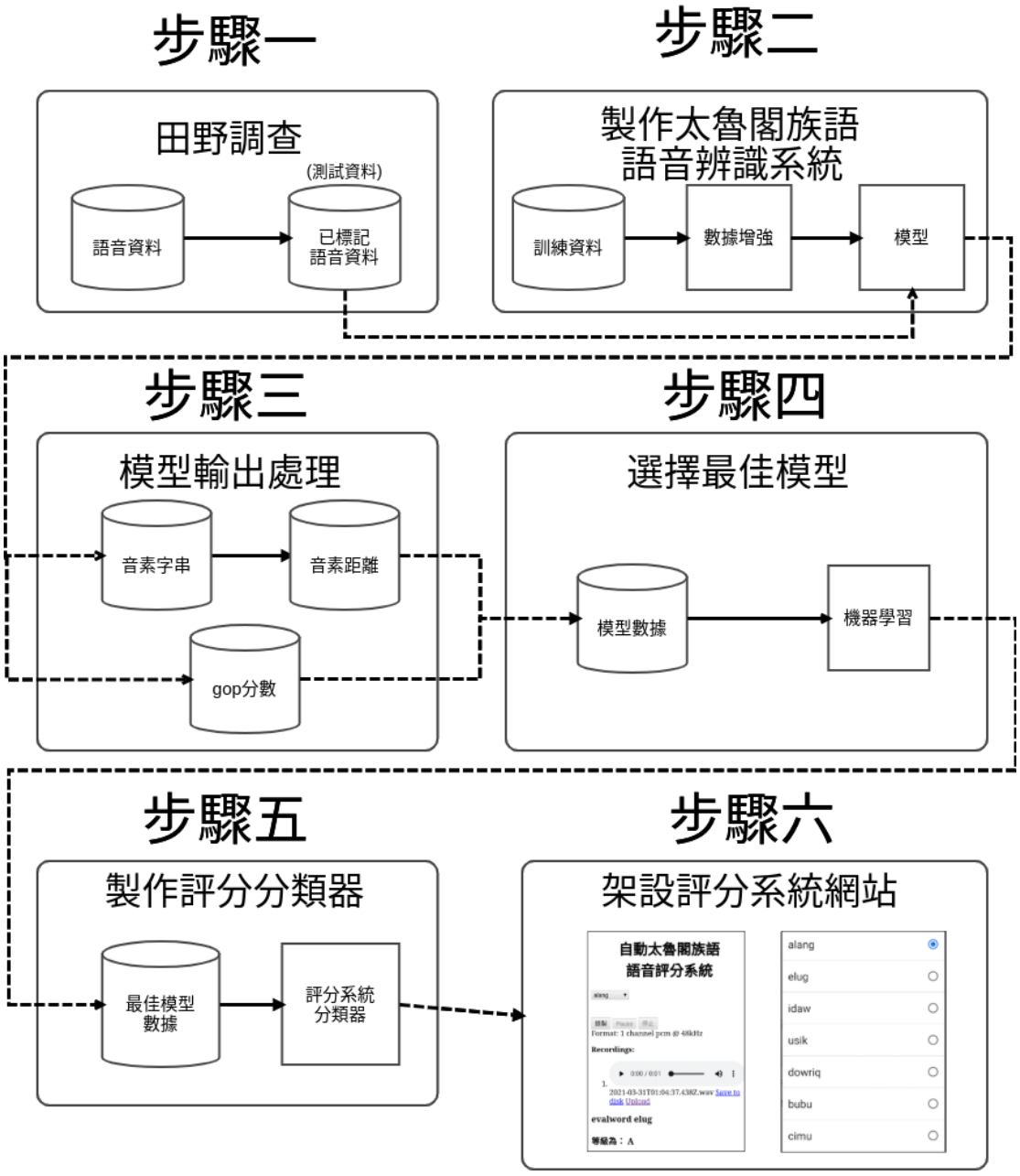
六、 Waikato Environment for Knowledge Analysis

七、 Flask 網頁伺服器模組

八、 Recorder.js 網頁錄音套件

九、 Python 程式語言

肆、 研究過程與方法



圖二、研究架構圖

圖二為研究架構圖，步驟一首先以田野調查方式，錄製三位研究學生與三位一般同學(非母語)和三位原住民同學(母語)的 31 個太魯閣族語單字發音，然後請族語老師評分，同時也教導我們如何學習太魯閣族語。然後步驟二使用 Kaldi 語音辨識系統來辨識產生音素與發音品質分數(GOP)，Kaldi 語音辨識系統的訓練資料為原住民委員會族語 E 樂園的太魯閣族語資料[3]，經由調整不同族語訓練

資料輸入與設定方式，可以產生各種不同的 Kaldi 辨識模型，在這階段總共產生 37 種辨識模型。步驟三再以字串相似演算法，計算模型輸出的音素品質。步驟四則是利用 python 進行線性迴歸、類神經網路、隨機森林法、支援向量機、梯度上升法、極限梯度上升法等，這六種方法來比較我們的田野調查結果和各種模型辨識的音素品質與 GOP 分數，來選擇最佳模型，也就是找出最接近我們田野調查結果的模型，步驟五再以最佳模型產生的數據來產生評分系統分類器。步驟六則是以最佳模型與分類器為根據架設主從式評分系統網站，並且可以在手機上操作此評分系統，同時也可藉此網站收集更多的太魯閣族語語料。以下詳述我們的每一個研究步驟：

一、步驟一：利用田野調查及文獻探討了解太魯閣族語的傳承狀態以及蒐集太魯閣族語的語音資料、評分的方式等。

為了蒐集太魯閣族語的族語語音資料、評分的方式，我們針對太魯閣族語的所有各種音素，挑選了 31 個單字，如表一。然後由九位受試者練習發音錄製太魯閣族語，總共有 279 筆族語語音資料，然後再請族語老師專家對於這 279 筆資料進行評分。對於我們的評分系統來說，279 筆族語語音資料為輸入，而我們的目標是：評分系統能夠自動輸出接近族語老師的評分。

我們實際訪問了兩位族語老師，這兩位族語專家的母語都是太魯閣語，主要評分者有二十多年的太魯閣族語教學經驗。在評分過程中我們了解到太魯閣族的某些發音與漢人差距很大。族語老師也提到，近年來願意學族語的年輕人愈來愈少，衷心盼望族語傳承狀況能有改善。

我們有九位受試者，分為三組(在研究結果的編號為 1-3)，第一組為母語者，錄音前，會先聽「族語 E 樂園」這個網站的範例[3]，再進行錄音；第二組為研究者，與第一組相同，會先聽範例音檔再行錄製；第三組為非研究者且非母語者，在錄製前，沒有聽過標準音檔，僅以羅馬拼音內容進行發音錄製。我們的錄製軟體為 AVR-X。

表一、本研究挑選的族語單字列表

編號	字詞	中文	編號	字詞	中文	編號	字詞	中文
1	alang	部落	12	kari	話	22	waqit	獠牙
2	elug	道路	13	limuk	鍋子	23	xiluy	鐵
3	idaw	白米飯	14	mkan	吃	24	yayung	河流
4	usik	辣椒	15	nalaq	膿	25	empeuyas	歌手
5	dowriq	眼睛	16	ngiyaw	貓	26	qowlit	老鼠
6	bubu	母親	17	payi	祖母	27	ciyux	梳子
7	cimu	鹽巴	18	qsiya	水	28	seejiq	人
8	duhung	臼	19	rhngun	門	29	btunux	石子
9	glu	喉嚨	20	sari	芋頭	30	kjiyay	蟬
10	hakaw	橋	21	tama	爸爸	31	towrah	肚兜
11	jiyum	使用						

二、步驟二：使用 Kaldi 製作太魯閣族語語音辨識系統。

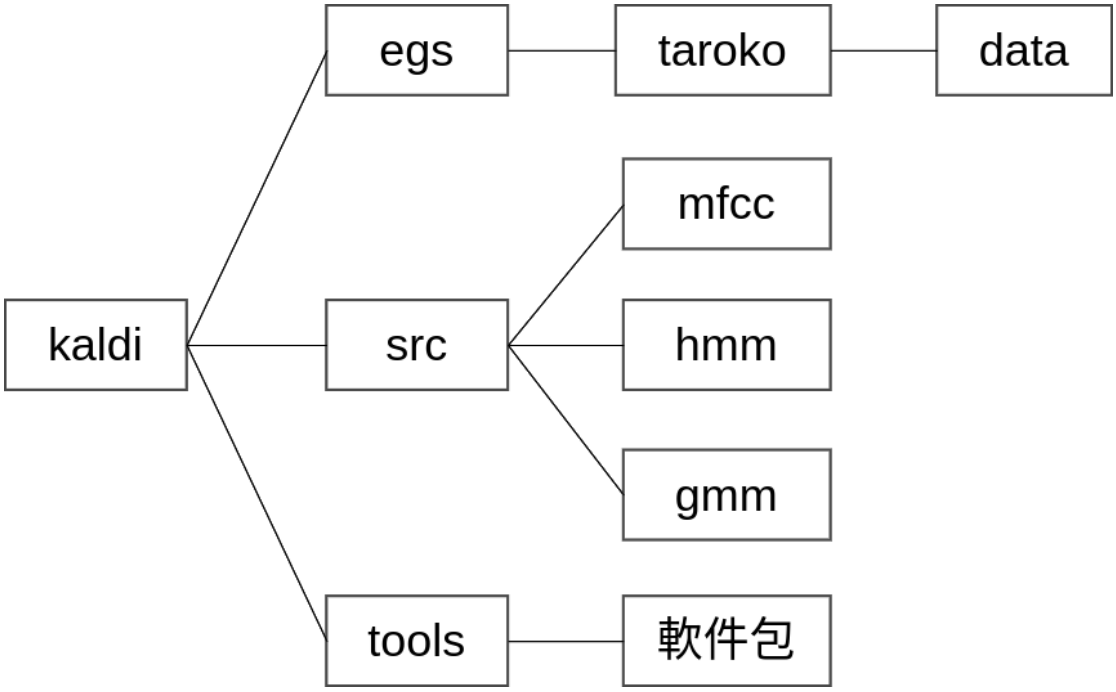
(一) kaldi 介紹

Kaldi 的命名由來，據說是來自於一個發現咖啡樹，來自埃塞俄比亞的牧羊人之名，在 Kaldi 尚未成熟時，研討會的參加人員多數為咖啡的愛好者，因此以此命名。

而此工具起源於 2009 年約翰霍普金斯大學夏季研討會，研究人員為了方便驗證實驗結果，開發了一個有限狀態轉換器為基礎的語音辨識解碼器，以及一些以語音辨識工具套件 HTK 為基礎的的語音辨識解碼器，這便是 Kaldi 的前身。在 2011 年 5 月 14 日，Kaldi 的出版程式庫正式發佈。[4]

而此套件之所以能獲得成功，主要來自於它容易閱讀的程式、易於

在不同線性代數庫之間轉換、以及通用的演算法等。最重要的，它是開放原始碼，吸引了很多企業及開發者，不斷的優化效能，才有今日的成功。圖三為我們此次研究使用的 Kaldi 之架構。



圖三、Kaldi 的架構圖

Kaldi 的一級主目錄中包括：egs(Kaldi 的實例集)、src(存放 Kaldi 的源代碼，包括 GMM、HMM…等一系列的傳統語音識別算法)、tools(存放 Kaldi 安裝的軟件包)等文件夾。我們使用的 hmm、gmm 和 mfcc 算法便是存在 src 中，egs 中的 data 為存放語言模型、發音字典和音素信息。我們研究的語音辨識框架有參考東華大學的研究生 Serkan 開發的語音辨識框架[6]。

我們使用 Kaldi 評分的流程如下:

1.特徵提取

一般來說，我們使用 MFCC 為訊號做特徵提取。MFCC 稱為梅爾倒頻譜係數，將原訊號進行傅立葉轉換，並取對數、通過 Mel 濾波器，最後透過離散餘弦轉換得到梅爾倒頻譜係數。詳細步驟如下：

(1)預處理

使用高通濾波器進行濾波，使高頻部分更加明顯。濾波器的濾波方式為:$H(z)=1-az^{-1}$ (z 代表原始信號，H 代表濾波後的信號)也可寫為：x[n]=x′[n]-ax′[n-1](x′代表原始信號，x 代表濾波後的信號)，其中 a 為一個介於 0-1 之間的值，通常取 0.95 或 0.97。

(2)Endpoint detection

將原始訊號分為數個幀(一般幀長為 20-40ms，且相鄰的兩幀有 1/2 或 1/3 的重疊)。

(3)MFCC[5]

a.對每個幀進行快速傅立葉轉換，將時域信號轉為頻域

$$Xa(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\frac{\pi k}{N}},0\le k\le N$$

公式中的 N 一般取 13。

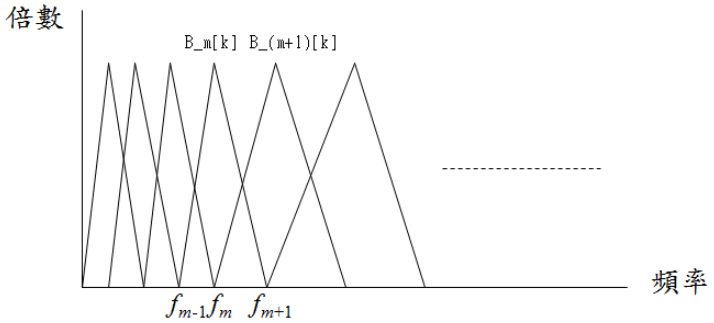
b.通過 Mel 濾波器:

人耳對不同頻率聲音的敏感度並不相同，在聽到某些頻率的聲音時會比聽到其他頻率的聲音更容易注意。因為此特性，有時某些頻率的聲音會把其他聲音給遮蓋掉。且聽覺並不是線性，而是指數性(如低音 Do、中音 Do 與高音 Do 的頻率相差兩倍，在頻譜上不是線性關係)。Mel 頻譜便是對應這些人耳特性所發明，使用 fmel= 2595 ∗log10(1 +f/700)將原本的頻譜轉換，能表現出更多的聽覺特性，因此提取特徵是在 Mel 頻譜上進行，使用 Mel 濾波器表現不同頻率的掩蔽現象。Mel 濾波器是由許多三角濾波器組形成的集合，如圖四，詳細的濾波方式如下:

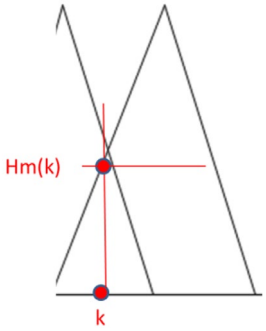
$$\mathbf{Hm(k)\{0}\boldsymbol{k}<\boldsymbol{f}(\boldsymbol{m}-\mathbf{1})\ \frac{\boldsymbol{k}-\boldsymbol{f}(\boldsymbol{m}-\mathbf{1})}{\boldsymbol{f}(\boldsymbol{m})-\boldsymbol{f}(\boldsymbol{m}-\mathbf{1})},\boldsymbol{f}(\boldsymbol{m}-\mathbf{1})<\boldsymbol{k}<\boldsymbol{f}(\boldsymbol{m})\ \mathbf{1}:\boldsymbol{k}$$

$$=\boldsymbol{f}(\boldsymbol{m})\ \frac{\boldsymbol{f}(\boldsymbol{m}+\mathbf{1})-\boldsymbol{k}}{\boldsymbol{f}(\boldsymbol{m}+\mathbf{1})-\boldsymbol{f}(\boldsymbol{m})}\boldsymbol{f}(\boldsymbol{m})<\boldsymbol{k}<\boldsymbol{f}(\boldsymbol{m}+\mathbf{1})\ \mathbf{0}\boldsymbol{k}>\boldsymbol{f}(\boldsymbol{m}+\mathbf{1})$$

可以將 Hm(k)理解為原始頻率 k 在高為 1、左右頂點 x 座標為 f(m-1), f(m+1), 中點座標為 f(m)的三角形邊上的投影，如圖五。



圖四、mel 濾波器



圖五、Hm(k)

c.提取對數能量

對於每個 Mel 濾波器，計算對數能量：

$s(m)=\ln\left(\sum_{K=0}^{N-1}|X_a(k)|H_m(k)\right)$, $0 \leqq m \leqq M$

離散餘弦轉換

DCT）得到 MFCC 係數：

$$C(n)=\sum_{m=0}^{N-1} s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),\ \mathbf{n=0,1,2..L}$$

將上述的對數能量帶入離散餘弦變換，求出 L 階的 Mel 參數。L 階指 MFCC 係數階數，通常取 12-16。這裡 M 是三角濾波器個數。

2.隱含馬爾克夫模型(HMM) [7][8]

隱含馬爾克夫模型(HMM)是美國數學家鮑姆(Leonard E.Baum)在 1960 年代所發表的一系列論文所提出的。先從馬爾可夫鏈說起，假設時間為 t，狀態為 S_t，假設所有狀態都是隨機的，任何狀態的取值都與周圍的狀態有關(S_t 與 S_{t-1}、S_{t-2}…有關)因此這個隨機過程就有二維的不確定性，因此馬爾可夫提出了簡化的假設 S_t 的狀態止於前一個狀態 S_{t-1} 有關($P(S_t|S_1,S_2,S_3\ldots S_{t-1})=P(S_t|S_{t-1})$，此稱為馬爾可夫鏈。而 HMM

便是馬爾可夫鏈的延伸，屬於圖機率模型，是結合圖和機率的學習方法，包含了：

(1)Represention：有隨機變數間的關係

(2)Inference：給定一個狀態，能估計其他隨機數

(3)Learning：給定資料能夠學習出圖形結構或參數

但 HMM 在任意時刻的狀態是不可見的，無法透過觀察得到轉移機率等的參數，但 HMM 在每個時刻會有個僅跟 S_t 相關的獨立輸出 O_t。我們可以計算出某個特定的狀態(S_1,S_2,S_3…)產生出獨立輸出(O_1,O_2,O_3…) 的 頻 率 ： P(O_1,O_2,O_3…|S_1,S_2,S_3…)= $\prod p(S_t|S_{t-1}) * p(O_t \vee S_t)$，很多語言的處理都適用此方式來解決。

而 HMM 模型有三個相關問題：

(1)給定一個模型，計算出某個特定輸出(O_t)的概率

(2)給定一個模型和一個特定輸出(O_t)，找出最可能找出這個輸出的狀態(St)

(3)給定足量的數據，估計 HMM 模型的參數

而第二個問題便是我們此次研究會用到的，可以使用維特比算法來解決。對於語音信號處理，Q 通常為不同的音(如「9」念作「ㄐ ㄧㄡ」，我們便可以把「ㄐ」、「一」、「ㄡ」劃分為不同的狀態)。O 是輸入的訊號。而 A(轉移矩陣)、B(觀察概率)可以透過訓練獲得：

3.求得 B(使用 GMM)[2]

由於每個人講話的音調、速度都不相同，每種單字的標準念法也都不只一種，因此求出 HMM 中觀察概率(B)的方式便非常重要。一般而言，我們使用高斯混合模型(GMM)來求出 HMM 中的觀察概率。

我們認為音檔的特徵遵循常態分佈法則，因此對於狀態 Si，只要有足夠的訓練資料，便可建構出分曲線，透過常態分佈公式求得該狀

態的 Bi(o)函數:

$$\mathbf{N}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi|\Sigma|}}\boldsymbol{exp}\left[-\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

注意這邊的 x,u 皆為向量，x 表示的是樣本向量，u 則是訓練資料的期望，Σ則是模型方差矩陣。但對語音評分(如單字辨識)來說，期望值(u)代表的不一定是單字的標準念法，例如以聲調頻率來說，女性的一般比男性高。若單純以唯一的 u 計算分布概率，效果便會相當差。即使是非常標準的男性音檔，也會跟 u 有相當大的差距(因 u 包括女性的聲檔)。而 GMM 便是為了解決此問題而生。GMM 的功能與一般高斯模型相同，但不使用唯一的高斯模型當作基準。為了對應不同的人聲，GMM 是多個高斯模型的集合，因此可以涵蓋較廣的聲音種類(GMM 可對男聲和女聲各自建立高斯模型)，並且對每個不同的高斯模型設定權值(所有權值和應為 1)。計算一組音素在 GMM 上的分布概率時，便分別算出該組音素在每一組高斯模型上的分布概率，再進行累加即可:

$$\mathbf{p(x)} = \sum_{j=1}^{P}(j)\boldsymbol{p}(\boldsymbol{x}\vee\boldsymbol{j}),\ \boldsymbol{P} = \sum_{j=1}^{P},\boldsymbol{P}(j)\boldsymbol{N_j}$$

4.訓練[8]

對於一個已給定 Q、B 的 HMM，我們使用向前-向後算法為 HMM 訓練，而他為何叫做向前-向後算法，主要是因為此算法既可用遞歸算出前向的機率，也可以計算此模型最終狀態的後向機率，因而命名。向前-向後算法是一種基於動態規劃的算法，首先對於隱藏馬爾可夫模型的參數進行一個初始的估計，一開始這可能是完全錯誤的，然後對於給定的數據來減少參數所造成的錯誤。是以梯度下降的方式尋找一種錯誤的最小值。

5.解碼[8]

若我們已訓練好一組 HMM，對於任意觀察序列 O，我們通常用

維特比演算法(Viterbi)求出對於 O 的最適觀察序列 Q。維特比演算法同樣是基於動態規劃的演算法。我們定義 V 為維特比演算法的網格，每隔 $V_t(j)$代表的是 HMM 在經過最適狀態序列的前 t-1 個 state $(q_1,q_2,..,q_{t-1})$後，停留在 q_j 的機率。我們可寫出狀態轉移方程：

$$\mathbf{Vt(j)} = \boldsymbol{max_{i=1}^{N} V_{t-1}(i) * a_{ij} * b_j(o_t)}$$

與向前算法的網格非常相似。我們可用此狀態轉移方程算出所有 $V_t(j)$，並用此求出最適序列。同時，對於每個 $V_t(j)$，我們用一個反向指標 $P_t(j)$表示 $V_t(j)$的轉移點$P_t(j) = argmax_{i=1}^{N} V_{t-1}(i) * a_{ij} * b_j(o_t)$。在最後選出最大的 $V_{t(end)}(j)$，即為最適序列的終點。再透過 P 反向推出最適序列。

(二) 訓練資料

我們將「族語 E 樂園」的新九階教材所有音檔作為基本訓練集，再加上步驟一所提到的 31 個單字之標準音檔為此研究之訓練集。

(三)數據增強

由於蒐集的族語語料不多，因此我們加入**數據增強**的實驗，數據增強是指在有限的訓練集中，透過適度的調整產生等價的訓練集，以克服訓練集不足的情況，由於我們研究的太魯閣族語語料稀少，因此我們決定採用三種方式：

1. 速度增強：將音檔速度進行 0.5 倍、1.5 倍的處理。
2. 加入噪音：將音檔前後加入白噪音。
3. 頻率增強：音檔的特徵放大，頻率提升，時長不變。

三、 步驟三：使用不同訓練集產生 37 種 Kaldi 模型，根據田野調查資料，計算模型分數，並挑選最佳模型。

每個模型評分方法如下：

(一)以模型辨識測試音檔資料，產生音素(text)與 GOP 分數

1. GOP(Goodness of Pronunciation)介紹[4]：

錯誤發音檢測的早期研究中，有學者將模型算出的輸出改良並稱作 GOP，也是最常被使用的發音檢測方法。將 GMM-HMM 算出的輸出使用 GOP 進行發音檢測。

GOP 的計算方式如下：

$$\mathrm{GOP(u,n)} \equiv \frac{1}{T_{u,n}} \log P(q_{u,n}|O_{u,n})$$

其中 GOP 是音素段落$O_{u,n}$對應目標音素$q_{u,n}$的事後機率，u 與 n 表示第 u 個語句的第 n 個音素。

(二)使用字串相似演算法計算比較產生的音素與原始音素，方法為 Hamming 以及 Levenshtein[11]

1.Hamming distance：兩個字串對應位置的不同字符的個數

2.Levenshtein：指兩個字串之間，由一個轉成另一個所需的最少編輯操作次數，允許的編輯操作包括：(1)將一個字符替換成另一個字符 (2)插入一個字符 (3)刪除一個字符。

四、 步驟四：將模型數據用不同的機器學習方法訓練出最佳模型

(一)將模型數據的評分數字轉換成三種類別

原本評分數字分為 1 分、0 分、-1 分，轉換為 A、B、C 三個等級，分別表示良、普、差。將此結果修改最佳模型數據。

(二)訓練模型的方法

以下前五種方法都是使用 python 的 scikit-learn 模組進行實作。而第六種方法是使用 python 的 XGBoosting 模組進行實作。[12] [14]

1.以線性迴歸(Linear Regression)訓練最佳模型

線性迴歸是利用稱為線性迴歸方程的最小平方函數對一個或多個自變量和因變量之間關係進行建模的一種分析方式。

2.以類神經網路(Neural Network，NN)訓練最佳模型

神經網路是一種模仿生物神經網路結構和功能的數學模型或計算模型。神經網路由大量的節點和節點之間互相連結而組成，每一種節點代表一種特定的輸出函數，稱為激勵函數。每兩個節點間的連結都代表一個對於透過該連接訊號的加權重，稱為權數，這相當於神經網路的記憶。網路的連接方式、激勵函數和權數值決定網路的輸出。

3.以隨機森林(Random Forest)來訓練最佳模型

隨機森林模型是一個由數百棵沒有關聯的決策樹所組成，決策樹是由不同的狀況所組合而成的有向無環圖，可以通過輸入變量，找出最符合的獨立結果，而一般的決策樹演算法有一個重要的操作-剪枝，以避免過於龐大的子葉而造成過擬合的情形。

而在隨機森林模型中，透過自助法重取樣技術，從原始 N 個樣本集中挑取 k 個樣本，根據這 k 個樣本產生 k 個分類樹組的隨機森林，每個決策樹皆會產生一個結果，在迴歸問題中，隨機森林的輸出會是所有決策樹輸出的平均值；而在分類問題時，會將結果進行投票，票數最多者即為輸出。

4.以支援向量機(Support Vector Machine)來訓練最佳模型

支援向量機是由 Vladimir Naumovich Vapnik 和 Alexey Yakovlevich Chervonenkis 於 1963 年發明，為一種可推展到線性不可分問題的線性分類器，以 margin(切割線中與分堆的最短距離)最大化的方式建構模型，並透過選擇不同的核函數來預測處理不同的資料。此外，決策函數是由少量的支持向量決定，預測效率高。

5.以梯度上升法(Gradient Boosting)來訓練最佳模型

Boosting 是指結合許多弱分類器，藉由改進每一次的錯誤，而獲得一個強分類器的方法，而 Gradient Boosting 可以透過不同損失函數

的使用，可以處理排名、分類等不同的問題。其產生的預測模型是弱預測模型的集成，例如採用典型的決策樹作為弱預測模型，可稱為梯度提升樹（GBT 或 GBDT）。像其他提升方法一樣，它以分階段的方式構建模型，但它通過允許對任意可微分損失函數進行優化，可作為一般提升方法的推廣。[13]

6.以極限梯度上升法(eXtreme Gradient Boosting)來訓練最佳模型

XGBoost (eXtreme Gradient Boosting) 是基於 Gradient Boosted Decision Tree (GBDT) 改良與延伸。它的目的在於提供一個可擴展、可移植和分布式梯度提升(GBM、GBRT、GBDT)庫。近幾年，由於這個演算法受到許多在機器學習競賽中獲獎團隊的青睞，因而受到了廣泛的歡迎和關注。[14]

(三) 選擇最佳算法

透過實驗所獲得的數據，能整理成各模型數據的表格，模型數據屬性如表二。前兩個屬性為使用字串相似演算法計算比較產生的音素與原始音素的結果，GOP 分數為 Kaldi 語音辨識所產生，預設有 12 個屬性。透過**上述算法**分析，如果平均準確度越高，表示此算法容易訓練出接近族語老師評分的模型，我們將所有模型的準確度進行平均，以找出最好的算法。

表二、模型數據屬性內容

Hamming distance	**Levenshtein**	**Gop 分數 1**	**Gop 分數 2**	**…**	**族語老師評分(輸出)**

(四) 選擇最佳模型

透過實驗所獲得的數據，能整理成各模型數據的表格，模型數據屬性如表一。透過**上述算法找出的最佳算法**分析，如果準確度越高，表示輸入語音的分析內容與族語老師的評分相關性越高，也就是其模型的品質越好。得到最佳模型後，就可計算出最佳模型數據，作為訓練

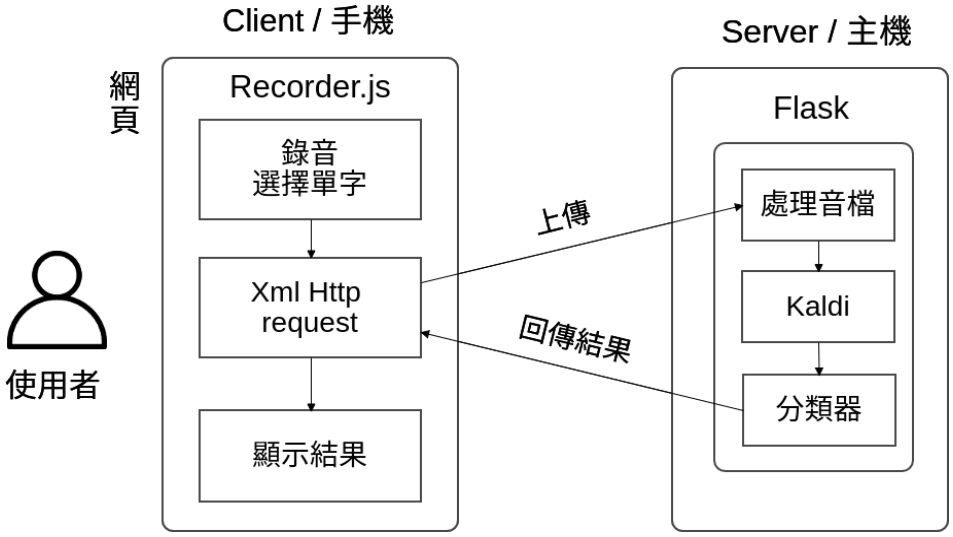
評分分類器的依據。

五、步驟五：以最佳模型搭配最佳算法產生評分分類器，並且驗證分類器準確度

使用上一步驟選擇的最佳模型，計算田野調查資料，輸出音素比較分數與 GOP 分數，作為評分分類器的訓練資料，訓練出評分分類器。評分分類器的品質則是使用交叉驗證法(k-fold corss validation)計算，交叉驗證是一種將樣本切割成多個小子集的做測試與訓練。本實驗將採用十折交叉驗證法（10-fold corss validation）作為交叉驗證。

六、步驟六：設計評分系統

圖六為我們設計的評分系統架構，此為主從式架構，客戶端使用 Javascript 模組 Recorder.js 透過瀏覽器進行錄製，再以加密通訊協定 https 上傳錄音結果至伺服器，此伺服器是以 python 的 Flask 模組設計的網頁伺服器，經過處理音檔後，輸入我們事前訓練好的 Kaldi 模型，得到 GOP 分數與發音音素，再輸入至評分分類器當中，得到評分結果後，再回傳至客戶端的網頁中輸出結果。



圖六、評分系統架構圖

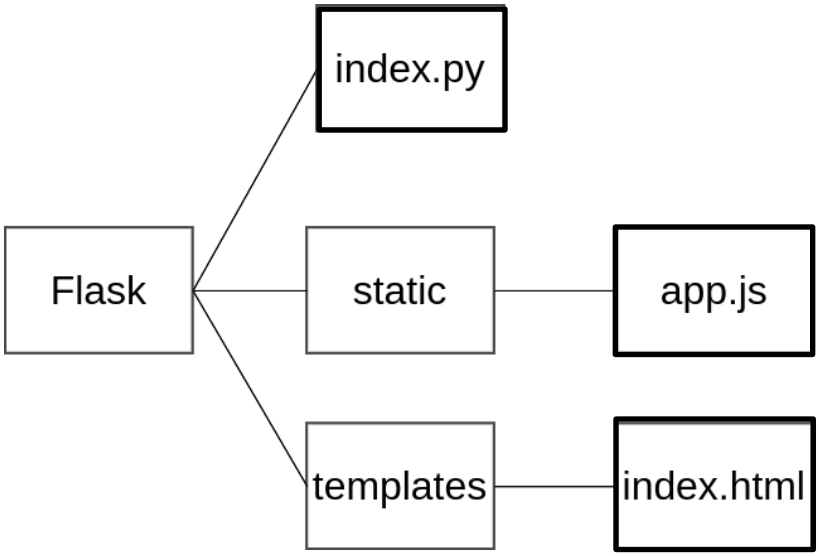
我們的系統為 python 的 flask 模組所設計，如圖七，此系統主要由三個主程式所構成：

(一)index.py：Flask 的主程式，使用 python 所撰寫的微框架，核心十分簡單，

主要是由 Werkzeug WSGI 工具箱和 Jinja2 模板引擎所組成。

(二)static/app.js：recorder.js 的程式，以 javascript 撰寫的錄音套件。

(三)templates/index.html：網站的圖形化介面和表單設計。



圖七、Flask 架構

伍、研究結果

我們的實驗資料為研究方法(一)所蒐集到的族語語音資料和族語老師評分，測試的部分，我們將「族語 E 樂園」的新九階教材所有音檔作為訓練集(以下簡稱基本訓練集)，再分別加上此次研究使用單字的標準音檔作為訓練集，我們的實驗結果如下：

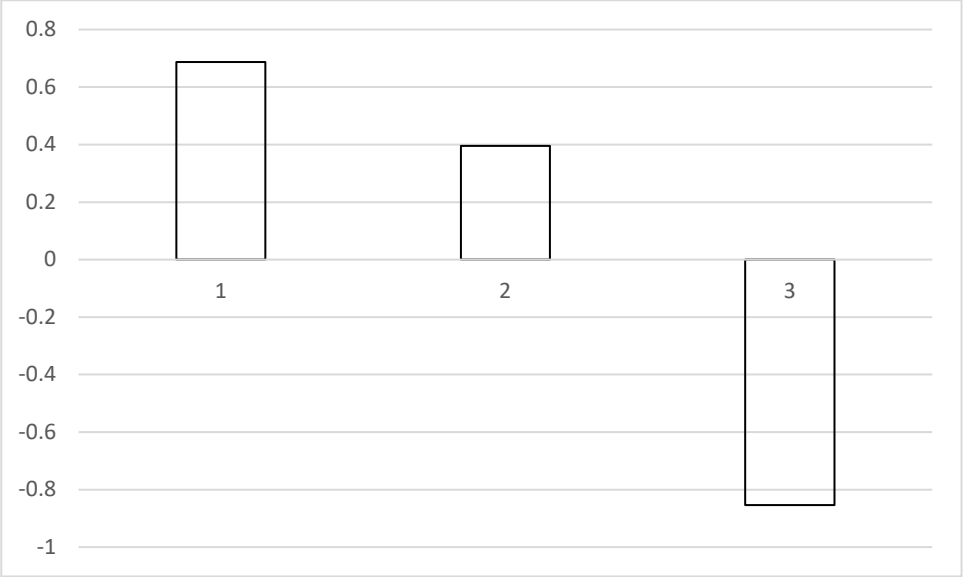
一、母語和非母語發音太魯閣族語音素的差別

我們有九位受試者，分為三組(在研究結果的編號為 1-3)，第一組為母語者，錄音前，會先聽「族語 E 樂園」這個網站的範例[3]，再進行錄音；第二組為研究者，與第一組相同，會先聽範例音檔再行錄製；第三組為非研究者且非母語者，在錄製前沒有聽過標準音檔，僅以羅馬拼音內容進行發音錄製。圖中的數字表示該組在特定單字下的平均分數。

我們事先錄好測試集，並且現場給太魯閣族語教學者進行評分，分數

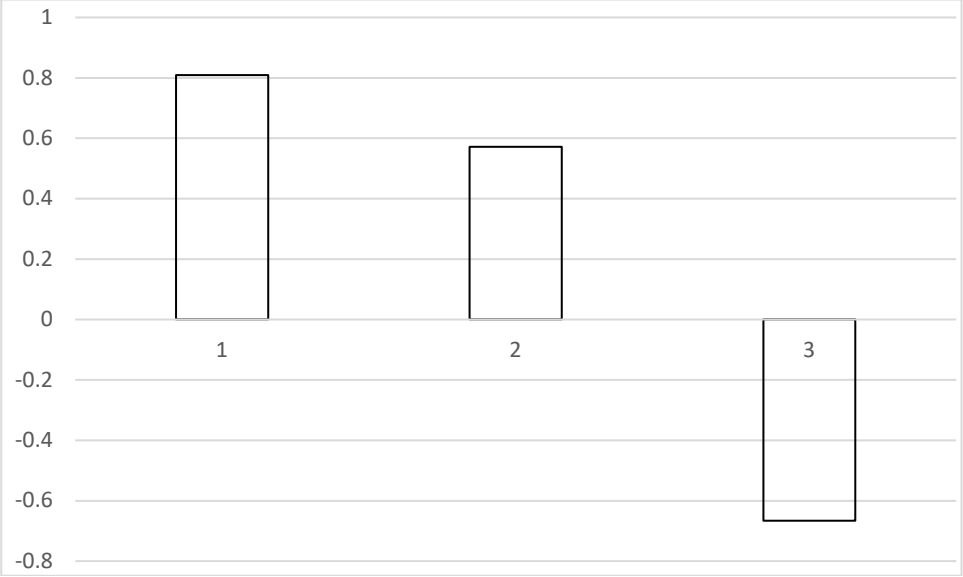
由好至差分類為三種(1 為良、0 為普、-1 為差)。而太魯閣族語教學者也提出：「l」、「ng」、「x」、「t」、「h」是一般人比較不容易念好的，我們將其分別和沒有上述音標之單詞進行比較，以下為我們的實驗結果，此結果可以做為設計評分系統的參考。

從圖八可以看到，在單詞出現上述之音標時，第一組分數明顯較其他兩組高，而第三組分數為最低。



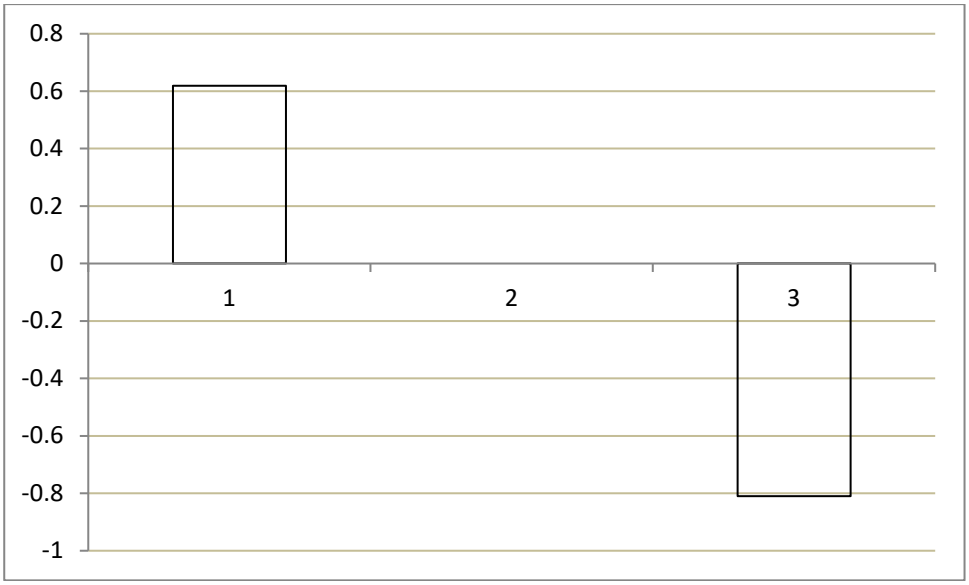
圖八、在單詞出現上述之音標時，專家的評分情形

從圖九可以看到，在單詞未出現上述之音標時，第一組分數明顯較其他兩組高，第二組的分數有明顯提升，而第三組分數仍為最低。



圖九、在未出現上述之音標時，專家的評分情形

從圖十可以看到，在單詞出現「l」音標時，第一組分數明顯較其他兩組高，第二組的分數有明顯下降，而第三組分數仍為最低。



圖十、在出現「l」音標時，專家的評分情形

二、比較使用線性迴歸、類神經網路、隨機森林法、支援向量機、梯度上升法、極限梯度上升法比較模型，找出其中最適合選擇最佳模型的算法

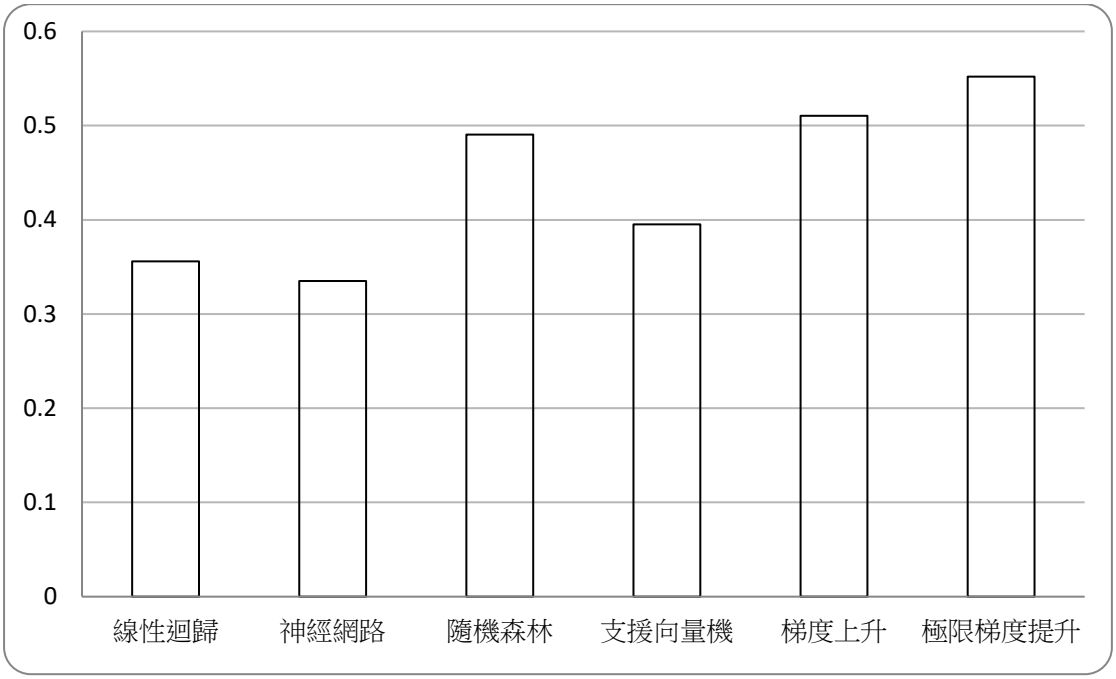
此實驗有 37 個模型，訓練集除了基本訓練集外，也會加入其他單字的標準音檔進行訓練，如表三。

表三、訓練集加入音檔列表

模型名稱	加入訓練集的音檔編號
1-31	對應編號之單字音檔
l	1、2、9、13、15、23、26
ng	1、8、16、19、24
x	23、27、29
t	21、22、26、29、31
h	8、10、19
all	所有單字

此外，我們在訓練模型時，也會藉由調整參數，來做出不同的模型。而這個

實驗的目的是藉由訓練模型的過程中，找出最適合我們研究的算法，我們將以十折交叉驗證所得的準確度(評分對應於族語老師的正確率)作為主要的比較方式。從圖十一中的實驗結果可以發現，隨機森林、梯度上升和極限梯度提升準確度較其他演算法高，趨近於 0.5。



圖十一、各算法進行模型比較時的準確度總平均關係

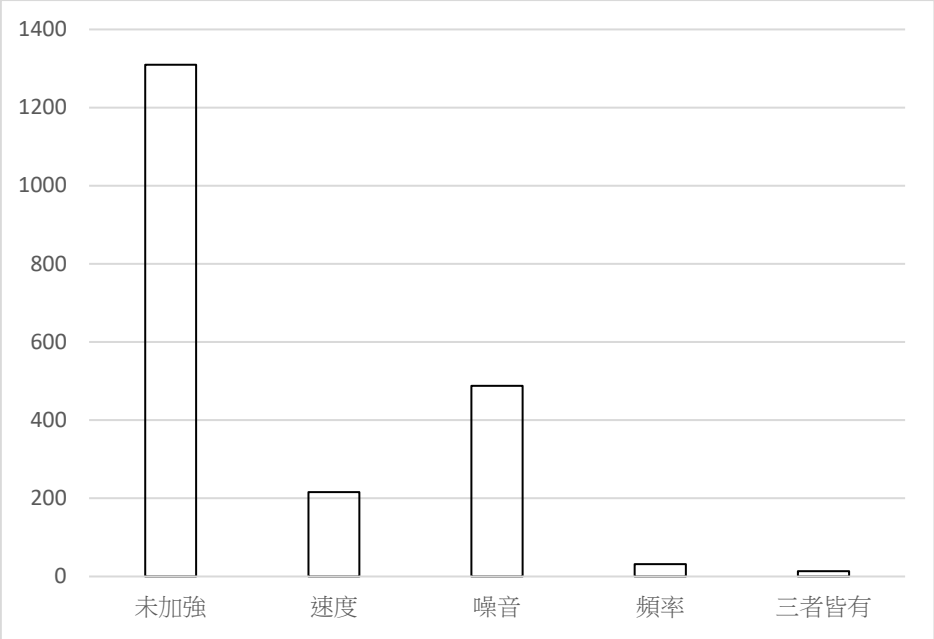
三、數據增強實驗結果

在模型進行語音辨識時，若錄製語音時出現背景雜訊、速度不一等不預期狀況時，辨識出的音素常為空白(辨識不出音素)，因此在這個實驗，我們想要加強訓練集的數據，進而提升模型辨識率。

我們將實驗二的 37 種訓練集進行速度、頻率、噪音的處理。而數據加強前後對比的方式有兩種，分別為音素辨識的空白數和隨機森林、梯度上升和極限梯度提升訓練後的平均準確度比較，以下為我們的實驗結果：

(一)數據加強前後，辨識音素空白的數量比較

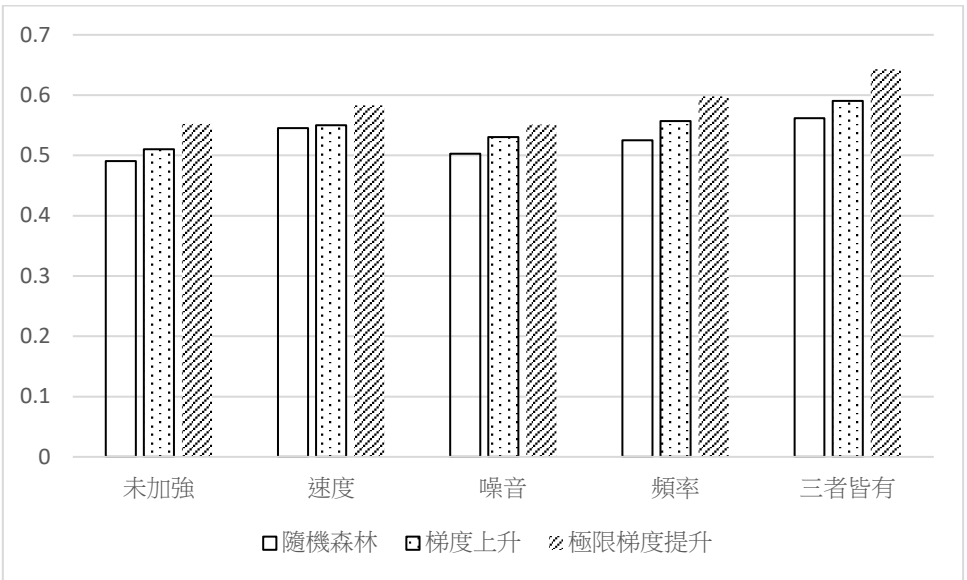
從圖十二可以得出，經過數據加強後，無論是這三種的何種方法，辨識度皆有明顯的提升，而速度增強、加入噪音、頻率增強這三種數據增強的方式皆加入訓練集時的改變最為顯著。



圖十二、數據加強前後的空白辨識音素數量比較

(二)數據加強前後，兩算法的準確度比較

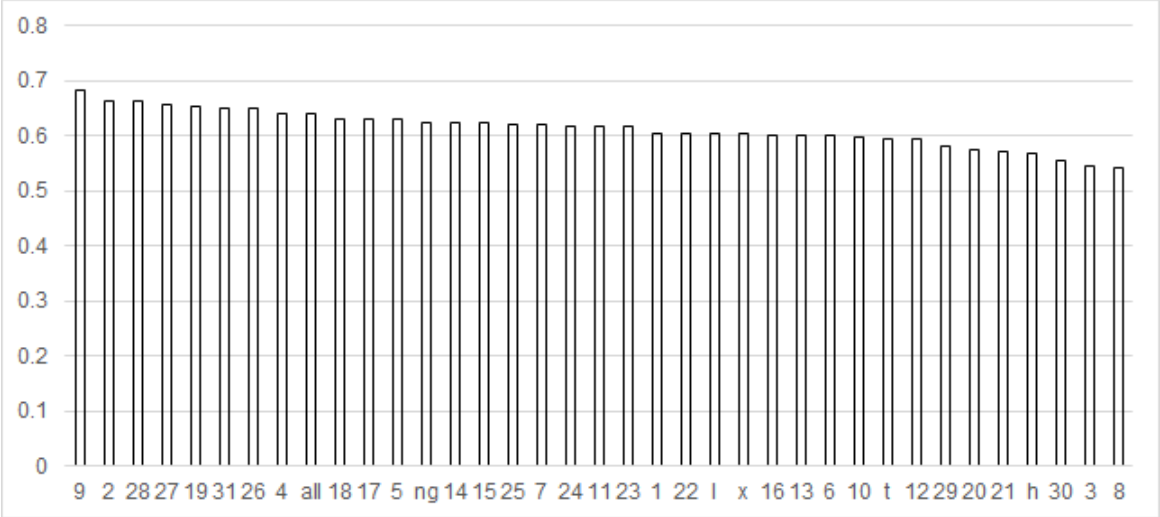
從圖十三可以得出，在所有數據加強資料中，極限梯度上升的平均準確率皆大於隨機森林和梯度上升，而三者皆加入訓練集時的效果最為優秀。歸納實驗結果，我們決定以極限梯度提升作為評分分類器的機器學習算法，並以對於速度、頻率、噪音擴增的數據增強訓練集作為模型訓練集。



圖十三、各數據加強以隨機森林、梯度上升和極限梯度上升準確度比較

四、 使用極限梯度提升判斷模型品質，並選擇最佳模型

透過極限梯度提升對各模型數據進行機器學習，所得到準確率作為對比。從圖十四可以得出，在所有模型數據中，9 號模型準確度最高。因此我們從 37 種模型中，選擇出 9 號模型作為最佳模型，並以此模型計算田野調查收集的語料，產生的數據做為評分分類器的訓練資料，分類方法則是採用極限梯度上升訓練。



圖十四、各模型數據以極限梯度提升訓練所得準確度比較

五、評分分類器的結果

田野調查族語老師的評分結果轉換為 A, B, C，分別表示良、普、劣。以 9 號模型輸出的數據為訓練資料，然後使用極限梯度提升訓練出評分分類器，作為評分系統的依據。訓練結果以 10 折交叉驗證法(10 folder cross-validation)分析，結果如在表四的混淆矩陣(confusion matrix)，279 筆資料中，分類準確率的有 196 筆，分類正確率高達 71%。表五列出準確率、精確率(Precision)、召回率(Recall)、f1-score，A 的精確率為 81%，顯示若評分結果為 A 時，有 0.81 的機率專家評分也是 A。而 C 的召回率是 88%，表示如果專家評分為 C，則有 0.88 的機會本評分系統也會評分為 C。

表四、評分分類器的混淆矩陣

		評分系統分類器計算結果		
專家評分		A(良)116	B(普)89	C(劣)74
	A(良)	89	21	0
	B(普)	21	43	9
	C(劣)	6	25	65

表五、評分分類器的準確率、精確率、召回率、f1-score

	precision	recall	f1-score	support
A	0.81	0.77	0.79	116
B	0.59	0.48	0.53	89
C	0.68	0.88	0.76	74
accuracy			0.71	279

六、評分系統網站

為了推廣我們的評分系統，我們實作了一個的主從式評分系統網站，可在行動裝置上透過網路操作我們的評分分類器。圖十五與圖十六是我們的評分系統網站截圖，在圖十五，進入網站後，會看到選擇欲評分單詞的表單，以及錄製、暫停和停止的按鈕，選擇單字後，按下錄製鍵進行錄製。在圖十六，錄製完成後，可進行 Save to disk(下載)或 Upload(上傳)，若按下了 Upload，則會將聲音檔傳至伺服器，透過伺服器的一系列處理後，會回傳辨識音素和評分結果，並顯示在網路上，進行辨識後輸出辨識音素和評分結果。

圖十五、左圖、進入網站的的介面，右圖、選擇要測試的單字

圖十六、左圖、錄製完畢的結果，右圖、上傳後輸出評分和辨識音素

陸、討論

一、 以田野調查結果與語言學分析太魯閣族語難念的音素

太魯閣族語教學者提出：「l」、「ng」、「x」、「t」、「h」是一般人比較不容易念好的，我們將其分別和沒有上述音標之單詞進行比較，由實驗一與結果得知，非母語學生在此狀況下的評分較低，顯示非母語學生的確不容易學習這五個太魯閣族語發音音素，其中「l」的單字最為明顯。在給予太魯閣族語教學

者的評分中，我們發現非母語者在特別的音標中，即使盡量模仿標準的語音，仍與母語者有區別。

因為與英文的發音不同，使非母語者容易有錯誤的讀音，雖然差距對於非母語者而言並不明顯，但在母語者的日常談話中，是容易造成誤解的。

相反的，在沒有單詞沒有特殊的音標時，非母語者聽過標準音檔後能夠得到較高的分數，顯示了非母語者應能透過聆聽標準音檔來做出好的發音。

二、 如何正確評估最佳模型

使用 Kaldi 語音辨識系統來辨識產生音素與發音品質分數(GOP)，Kaldi 語音辨識系統的訓練資料為原住民委員會族語 E 樂園的太魯閣族語資料，經由調整不同族語資料輸入與設定方式，可以產生各種不同的 Kaldi 辨識模型。不同的辨識模型會產生不同的模型數據。再以字串相似演算法，計算模型輸出的音素。

最後利用線性迴歸、類神經網路、隨機森林法、支援向量機、梯度上升法、極限梯度上升法，比較我們的田野調查結果，來選擇最好的算法和最佳的模型，也就是最接近我們田野調查結果的模型，然後可以產生最佳的模型數據。

三、 數據增強後的影響

由於太魯閣族語語料資源稀少，我們以速度增強、加入噪音、頻率增強這三種數據增強的方式，將原本的訓練集音檔進行處理，並以音素辨識的非空白數、訓練後的準確度比較來判斷數據增強的結果。結果發現三種數據增強方式都使用，在極限梯度上升機器學習方法上面會有最好的效果。

此外，從圖十二和圖十三可以看到，頻率增強的模型和三者皆增強過後模型的結果相差無幾，推測是因為頻率增強將音檔的特徵放大，時長不變的情況下，特徵更加明顯，此舉或許讓訓練模型時辨識的效果變好，這有待日後的驗證。

四、 評分分類器的製作與結果

透過更改不同的訓練集，各模型的準確度在 0.7 到 0.55 之間，標準差約為 3%，模型之間的準確度不同，推測是田野調查所取用的測試集單詞數量太少所造

成的誤差，模型準確度高低和模型訓練集並沒有找出明顯的關聯，主要是因為我們很難解釋模型內部確切的運作方式，只能透過模型輸出來辨別模型的好壞。

五、 評分系統架設對此研究和未來的影響

我們實作的主從式評分系統網站，可在行動裝置上透過網路操作，也有不錯的正確率，已經可以做為簡單教學使用。一般來說，一個九成以上正確率的語音模型，需要上萬個真實語料，而我們此研究的模型，加上數據增強後，僅有上千個語料，而達到七成的正確率，顯示了我們的語料不足。本研究以機器學習方式，儘量提升正確率，透過一些加強語料提升正確率，而達到不錯的效果。期待在我們實作的評分系統網站推廣之後，除了幫助想學習太魯閣族語的人士檢視自己的發音外，也能讓我們大量的蒐集訓練語料，以便加強我們評分系統的強度。

柒、結論

一、總結

(一) 特別的音標「l」、「ng」、「x」、「t」、「h」是非母語較難學習的。

(二) 隨機森林、梯度上升和極限梯度上升是較適合本研究的算法，其中又以極限梯度上升為最佳。

(三) 將訓練音檔進行數據增強時，在辨識率、準確度皆有所提升，其中以同時進行速度增強、加入噪音和頻率增強效果最佳，只進行頻率增強時效果次佳。

(四)以極限梯度上升的算法對各模型數據進行機器學習，得到的模型以加入九號單字(glu)的模型為最佳。

(五) 以 9 號模型輸出的數據為訓練資料，然後使用極限梯度上升訓練出評分分類器，其分類準確度高達 71%，契合我們的研究目的。

二、未來展望

(一)利用線上評分系統蒐集更多語音資料，以擴展我們的訓練資料，進而提

高評分系統的正確率。

(二)進一步探討頻率增強對模型訓練的影響。

(三)嘗試更多機器學習的方式，並混合不同的機器學習模型(hybrid)，以提升分類器的準確度。

捌、參考資料及其他

[1] 本土語言資料網。本土語言使用情況說明。https://mhi.moe.edu.tw/faqList.jsp(2020/11/5)。

[2] 聯合國永續發展目標(SDGs)說明。聯合國永續發展目標 SDGs 的第 11.4 條。檢自 https://www.ait.org.tw/wp-content/uploads/sites/269/un-sdg.pdf (2020/11/5)。

[3] 族語 E 樂園。太魯閣族語字母篇、新九階教材。檢自 http://web.klokah.tw/ (2020/11/29)。

[4] 陳果果、都家宇、那興宇、張俊博(2020)。*AI 辨識-用 Kaldi 實作應用全集*。台北市：深智數位股份有限公司。

[5] Daniel Jurafsky & James H. Martin(2019)。*自然語言處理綜論(第二版)*。北京市：電子工業出版社。

[6] Serkan Kavak(2020/08/24)。*"Kaldi Toolkit, Automatic Speech Recognition and Goodness of Pronunciation"PowerPoint 演示文稿*。東華大學。

[7] 李琳山(2020)。Introduction to Digital Speech Processing 2020 Autumn。檢自 http://speech.ee.ntu.edu.tw/DSP2020Autumn/?fbclid=IwAR0hsa-dU27J0wCyZXV419K34TtCgE5juu69rWrtQjaMmlTvxGhwtMw8JN8 (2020/12/22)。

[8] 吳軍(2016)。*數學之美*。北京市：人民郵電出版社。

[9] 袁梅宇(2015)。*王者歸來 WEKA 機器學習與大數據聖經*。臺北市：佳魁資訊。

[10] 許曜麒、楊明翰、洪孝宗、林奕儒、陳冠宇與陳柏琳(2016/12/2)。*Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection*。

[11] Yoylee_web(2019/01/23)。*資料對齊-編輯距離演算法詳解(Levenshtein distance)*。

[12] Machine Learning in Python, scikit-learning, 2021/1, https://scikit-learn.org/stable/

[13] 維基百科，梯度的提升技術。

[14] Xgboost GitHub project webpage, 2019, https://github.com/dmlc/xgboost

【評語】052501

此作品可協助太魯閣族語的學習，成果具有教育價值。但實驗分為三組各只有 3 人，統計上的意義代表性不高，建議可多增加實驗人數。報告書中圖表的呈現也可多增加一些說明，例如 Y 軸應註明單位等。另外，目前辨識正確率雖有 0.7，建議未來更深入研究後，找出其它較好的方法來提昇正確率。

作品簡報

原音重現—
自動太魯閣族語語音評分系統

組別：高級中等學校組
科別：電腦與資訊學科

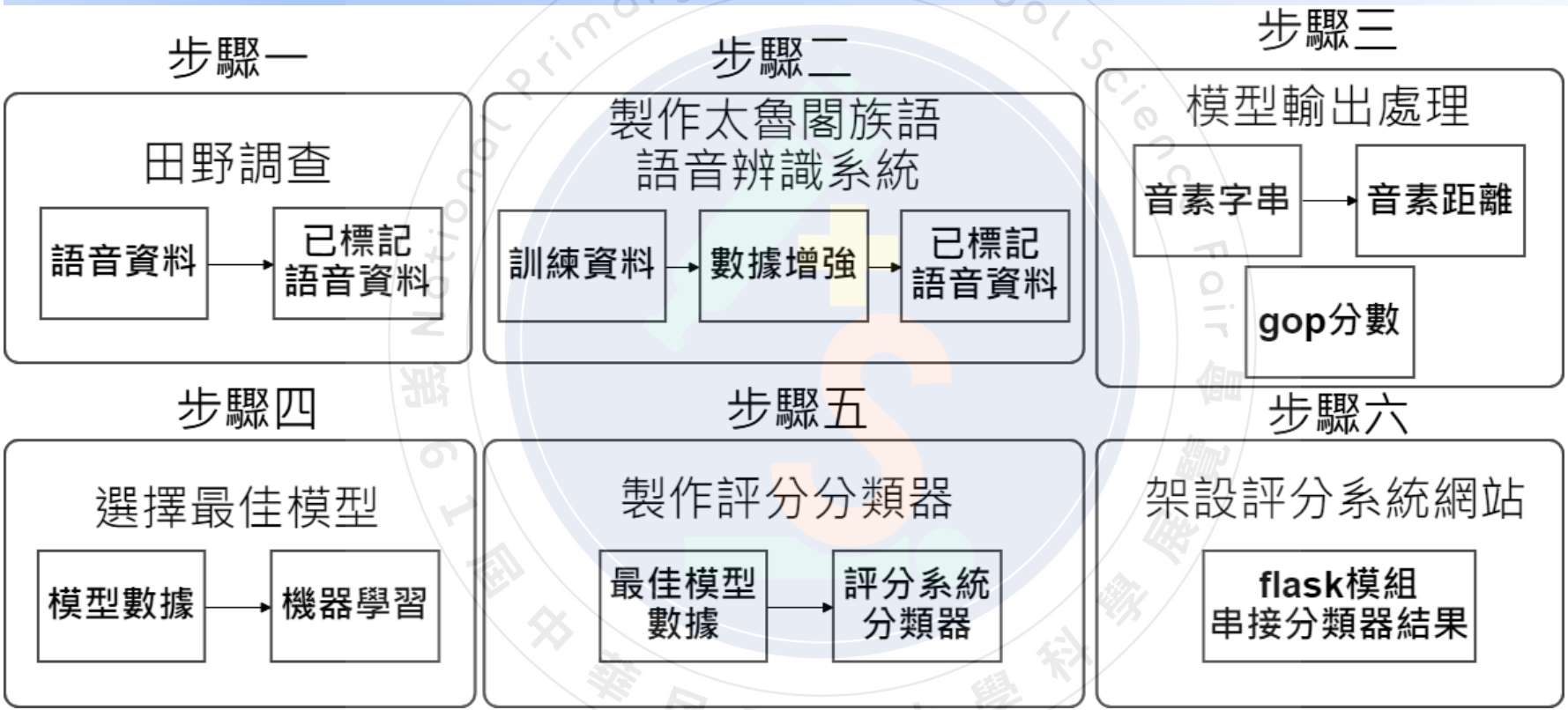
前言

- **研發**在行動裝置操作的自動太魯閣族語語音評分網站，對於族語單詞的發音，在評分等級為A, B, C(良、普、差)時，與族語老師比較，**正確率在七成以上的行動學習工具**。

- 以**田野調查**的經驗與本評分系統，**探討太魯閣族語的特性**，並討論母語學生與非母語學生學習太魯閣族語的差別。

圖一：評分系統網站

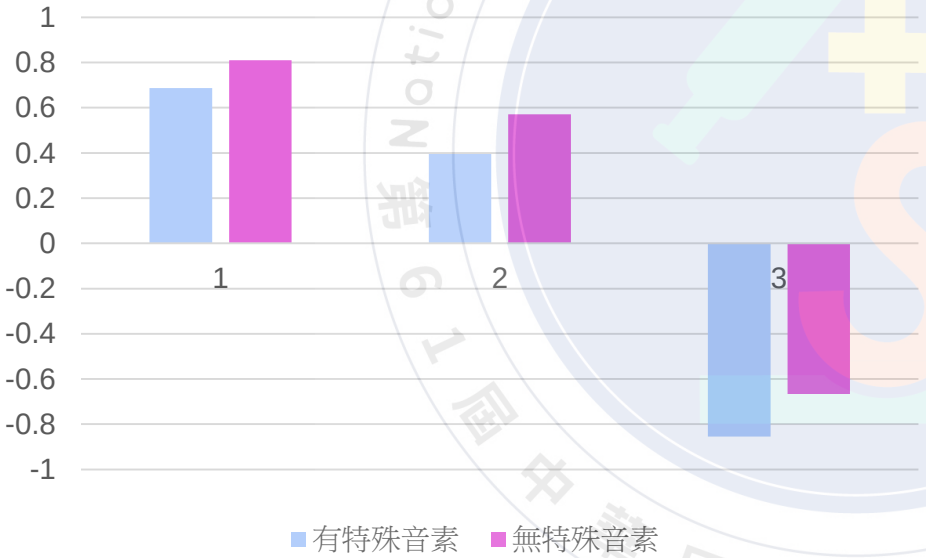
研究方法與過程



圖二：研究流程圖

實驗一：母語和非母語發音太魯閣族語音素的差別

編號	1	2	3
對應的測試者	母語者	聽過標準音檔的非母語者	未聽過標準音檔的非母語者



圖三：有無特殊音素的單字分數比較

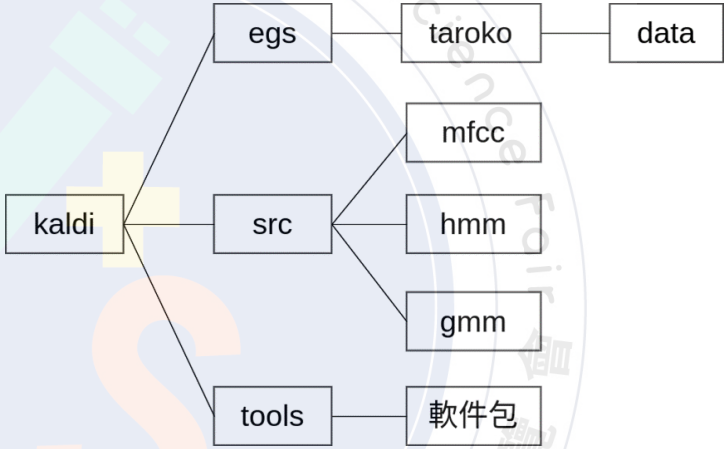
- **無特殊音素**：母語者和非母語者的差別較小
- **有特殊音素**：母語者和非母語者有差別
(l、ng、t、x、h)

田野調查
製作辨識系統
輸出處理
選擇模型
製作評分分類器
架設網站

語音評分系統框架和選擇實驗的單字

表一：訓練集加入的音檔列表

模型	加入訓練集的音檔編號
1-31	對應編號之單字音檔
l	1、2、9、13、15、23、26
ng	1、8、16、19、24
x	23、27、29
t	21、22、26、29、31
h	8、10、19
all	所有單字



圖四：Kaldi系統架構

表二：模型數據屬性內容

Hamming distance	Levenshtein	Gop 分數1	Gop 分數2	…	族語老師評分(輸出)

原始單詞：alang

輸出音素：ana

$$\mathrm{GOP}(u,n) \equiv \frac{1}{T_{u,n}} \log P(q_{u,n}|O_{u,n})$$

田野調查

製作辨識系統

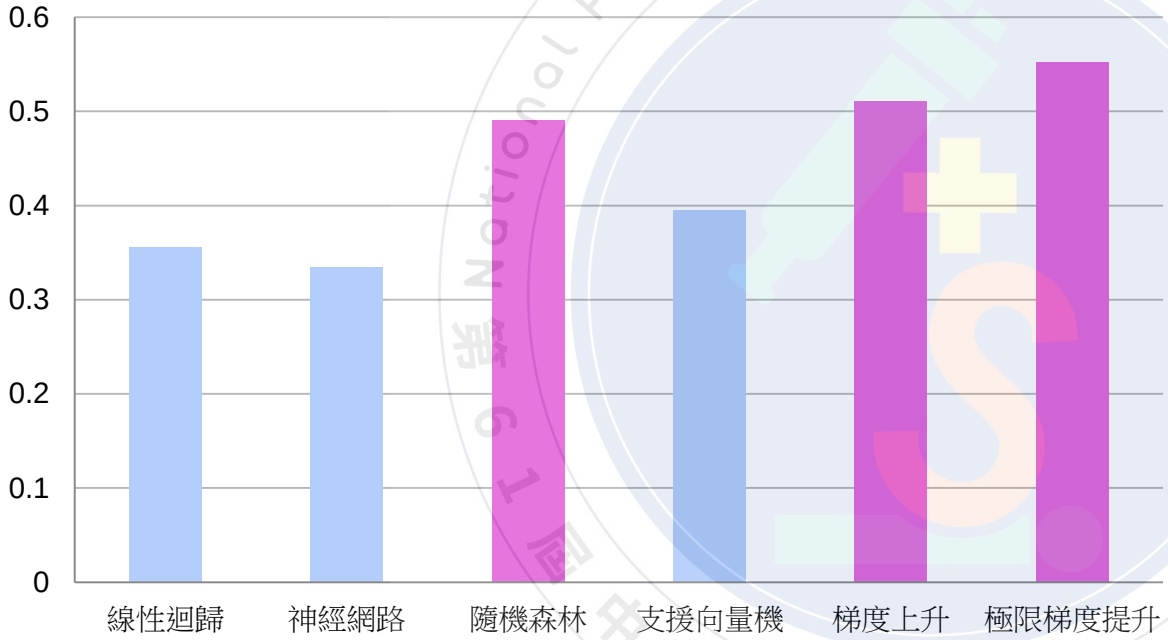
輸出處理

選擇模型

製作評分分類器

架設網站

實驗二：找出最適合的算法



圖五：各算法的準確度比較

下列算法準確度皆趨近 0.5

- 隨機森林
- 梯度上升
- 極限梯度提升

田野調查

製作辨識系統

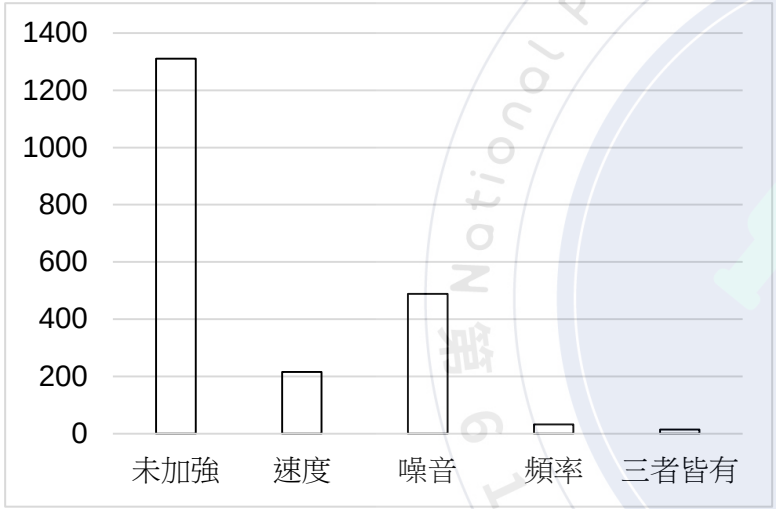
輸出處理

選擇模型

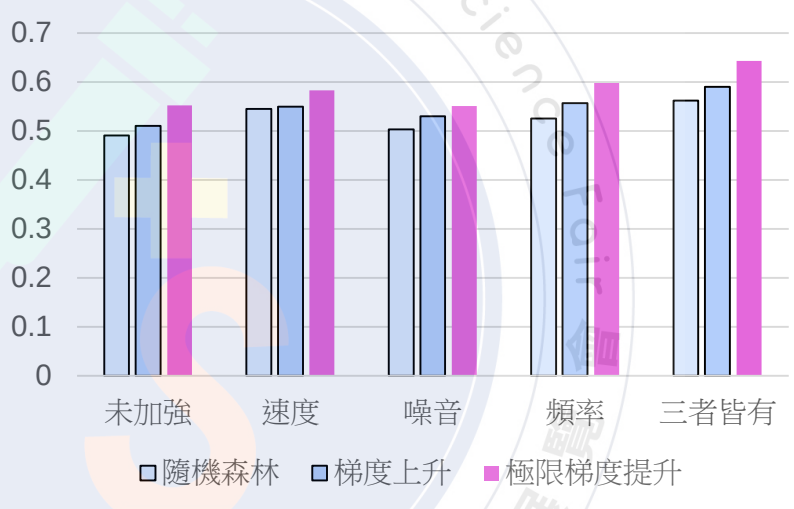
製作評分分類器

架設網站

實驗三：數據增強實驗結果



圖六：數據加強前後的空白辨識音素數量比較



圖七：數據加強後三種算法準確度之比較

數據增強挑選結果

- 三種數據增強都使用

最佳算法挑選結果

- 極限梯度提升

田野調查

製作辨識系統

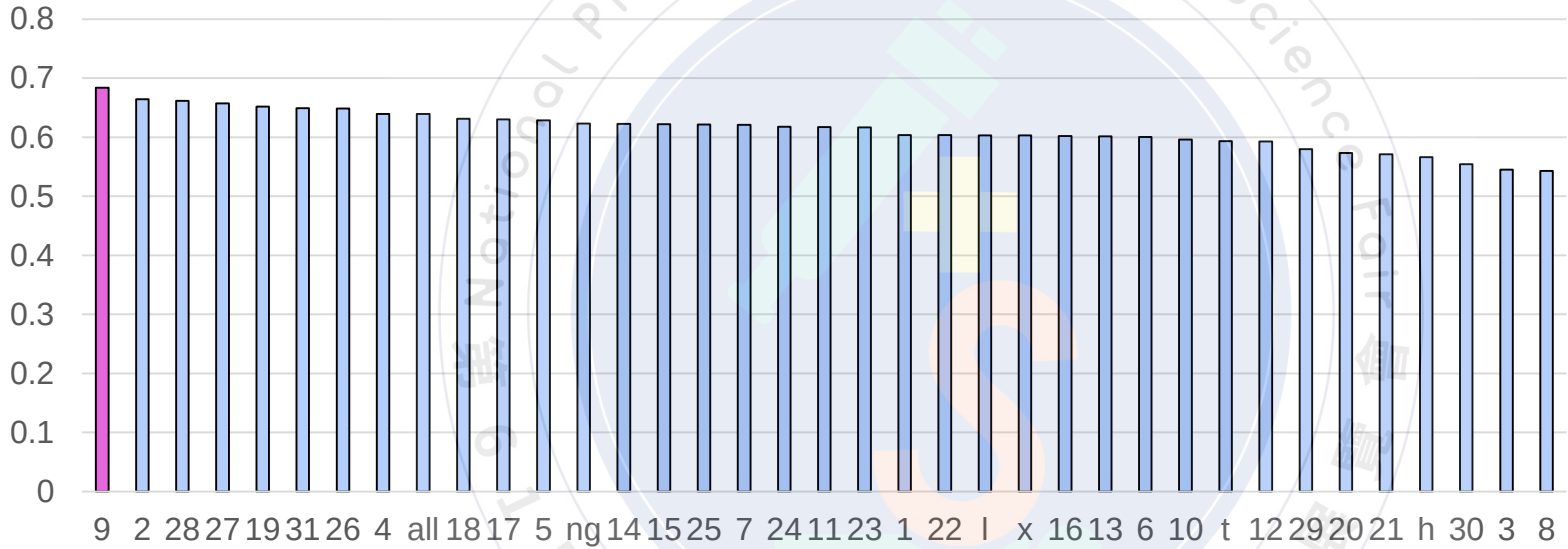
輸出處理

選擇模型

製作評分分類器

架設網站

實驗四：評分分類器使用的最佳模型



圖八：各模型數據以極限梯度提升訓練所得之準確度比較

- **9號模型**(基本訓練集+glu單字的標準音檔、三種數據增強、極限梯度提升)效果最佳

實驗五：評分分類器的結果

表三：評分分類器的混淆矩陣

		評分系統分類器計算結果		
		A(良) 116	B(普) 89	C(差) 74
專家評分	A(良)	89	21	0
	B(普)	21	43	9
	C(差)	6	25	65

表四：評分分類器的
準確率、精確率、召回率、f1-score

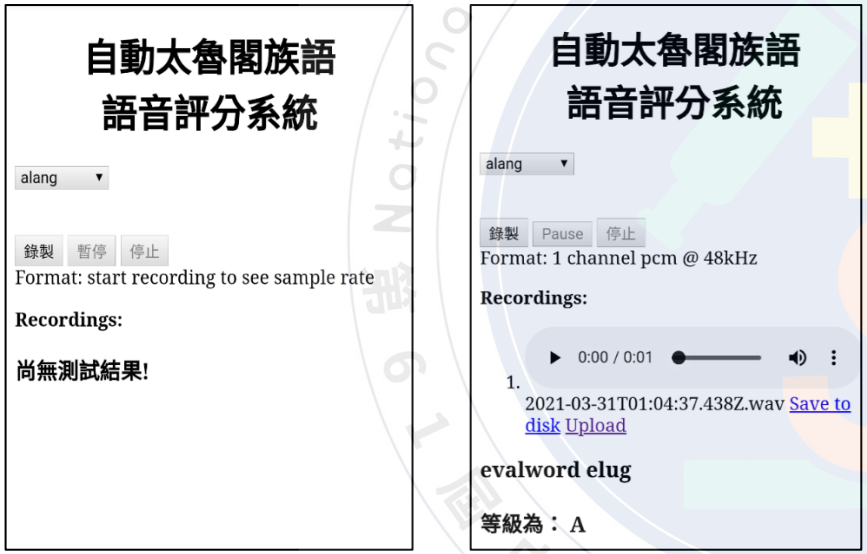
	precision	recall	f1-score	support
A(良)	0.81	0.77	0.79	116
B(普)	0.59	0.48	0.53	89
C(差)	0.68	0.88	0.76	74
accuracy			0.71	279

- 專家評分共 279 筆樣本
- 機器分類結果共 197 筆分類正確
- **分類正確率為 71%**

- A的精確率為 **81%**
- C的召回率是 **88%**
- B的分類效果較差

田野調查
製作辨識系統
輸出處理
選擇模型
製作評分分類器
架設網站

實驗六：評分系統網站



圖九：評分系統網站

此系統主要目標

- 幫助想學習太魯閣族語的人士檢視自己的發音
- 透過系統可廣度地蒐集訓練集

田野調查

製作辨識系統

輸出處理

選擇模型

製作評分分類器

架設網站

結論與未來展望

- 特別的**音標**「l」、「ng」、「x」、「t」、「h」是非母語較難學習的。
- 隨機森林、梯度上升和極限梯度上升是較適合本研究的算法，其中又以**極限梯度上升**為最佳。
- 將訓練音檔進行數據增強時，在辨識率、準確度皆有所提升，其中以**同時進行速度增強、加入噪音和頻率增強效果最佳**。
- 以極限梯度上升的算法對各模型數據進行機器學習，得到的模型以**加入九號單字(glu)的模型**為最佳。
- 以9號模型輸出的數據為訓練資料，然後使用極限梯度上升訓練出評分分類器，其**分類準確度高達71%**，契合我們的研究目的。
- **未來**我們將利用線上評分系統**蒐集更多語音資料**，並且**嘗試更多機器學習**的方式，以**提升分類器的準確度**。

參考資料

[1]族語E樂園。太魯閣族語字母篇、新九階教材。

[2]陳果果、都家宇、那興宇、張俊博(2020)。*AI語音辨識-用Kaldi實作應用全集*。台北市：深智數位股份有限公司。

[3] Serkan Kavak(2020/08/24)。*"Kaldi Toolkit, Automatic Speech Recognition and Goodness of Pronunciation"PowerPoint演示文稿*。東華大學。

[4] Yoylee_web(2019/01/23)。*資料對齊-編輯距離演算法詳解(Levenshtein distance)*。

[5] Machine Learning in Python, scikit-learning, 2021/1

[6] Xgboost GitHub project webpage, 2019