淡江大學學校財團法人淡江大學
資訊工程學系

日文語音評分系統

專題成果報告書

組員

資工三 A 410410103 林聿朔
資工三 A 410411036 施吉益
資工三 A 410411788 陳緯榛
資工三 A 410410707 吳天宇
資工三 A 410410137 蘇柏修

目錄

動機與目的 - 1 -

研究背景 - 1 -

研究動機 - 1 -

研究目的 - 1 -

相關研究 - 2 -

降噪處理 - 2 -

語音辨識 - 2 -

語音特徵提取 - 4 -

BLSTM - 5 -

評分模型 - 5 -

網頁處理 - 5 -

評估指標 - 6 -

模型建構 - 7 -

訓練流程概述 - 7 -

資料集介紹 - 8 -

音檔前處理 - 10 -

SSL 模型 ASR 微調 - 10 -

評分模型的建立與訓練 - 11 -

系統呈現 - 14 -

整體流程介紹 - 14 -

Client Side - 15 -

Server Side - 17 -

結論 - 17 -

展示海報 - 19 -

參考文獻 - 20 -

動機與目的

研究背景

隨著全球化的加速發展，日語作為一門重要的國際語言，其學習需求在全球範圍內顯著增長。日語的流利程度，特別是發音的準確性，對於學習者在商業、文化及教育等多方面的交流具有至關重要的作用。然而，在日語教育實踐中，精確評估學生的發音並提供即時反饋仍然是一項挑戰。目前，許多學校和教育機構缺乏有效的工具來支援學生在日語發音上的自我改進，同時也增加了教師在語音評估過程中的工作負擔。

研究動機

本研究的動機源於解決學校在提供精確的日語發音評估和快速反饋方面的現有空缺。我們觀察到教師在評估學生發音標準時面臨著時間上巨大的壓力，且學生缺少足夠的資源來指導他們獨立改進發音，這促使我們尋求一種新的解決方案。此外，傳統的發音評估方法通常依賴於耗時的特徵工程和人工評估，這不僅效率低下，而且往往缺乏客觀性和一致性。

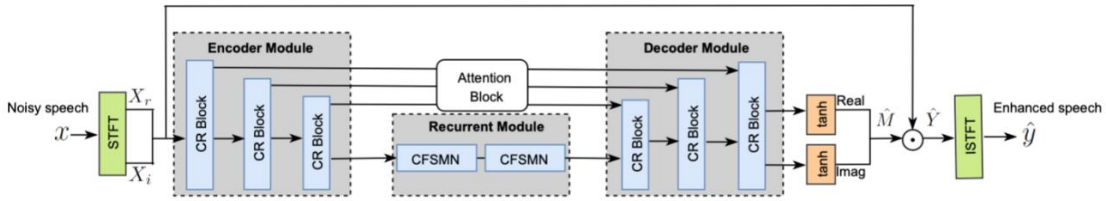
研究目的

本研究旨在開發一套自動發音評估系統，特別針對日語學習者，以支持我們學校的教育需求。此系統基於自監督學習(SSL)模型，能夠從原始語音檔案中直接學習豐富的語音特徵，從而免去傳統語音處理中對特徵工程的依賴。此方法不僅旨在提高發音評估的精確度，而且簡化了評估過程，使反饋更加即時和準確。系統的開發將為教師提供一個強有力的教學輔助工具，使他們能夠更有效地進行教學，同時也為學生提供一個隨時可用的練習工具，以期在課餘時間更有效地提升其日語發音能力。透過這種方式，本研究期望為日語教育領域帶來創新的教學方法和技術解決方案。

相關研究

降噪處理

FRCRN 語音降噪模型[1]是基於頻率循環 CRN (FRCRN) 新框架所發展出來的(圖一)。此框架是在卷積編-解碼(Convolutional Encoder-Decoder)架構的基礎上，透過進一步增加循環層所獲得的卷積循環編-解碼(Convolutional Recurrent Encoder-Decoder)新型架構，可以明顯改善卷積核的視野局限性，提升降噪模型對頻率維度的特徵表達，尤其是在頻率長距離相關性表達上獲得提升，可以在消除噪音的同時，對語音進行更針對性的辨識和保護。模型輸入與輸出皆為 16kHz 取樣率單聲道語音時域波形訊號，輸入訊號可由單聲道麥克風直接進行錄製，輸出為雜訊抑制後的語音音訊訊號。模型輸入訊號透過 STFT 變換轉換成複數頻譜特徵作為輸入，並採用 Complex FSMN 在頻域上進行關聯性處理和在時序特徵上進行長序處理，預測中間輸出目標 Complex ideal ratio mask, 然後使用預測的 mask 和輸入頻譜相乘後得到增強後的頻譜，最後透過 STFT 逆變換得到增強後語音波形訊號。

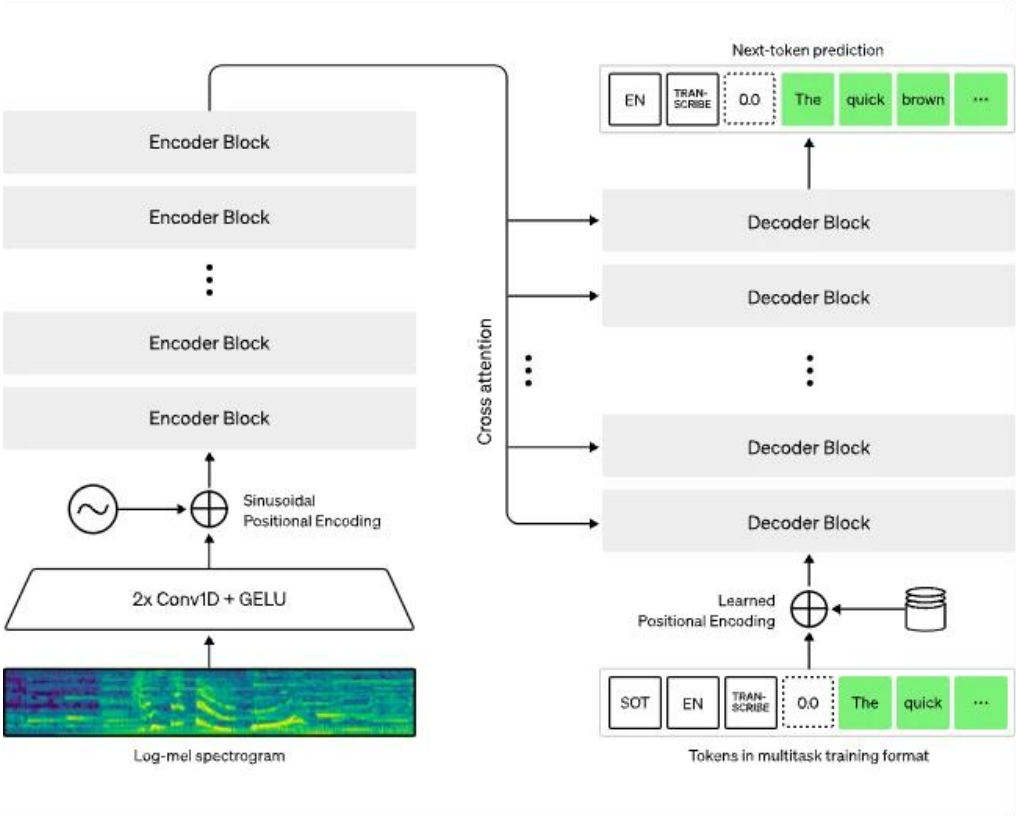


圖一、FRCRN 神經網路結構[1]

語音辨識

Whisper：

Whisper[2]是採用 Sequence to Sequence model 訓練，將輸入的音檔分成每 30 秒一個的 chunks，然後將每一段 chunk 轉成梅爾頻譜(Log-Mel spectrogram)，再經過兩個一維卷積層(Conv1D)和 Gaussian Error Linear Units(GELU)啟動函數進行訓練，再經由正弦波位置編碼(Sinusoidal position encoder) 加入關於 token 位置的資訊，接著進入 Transformer Encoder，然後透過交叉注意力機制(cross attention)，從不同序列中獲取資訊，最後由 Transformer Decoder 輸出。圖二為 Whipser 的架構示意圖。



圖二、Whisper 架構[2]

ESPnet：

ESPnet 語音辨識技術是一種基於深度學習的語音辨識技術[3]。它結合了連接主義時序分類 (Connectionist Temporal Classification, CTC) 和基於注意力的 ASR 編碼器-解碼器網路，透過訓練大規模的語音資料集，實現了高精度的語音辨識。該技術的核心是使用深度學習演算法對語音訊號進行處理，從中提取有效的特徵，進而進行分類和識別。

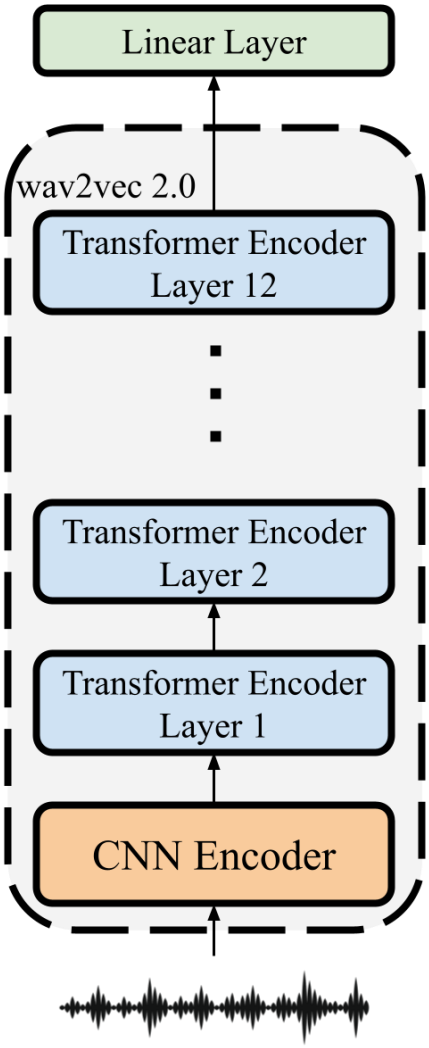
ReazonSpeech_v2：

這個模型[4]採用了從 Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition 改進的 Conformer 架構，並採用了 Subword-based RNN-T 模型。編碼器使用 Longformer attention，並且具有單一全域標記。解碼器擁有由 SentencePiece unigram tokenizer 建構的 3000 個 token 的詞彙空間。且按照 Noam annealing schedule 使用 AdamW optimizer 對模型進行訓練。

語音特徵提取

在 Ankita Pasad 等人於 2021 年提出對自監督學習(Self-Supervised Learning, SSL)模型的分析研究中，逐層分析了 wav2vec 2.0 模型所學到的特徵[5]。發現 wav2vec 2.0 會遵循著類似自編碼器的行為所學習的特徵隨著層數的深入，由基礎的聲學特徵到複雜的語義信息，並在最後幾層會傾向重構出輸入的語音。

在模型的最初幾層主要由卷積層(CNN)所組成的，專注於捕捉語音中的原始聲學信號的基本屬性，例如音高、音量等。這些層與 mel 頻譜圖特徵高度相關，表明模型在這一階段能有效學習到與傳統聲學特徵相似的表徵。隨著模型層次的加深，wav2vec 2.0 開始處理更高階的語音特徵。於 wav2vec 2.0 的中間層，能夠把語音信號識別具體語音單元，如音素，這是將語音信號轉換為可理解文本的關鍵步驟。進一步深入到模型的更深層，單詞識別和語義理解成為主要學習的特徵。這些層不僅能識別出單詞，還開始將單詞與其含義連結起來，這一能力透過與 GloVe 等文本基礎詞嵌入的關聯性分析得到驗證。在研究中對 wav2vec2 模型進行自動語音辨識(Automatic Speech Recognition, ASR)任務的微調，這顯著改變了模型最末幾層的表現，原本是類似自編碼器的行為，即重建出輸入的語音，在經過微調後，最末幾層變得專注於與 ASR 直接相關的資訊，如更精確的字詞辨別等。wac2vec 2.0 模型架構如下圖三。



圖三、wav2vec 2.0 base 模型架構

BLSTM

在 2015 年，Yu 等人的研究方法對自動發音評估系統提出了新的架構以優化過往的評分系統的性能[6]，這一架構中引入雙向長短期記憶神經網路(Bidirectional Long Short-Term Memory, BLSTM)，透過 BLSTM 的雙向結構使發音評分系統能夠同時考慮過去和未來的上下文訊息，更有效的捕捉時間序列的特徵，如韻律和梅爾頻譜倒頻係數(MFCCs)。這增強了評分系統對語音動態的理解，還能夠捕捉到更多細節和語音的變化特徵。在這項研究中，結合了 BLSTM 從韻律和 MFCCs 中提取的特徵和傳統的時間聚合特徵，與 MLP 模型一起使用時，在語音評分的準確性和與人類評分者的相關性上表現最佳。另外研究也提出了評分系統在只依靠 BLSTM 所提供的特徵而不依賴對 ASR 轉錄文本的分析下，也就是只捕捉時間序列數據的高級特徵，能與僅使用時間聚合特徵的模型達到可比的性能。

評分模型

在 Kim 等人提出的自動發音評估的方法[7]，透過應用自監督學習(SSL)模型，如 wav2vec 2.0 和 Hidden-Unit BERT (HuBERT)，提出了一種新的自動發音評估方法。這些模型首先透過連接主義時序分類(CTC)對先前訓練好的 SSL 模型進行微調，以適應非母語學習者的發音特點。進一步從 Transformer layer 中提取分層的特徵，搭配相對應的文本輸入到 BLSTM 模型中預測發音得分。這種利用 SSL 模型自動學習的語音特徵和 BLSTM 的動態時間序列處理能力的架構，相比於傳統的發音評分方法和基線模型，Kim 等人提出的方法在與專家標注分數的相關性方面(使用皮爾森相關係數衡量)表現出更好的結果。

網頁處理

HTML 與 CSS：

HTML 是網站的結構和內容，而 CSS 是負責設計網頁格式和佈局[8]。我們利用了 HTML 來組織我們網站的內容，包括標題大小、段落格式、網站內容的排版之類。而 CSS 我們用來選擇網站內呈現的樣式，包括背景顏色、文字大小、字形顏色，利用 HTML 和 CSS 是網站開發的兩大支柱，它們相輔相成，共同建構了優秀的網頁體驗。 透過靈活運用這兩種技術，我們可以創造出功能豐富、外觀精美的網站，為使用者提供更好的瀏覽空間。

JavaScript ：

JavaScript 是一種高階、動態、解釋型的程式語言[9]，主要用於在網頁上實現互動式和動態的功能。 作為前端開發的核心技術之一，JavaScript 可以讓網頁呈現更生動、互動性更強的使用者體驗。我們利用 JavaScript 進行錄音和輸出音訊文件，並設計了一些按鈕與使用者互動。 根據使用者的操作結果，網站的呈現會有所不同。這種互動式設計增強了使用者體驗，使用戶能夠更直觀地與網站互動。 JavaScript 的靈活性和強大功能使得我們能夠實現各種互動功能，從而為使用者提供更豐富的使用體驗。

Flask：

Flask 是一個基於 Python 程式語言的輕量級 Web 應用框架[10]，它簡潔而靈活，適用於建立各種類型的 Web 應用程式， Flask 提供了豐富的功能和擴展，使開發者能

夠快速地建立、測試和部署 Web 應用，具有靈活的架構，允許開發者根據專案需求選擇合適的擴充功能和函式庫。 Flask 非常適合建構各種類型的網路應用，從簡單的靜態網頁到複雜的 API 服務都能勝任。我們利用此技術來達成架設網站的目的。

評估指標

單字錯誤率(Word Error Rate, WER)：

單字錯誤率是自動語音識別(ASR)系統中常用的性能評估指標。WER 衡量的是在自動語音識別(ASR)過程中生成的文本與標準答案文本之間的差異。計算方法是將錯誤的單字數除以標準文本中的總單字數。這個指標反映了系統在轉錄單個詞彙時的準確性。

字元錯誤率(Character Error Rate, CER)：

字元錯誤率與單字錯誤率類似，不過它是基於單個字元而非整個單詞。CER 計算的是轉錄文本與標準答案文本之間在字元層面的錯誤數除以參考文本的總字元數。反映了系統在轉錄單個字元時的準確性。

準確率(Accuracy)：

準確率是評估常用的指標，用於衡量模型預測結果的總體正確性。它計算的是所有正確預測(True Positive 和 False Positive)與總樣本數的比例。這個指標提供了一個直觀的方法來判斷系統在所有預測中的總體正確率，Accuracy 計算為公式(1)。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

精確率(Precision)：

精確率關注於模型在預測為正類時的準確性。它是計算從所有正向預測(True Positive 和 False Positive)中確實是正類(True Positive)的比例。精確率特別重要於評估結果的可靠性，避免因過多的錯誤正向預測(False Positive)影響判斷。Precision 計算方式為公式(2)。

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

召回率(Recall)：

召回率專注於模型預測出實際正類(True Positive)的能力。它計算的是在所有真實正類(True Positive 和 False Positive)中，被模型正確識別為正類的比例。高召回率表示系統能夠有效識別大部分真實的正類案例。Recall 計算方式為公式(3)。

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score：

F1-Score 是精確率和召回率的調和平均(Harmonic Mean)，提供了一個統一的衡量標準來評價精確率和召回率的平衡。當需要同時考慮精確率和召回率，而且它們同等重要時，F1-Score 的公式計算如下公式(4)。

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

特異性(Specificity)：

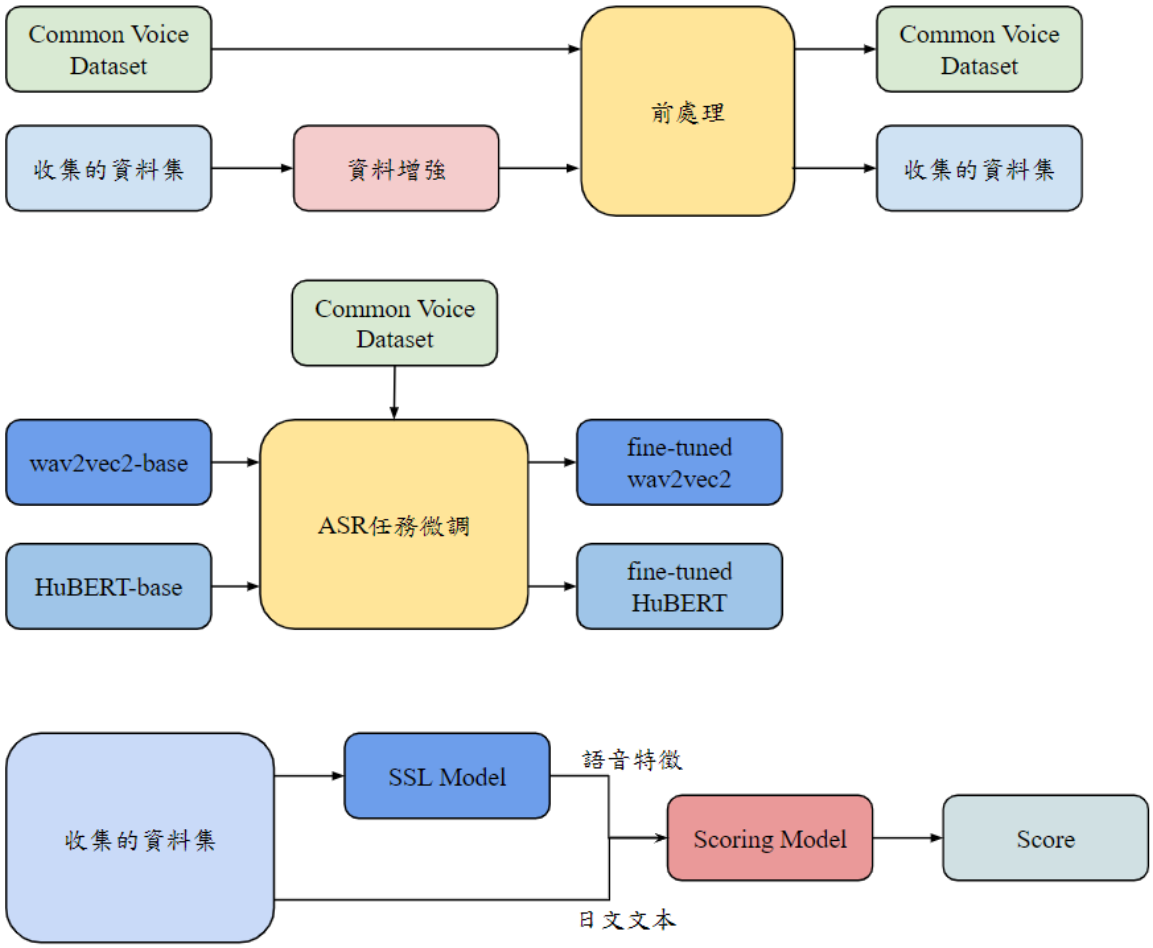
特異性是在所有實際為負類(True Negative 和 False Positive)的樣本中，模型能夠正確識別出負類的比例。它反映了模型識別出真實負例(True Negative)的能力，高特異性也代表減少了錯判真實負類為正向(False Positive)的發生。Specificity 計算方式為公式(5)。

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{5}$$

模型建構

訓練流程概述

第一步為了處理收集自日文系資料集的標籤不平衡問題，我們進行了資料增強，透過調整音高和語速等方法來增加不正確標籤的音檔多樣性，以平衡發音正確與不正確的比例。接下來，對音檔進行前處理，確保它們符合自監督學習(SSL)模型的輸入需求。包括透過降噪和重取樣技術來處理音檔，使其達到適合 SSL 模型輸入的格式。使用 Common Voice Dataset 對 wav2vec2 和 Hidden-Unit BERT (HuBERT) 模型透過適應 ASR 任務，使 SSL 模型能更精細地捕捉語音裡的特徵。之後，將微調過的 wav2vec 2.0 和 HuBERT 模型用作於語音特徵提取器，把提取的語音特徵及對應的文本輸入到評分模型中進行訓練，最後，我們比較了不同 SSL 模型的特徵提取能力與實際分數標籤的相關性，選擇表現最佳的模型組合作為後續系統使用的評分模型。下圖四為資料處理到模型訓練的整體流程。



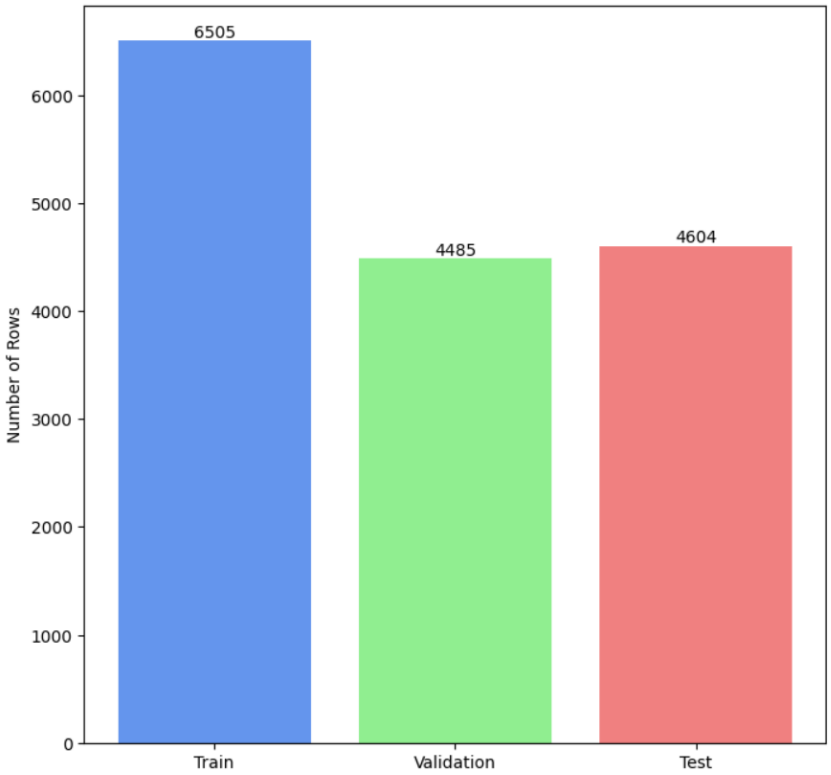
圖四、整體訓練流程

資料集介紹

我們的模型建構會使用到兩個資料集，分別是 Common Voice Dataset 和從日文系收集而來的資料集，以下會分別介紹：

Common Voice Dataset

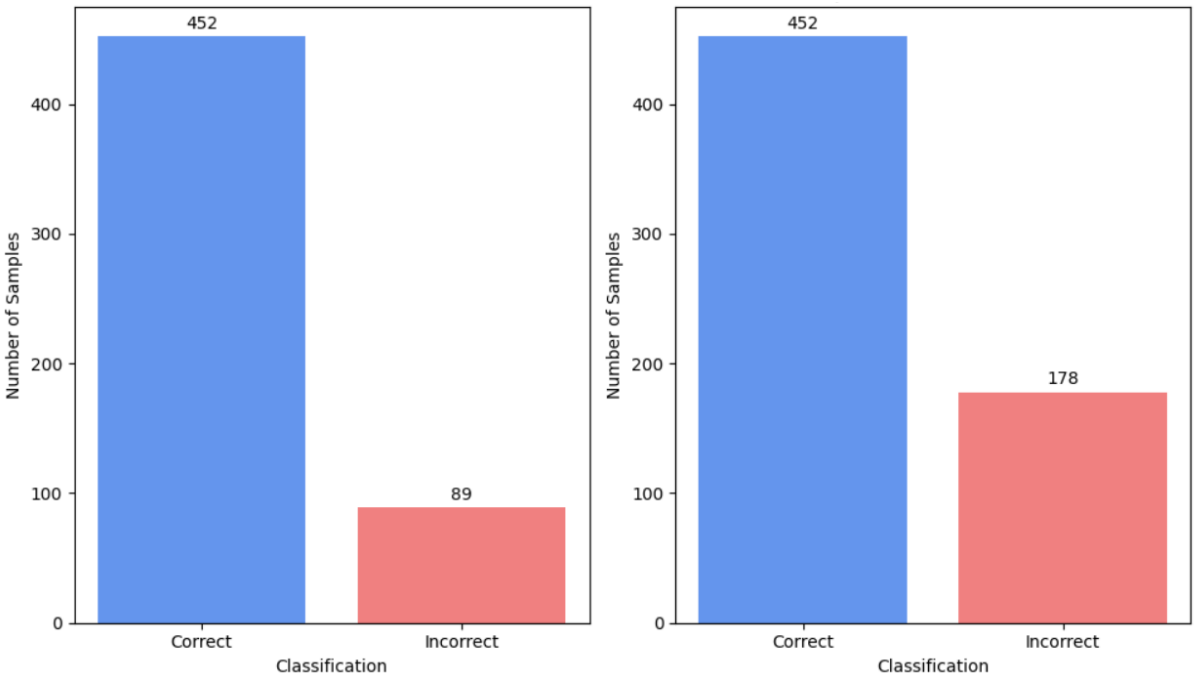
這是於 Mozilla 創建的多語言資料集，由世界各地志工貢獻的錄音所組成，資料集由 MP3 音檔和對應的文字檔案組成。我們選用其中的日語資料集作為 SSL 模型對 ASR 任務的微調使用。我們分別選用了 Common Voice Dataset 中的 train、validation 和 test，如圖五所示。



圖五、Common Voice Dataset 切分集

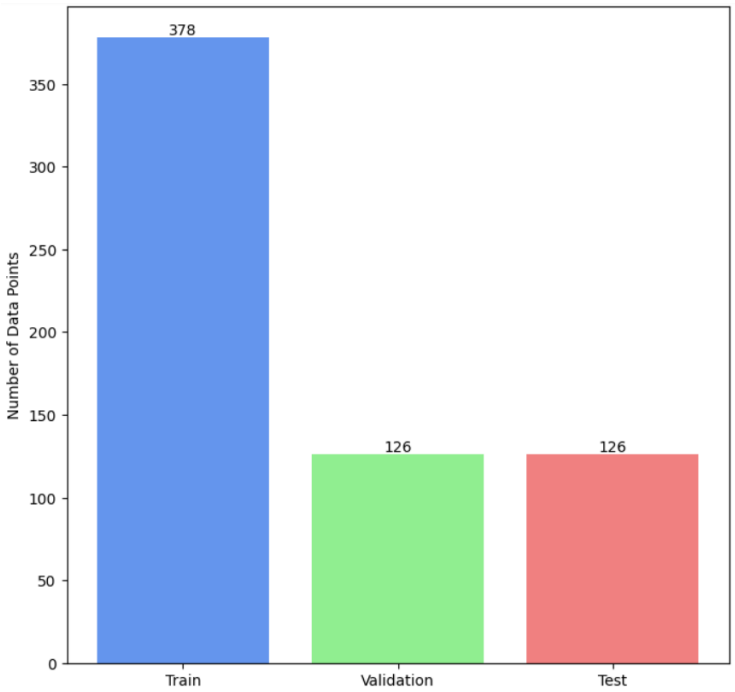
自日文系收集的資料集

我們使用此資料集當作評分模型的訓練，資料集由 MP3 音檔、對應的文字以及日文系教授對錄音發音的評分所組成。針對從日文系收集的資料集中的標籤不平衡問題，其中發音標記為標準的與標記為不標準的比例為 5：1，我們對於標記為非標準發音的語音採取了資料增強的方式來改善訓練資料的多樣性和均衡性。進行了音高和語速的隨機調整，調整幅度在正負 0.3 倍的範圍內。以下圖六為對自日文系收集的資料集做資料增量後發音正確和不正確比例的表示。



圖六、日文系收集的資料集總標籤比例(左圖為增量前，右圖為增量後)

隨後我們將從日文系收集到的資料集以 6：2：2 切分為 train、validation、test 三個子資料集以便後續評分模型訓練，如下圖七所示。



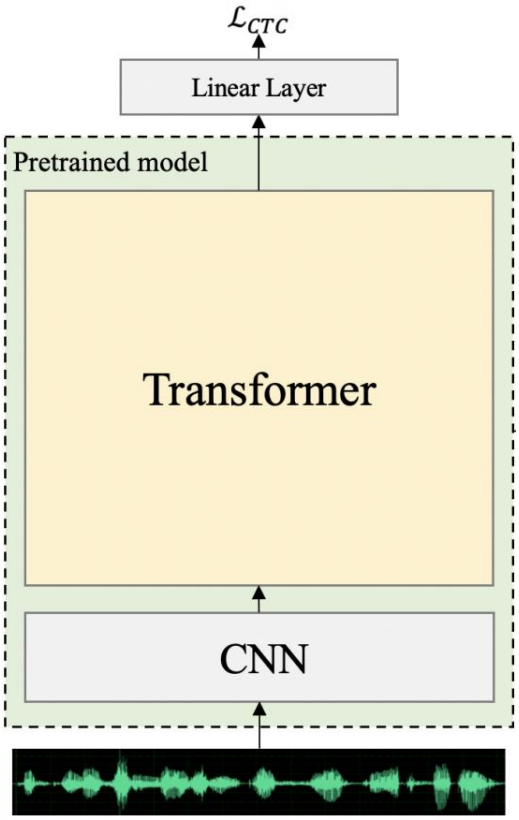
圖七、從日文系收集的資料集分割比例

音檔前處理

在自動發音評分系統和後續的語音評分模型中，音檔前處理是確保模型精確性的首步驟。為了降低背景噪音及雜訊對發音評估的干擾，本系統採用了 FRCRN 模型進行高效的降噪處理。此模型專為從複雜背景中提取清晰語音訊號而設計，能夠顯著提高語音資料的品質。為了配合自監督學習(SSL)模型的需求，所有處理過的音檔均進行了重取樣，以統一格式至 16000 HZ 的取樣率。這有助於模型更有效地學習和提取語音特徵。

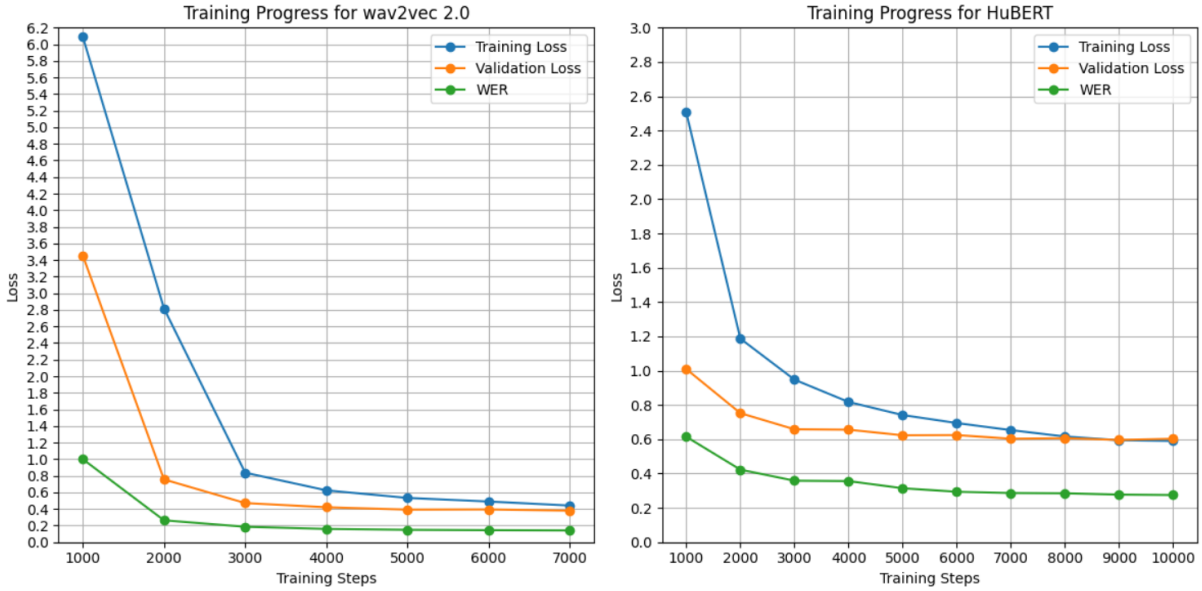
SSL 模型 ASR 微調

為了有效提取日語語音的特性，特地選擇專門以日文語音資料集訓練的 SSL 模型，同時考慮了系統運作的回應速度，我們挑選 wav2vec 2.0 和 HuBERT 的基礎版本，對兩個模型透過 ASR 任務進行微調，期待能更有效地捕捉語音中的細節特徵[5]。為了使 SSL 模型—即 wav2vec2 Base 與 HuBERT Base—能適應 ASR 任務，我們對模型結構上進行更改，額外加入了一層線性層(Linear Layer)以便模型能將輸出映射到日文平假名上，如下圖，為加入線性層後的 SSL 模型架構。隨後使用連接主義時序分類(CTC)損失函數對先前訓練好的 SSL 模型進行微調。CTC 透過動態對齊輸入語音與預期輸出文本，從而有效地訓練模型識別日語發音和相關文本。最後根據計算出的損失值來修正 SSL 模型參數，最後以詞錯誤率(WER)與字元錯誤率(CER) 來評估模型在測試資料集上 ASR 任務的性能。下圖為 SSL 模型為適應 ASR 任務所變更的架構。



圖八、SSL 模型的 ASR 架構[7]

在對 wav2vec 2.0 和 HuBERT 模型架構進行調整後，我們利用 Common Voice Dataset 對這些模型進行自動語音識別(ASR)任務的訓練。使模型打破在最後幾層會重建輸入語音的行為，從而更有效地捕捉語音特徵。以下是 wav2vec 2.0 和 HuBERT 在訓練過程中的表現。



圖九、wav2vec2、HuBERT 的 ASR 訓練過程

最後對 wav2vec 2.0 和 HuBERT 兩個模型在 Common Voice Dataset 上進行了最終的性能評估。這些模型的表現如下表一中：

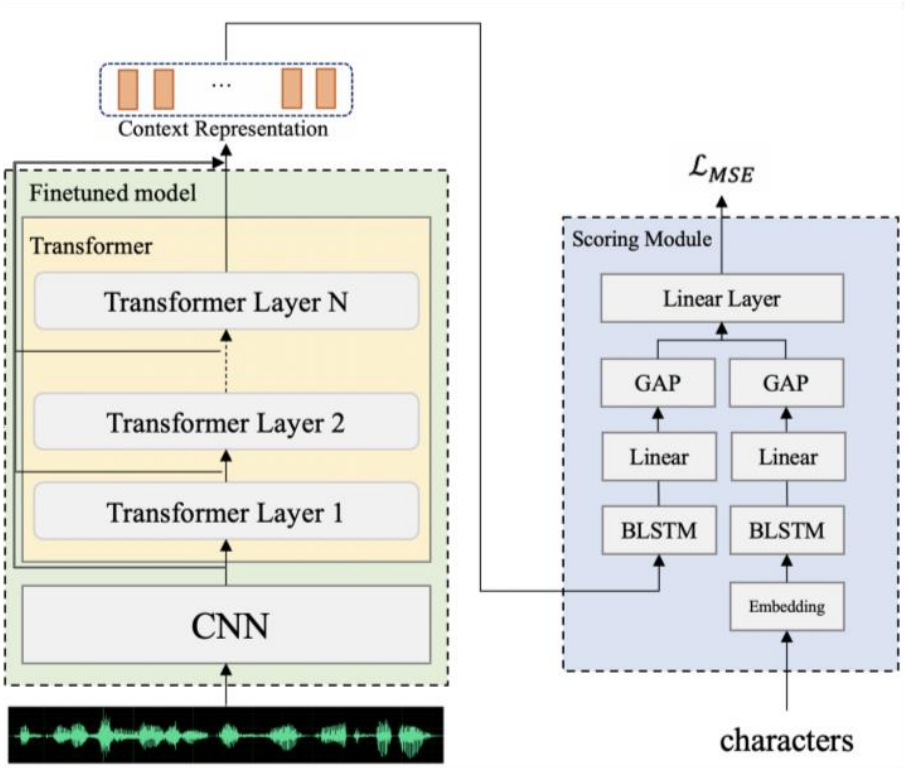
表一、模型在 Common Voice Dataset 測試集上的評估表現

Model	**CER**	**WER**
wav2vec2 Base	6.463%	14.177%
HuBERT Base	11.700%	27.512%

評分模型的建立與訓練

為了有效地評估日語發音的準確性，我們的評分模型是設計來接收由 SSL 模型提取的語音特徵及發音文本的嵌入向量[7]。採用雙向長短期記憶神經網路(BLSTM)結合線性層(Linear Layer)和 Mean Gap Layer 作為核心架構，以捕捉語音資料中的時間序列特徵和整體語音特性。

BLSTM 層是評分模型中的核心，從兩個方向—向前和向後—學習資料中的上下文相關性。這種雙向學習使模型不僅能理解語音片段本身的特性，還能把握前後語境中的變化，從而更精確地評估發音。線性層則在 BLSTM 處理過的複雜數據上進行整合，將高維度的時間序列數據轉化為更低維度，使特徵更適合進行後續的評分分析。透過精簡不必要的信息，突出最關鍵的特徵。Mean Gap Layer 則對線性層的輸出進行時間維度上的平均池化(Mean Pooling)，有效地縮減特徵維度，合併每個時步的資訊，捕獲整個時間序列的平均特性，有助於模型快速處理特徵。最後將經過上述層處理過後的語音和文本嵌入向量的輸入整合並輸入進線性層得到對日語發音的準確評分。



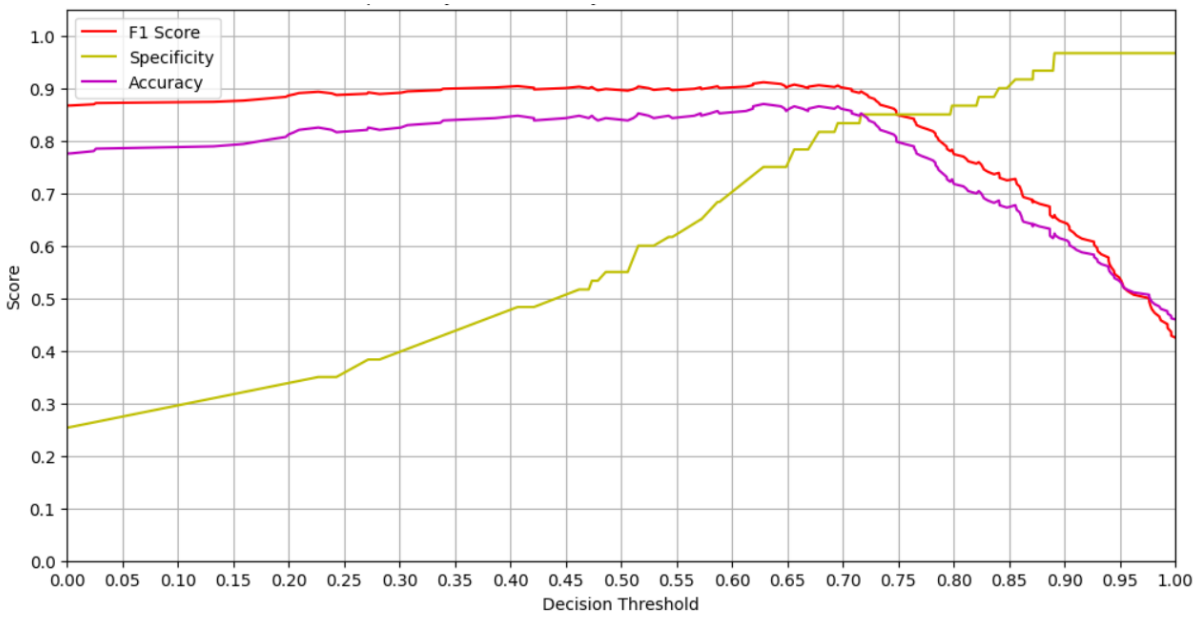
圖十、評分模型架構[7]

我們在從日文系收集的資料集上比較了搭配相同評分模型架構的 wav2vec 2.0 和 HuBERT 的性能。在評估過程中，我們使用了從日文系資料集中劃分出來的驗證集，並對 ASR 任務微調過的模型以及原始的 SSL 模型進行了比較。我們以 HuBERT 和 wav2vec 2.0 作為特徵提取器，分析評分模型的預測結果與實際標籤之間的相關性。以下是使用皮爾森相關係數（Pearson Correlation Coefficient, PCC）計算的結果，見表二。

表二、在驗證集上，評分模型預測與實際標籤的相關性

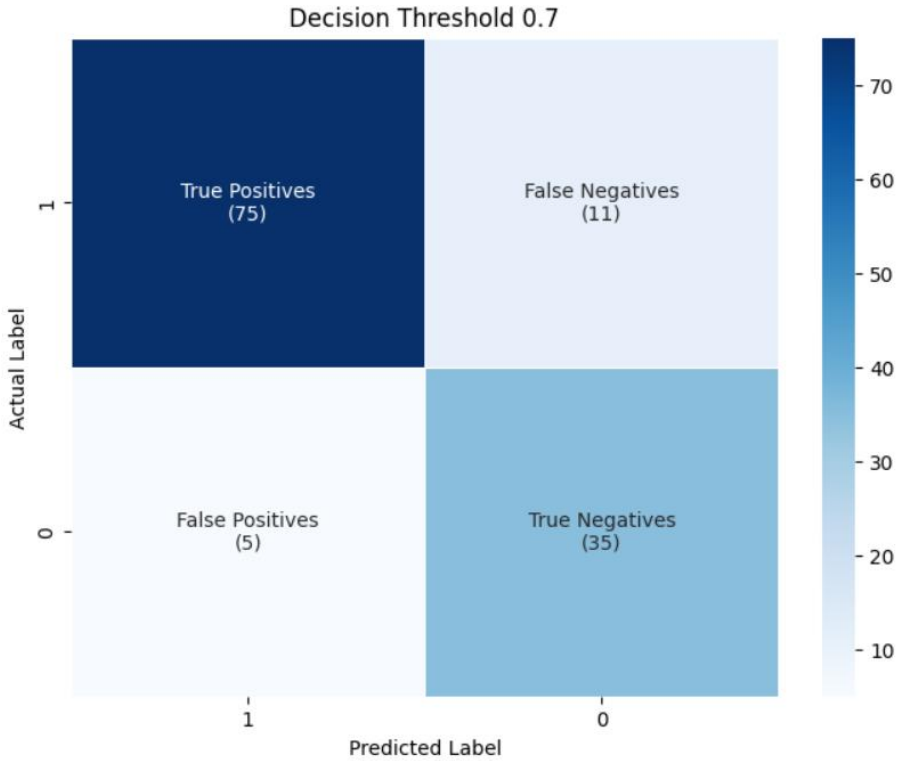
Model	**PPC**
wav2vec2 Base	0.606
HuBERT Base	0.668
fine-tuned wav2vec2 Base	0.651
fine-tuned HuBERT Base	0.701

透過表二可以得知提取 fine-tuned HuBERT 在語音中獲取的特徵搭配評分模型可以得到最大相關性，透過以下評估指標曲線(圖十一)，可以得知在閾值為 0.7 下能在 Accuracy、 F1-Score、Specificity 三項指標下獲得綜合最佳表現。



圖十一、在各個閾值下，模型在驗證集中的性能表現圖

在閾值為 0.7 時，評分模型的在測試集上的詳細表現為下圖十二的混淆矩陣，其準確值(Accuracy)為 0.87，特異值(Specificity)為 0.88，F1 Score 為 0.90。

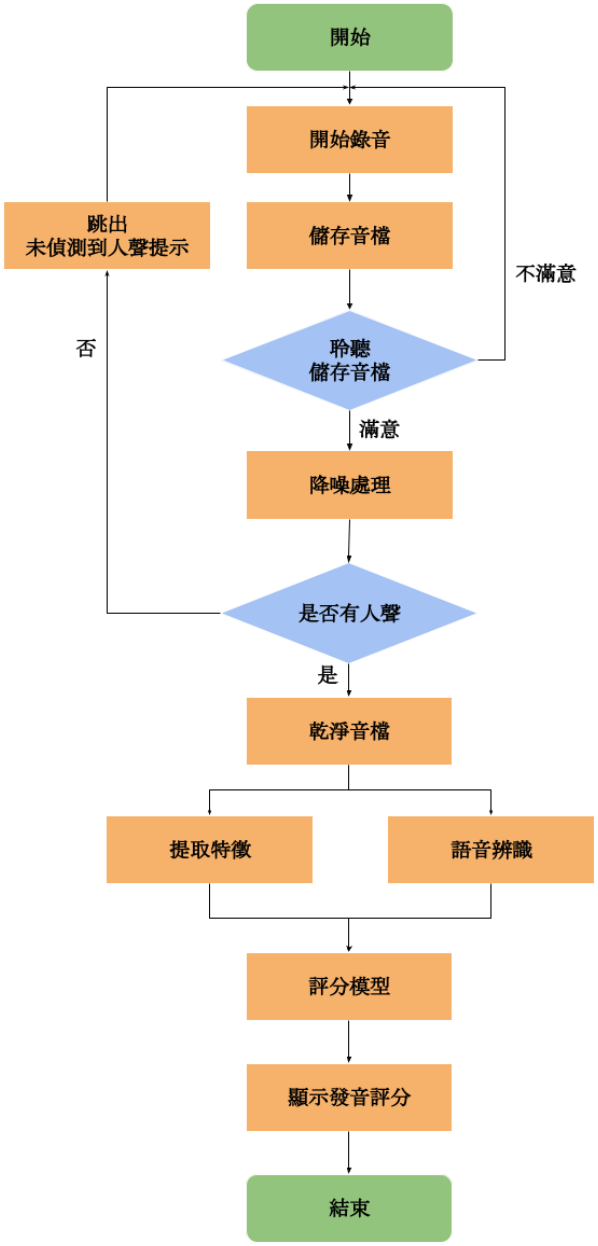


圖十二、閾值為 0.7 下，模型在測試集中表現之混淆矩陣

系統呈現

整體流程介紹

一開始進入初始頁面時，會先要進行麥克風的授權，當同意授權後，便可按下開始按鈕進入錄音介面並開始錄音，在錄音介面按下停止按鈕後，停止錄音，這時可以播放剛剛錄音的音檔，如果不滿意的話，可以按下開始按鈕並重新錄音。如果滿意的話，便可以按下上傳按鈕，按下上傳按鈕後，會將音檔傳送至後端，後端將對音檔進行降噪並判斷音檔是否有人聲，如果判斷沒有人聲的話，會進入未偵測到人聲頁面，並顯示返回重新錄音按鈕。如果判斷有人聲的話，會將音檔進行提取特徵和語音辨識，再將提取出的特徵和語音辨識出的文字送入評分模型，最後將評分模型得出的分數傳回前端，並利用結果畫面顯示分數。



圖十三、整體流程圖

Client Side

一開始進入初始頁面時，會先要進行麥克風的授權。圖十四顯示麥克風的授權。

圖十四、麥克風的授權

可以在初始頁面中點選日文文字聽取該文字的標準發音。如圖十五顯示。

圖十五、點選文字聽取正確發音

當同意授權後，便可按下開始按鈕進入錄音介面並開始錄音。圖十六顯示初始頁面。

圖十六、初始頁面

在錄音介面按下停止按鈕後，停止錄音。圖十七顯示錄音頁面。

圖十七、錄音頁面

這時可以播放剛剛錄音的音檔，如果不滿意的話，可以按下開始按鈕並重新錄音。如果滿意的話，便可以按下上傳按鈕。按下上傳按鈕後，會將音檔傳送至 server 端。圖十八顯示上傳頁面。

圖十八、上傳頁面

Server 端傳回結果並顯示評估結果和對應的文字。圖十九顯示結果頁面。

圖十九、結果頁面

Server 端傳回未偵測到人聲並顯示未偵測到人聲頁面。圖二十顯示未偵測到人聲頁面。

圖二十、未偵測到人聲頁面

Server Side

使用 Flask Framework 來進行資料的傳遞，我們設計了一個簡易的網站服務機制，使得客戶端可以透過 requests 模組以 POST 方法將音檔傳送到伺服器。伺服器端在接收到音檔後，會首先將之儲存至名為 audio 的資料夾中，待後續處理。

在音檔的初步處理中，伺服器端會使用 FRCRN 降噪模型對音檔進行降噪處理。主要目的是識別並去除音檔中的背景雜音，進而保障語音的清晰度和整體穩定性。隨後，會對音檔進行人聲偵測，這一步驟利用音檔內的振幅和音波的平均能量進行判斷。偵測過程是基於音頻的能量分佈分析技術，如果音檔的振幅或平均能量未達到特定標準，系統會認定該音檔中無人聲存在，並將此結果透過跳轉到指定的頁面提醒使用者重新錄音，如上圖二十。

若音檔中確認存在人聲，則會利用 HuBERT 模型從中提取聲學特徵。提取完畢後，進一步將音檔交由 ReazonSpeech_v2 模型處理，將語音信號轉寫成文字。這些文字隨後被輸入到一個 embedding layer 中，轉換成一系列的嵌入向量，以代表文本的內容。

結合從 HuBERT 模型獲得的聲學特徵和從 ReazonSpeech_v2 模型獲得的嵌入向量後，會被送入 BLSTM 評分模型。BLSTM 模型透過分析這些特徵來評估語音的發音標準性和表達的流暢性，最終生成一個評分。這個分數將被回傳至客戶端，並在結果頁面上顯示，使用者可以即時看到自己語音的評估結果，如上圖十九。

結論

我們專題計畫中，分為 AI 模型和系統實現兩組，所以結論將分為 AI 組與系統組分別講述：

AI 組結論

開發困難點：

面臨的主要困難有兩點，一是日文評分資料集的不足、二是強制對齊(align transcription)技術上的難度。欠缺較大且完整的日文評分資料集限制了評分模型的擴展應用，沒辦法進一步擴展評分模型的適用範圍。此外，將語音與文字進行強制對齊(align transcription)以提供針對性的發音提示，在技術上也有一定難度需要時間鑽研。

未來擴展：

期望能跟日文系合作獲取更完整的語音資料集，以提高模型的訓練水準和評分精準度。計畫將評分模型從單詞評分擴展到短句層次，將評分的範圍從只判斷標準與否到能判斷多個出評分等級，並透過強制對齊切割音檔以實現針對語音中的每個單詞提供評分，這將使評分系統更加全面和詳細。

心得：

在這次專題中，學習到許多關於音檔的前處理工作，其中包含了音檔的降噪與切割。學習了這些技術的操作過程，也理解了實際應用中的重要性。

對 Transformer 架構的自監督學習(SSL)模型有了更深刻的理解。這些模型可以適應各式各樣的語音下游任務，展現了其在處理語音方面的強大能力。這次經驗讓我們認識到了 SSL 模型在自然語言處理領域的潛力，以及它們在特徵學習中的高效性。也鑽研了 BLSTM 和 LSTM 這些處理序列資料的模型。透過對這些模型的學習和應用，加深了對於時間序列數據的處理機制的理解，也增強了我們在構建更複雜模型時的能力。學會如何透過合理的建置模型來改善語音評分的準確性。

系統組結論

開發困難點：

環境安裝是最難的，有時候真的不知道哪裡出錯，後面才知道是因為環境版本的因素導致錯的。由於使用免費可供人架設的網站，使得上傳的檔案不能太大，除非要購買會員，所以上傳檔案必須要注意大小，以免會漏失掉一些檔案，然後要和語音辨識模型的連結並不太容易，有時候會不知道有沒有成功從網站傳出音檔給模型去評分，這一點困擾我們很久。

未來擴展：

原本要利用 cordova 做出 android 用戶可以使用的 app，但由於環境因素(sdk)版本一直沒辦法好好的使用，並且現在由於資料偏少，導致評分分數沒辦法分的更詳細一點，不能根據分數的不同做出相對應的回饋給使用者體驗，所以希望未來可以成功讓手機用戶使用我們所做出來的 app，還希望未來如果可以把分數用得更詳細一點，讓我們網站所呈現的效果和動畫更加多元一點，讓使用者體驗更好。

心得：

這次是我們第一次做的專案，我們分成了系統組和 AI 組，系統組主要是在做網頁本身和如何去架設網站，過程中我們去翻閱了相關的文獻，想辦法做出我們要的成果，原本我們打算要做出 app，在做的過程中發現一開始的想法並不容易，而且網站出現各式的錯誤，才清楚瞭解到了想做和能做差別，雖然最後我們沒有如預期中做出 app，但這次讓我學到了很多相關 HTML、Javascript 和 Flask 的應用。

展示海報

日文語音評分系統

資工三A 410410707 吳天宇
資工三A 410410137 蘇柏修
資工三A 410410103 林聿朔
資工三A 410411036 施吉益
資工三A 410411788 陳緯榛

摘要

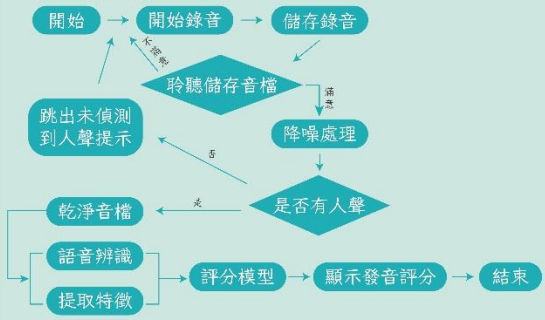
前端系統組與後端AI組合作共同開發。
系統組使用Flask建立網頁UI介面，提供使用者評分互動畫面，並使用API與後端串接。
AI組先降噪模型去除雜訊，再透過多個模型進行切割音檔、提取特徵等工作，最後由評分模型產出結果回傳至網頁。
本系統將協助學生學習正確日語發音，降低教師日常對於發音評判教學的工作負擔。

研究動機

發音與咬字需要長時間反覆地練習，才能穩定說出正確發音。傳統的發音評估依賴人工，效率低下、缺乏客觀性和一致性。因此我們決定開發一套日語評分系統，協助教師評判，也成為學生檢驗及練習的工具。

研究目的

目標開發一套自動日語發音評估系統，支持學校的教育需求。本系統基於自監督學習模型，從原始語音檔案直接學習語音特徵，免去傳統特徵工程。提高精確度，簡化評估過程，使用更便利。期望為日語教育領域帶來創新的教學方法和技術解決方案。



AI組結論

開發困難點：
欠缺較大且完整的日文評分資料集造成訓練過程困難，限制了評分模型的應用。
將語音與文字進行強制對齊難以解決，提供針對性的發音提示，技術上需要時間鑽研克服。

未來擴展：
期望能與日文系合作獲取更完整的語音資料集提高模型的訓練水準和評分精準度。計畫從單詞評分擴展至短句層次；判斷對錯拓展至給星評分，並透過強制對齊切割音檔以實現語音中的每個單詞提供評分，使評分系統更加全面和詳細。

系統組結論

開發困難點：
版本因素導致無法找出錯誤消耗許多時間。
由於使用免費資源，網站上傳的檔案大小有限制，需隨時注意以免遺失資料。
與後端串接時常有問題，難辨別網站是否成功傳送語音至後端評分。

未來擴展：
期望未來能完整做出手機端App，並將評分功能徹底完善，提供確切分數與修正建議，多元呈現各種結果，提升使用者體驗。

參考文獻

[1] Shengkui Zhao., Bin Ma., Karn N. Watcharasupat. & Woon-Seng Gan. (2022). FRCRN: Boosting Feature Representation Using Frequency Recurrence for Monaural Speech Enhancement. Retrieved from https://ieeexplore.ieee.org/document/9747578

[2] "Whisper 语音识别及 VITS 语音合成." (n.d.). EULA Club. Retrieved from https://www.eula.club/blogs/Whisper%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB%E5%8F%8AVITS%E8%AF%AD%E9%9F%B3%E5%90%88%E6%88%90.html

[3] "自然语言处理（NLP）中的语言模型预训练技术：BERT、GPT 和 T5." (n.d.). Zhihu. Retrieved from https://zhuanlan.zhihu.com/p/575873176

[4] "reazonspeech-nemo-v2." (n.d.). Hugging Face Model Hub. Retrieved from https://huggingface.co/reazon-research/reazonspeech-nemo-v2

[5] Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). https://ieeexplore.ieee.org/document/9688093

[6] Yu, Z., Deng, L., Gong, Y., & Acero, A. (2015). Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) https://ieeexplore.ieee.org/document/7404814

[7] Kim, E., Jeon, J.-J., Seo, H., & Kim, H. (2022). Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. arXiv preprint arXiv:2204.03863. https://arxiv.org/abs/2204.03863

[8] Mozilla. (n.d.). Mozilla 开发者网络 (MDN). Retrieved from https://developer.mozilla.org/zh-CN/

[9] Mozilla. (n.d.). JavaScript basics. In Mozilla 开发者网络 (MDN). Retrieved from https://developer.mozilla.org/zh-TW/docs/Learn/Getting_started_with_the_web/JavaScript_basics

[10] "How to use the format of python strings". (n.d.). IT Home. Retrieved from https://ithelp.ithome.com.tw/articles/10316731