The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Improving End-to-end Taiwanese-Speech-to-Chinese-Text Translation by Semi-supervised Learning
(通過半監督學習改進端到端台語語音至中文文字翻譯)

Yu-Chun Lin
Dept. of CSIE
National Taiwan Univ.
Taiwan
xup6m4rmp4@gmail.com

Chung-Che Wang
Dept. of CSIE
National Taiwan Univ.
Taiwan
geniusturtle6174@gmail.com

Jyh-Shing Roger Jang
Dept. of CSIE
National Taiwan Univ.
Taiwan
jang@csie.ntu.edu.tw

摘要

傳統台語語音辨識的主要問題，爲缺乏大量且公開的台語語料集，以及台語文字書寫系統不統一；前者導致進行語音辨識的任務上面臨資料不足，而後者則造成輸出格式不統一且不易讀解。因此，本研究以台語語音至中文文字的語音翻譯爲任務，透過預訓練語音模型結合端到端深度學習模型的架構，來建立台語語音至中文文字的語音翻譯模型。我們的方法是以少量台語語音配對中文文本的語料爲基礎，並透過大量蒐集未配對的台語語音資料，並設計各種演算法來利用大量未配對語料改善台語語音至中文文字的翻譯系統。研究探討主要分爲端到端語音翻譯模型、預訓練語音模型特徵、疊代訓練方法以及語料清洗四種改進方向。根據實驗結果顯示，上述方法皆能有效改善台語語音至中文文字的翻譯表現。

Abstract

The main challenges in Taiwanese speech recognition are the lack of abundant and publicly available Taiwanese speech corpora, and the inconsistency in the written system of Taiwanese. The former results in insufficient data for speech recognition tasks, while the latter leads to inconsistent output formats and difficulties in interpretation. Therefore, this study takes the speech translation from Taiwanese speech to Chinese text as the task, and builds a speech translation model from Taiwanese speech to Chinese text by combining the pre-trained speech model with the architecture of the end-to-end deep learning model. Our method is based on a small amount of Taiwanese speech paired with Chinese text, and by collecting a large amount of unpaired Taiwanese speech data, and designing various algorithms to use a large amount of unpaired corpus to improve the system of translating Taiwanese speech into Chinese text. The research and discussion are mainly divided into four improvement directions: end-to-end speech translation model, pre-trained speech model features, iterative training method and corpus cleaning. Experimental results show that the above methods can effectively improve the translation performance of Taiwanese speech to Chinese text.

關鍵字：端到端語音翻譯、半監督式學習、語料清洗

Keywords: End-to-end speech translation, Semi-supervised learning, Corpus cleaning

1 簡介

隨著資訊科技的演進，語音辨識結合自然語言處理的應用，已實際出現在我們日常生活中的許多地方，例如物聯網裝置控制 (Mehrabani et al., 2015)、車載語音助手 (Ivanko et al., 2022)，以及字幕生成 (Mathur et al., 2015) 等等。以這些應用在台灣的使用情境來看，民眾主要使用語言爲華語和台語，而其中大多數中高年齡層的長者又很可能以使用台語居多；而教育方面又有政府推動母語教育的相關政策，因此對於台語語音辨識技術的相關應用，其市場需求以及重要性相當明顯。

然而台語語音辨識有許多困難以及挑戰，主要問題可以分爲台語缺乏大量且公開語料，以及台語無統一書寫系統等問題；前者爲自動語音辨識任務中的低資源語言問題 (Zhou et al., 2018)，相較於主流語言，如英文、中文，低資源語言沒有大量且完備的語料集，導致訓練語音辨識模型無法發揮於從大量資料學習輸入輸出對應關係的能力；後者則是和台語語言本身的歷史背景和特性所導致的問題，因爲時代以及地理等因素，導致台語主要以語音爲載體，並沒有完整且獨一的書寫系統，造成在蒐集台語語料時可能有多種不同的標註形式，導致模型學習資料的處理和訓練困難增加；同時

多種不同書寫方法的台語也不易互相解讀，因此，如何解決語料問題和語言特性，在端到端語音模型訓練中也是一大挑戰。

相較傳統語音辨識的高斯混合模型-隱藏式馬可夫模型，目前最新端到端語音辨識模型可以達到更高辨識率和對環境更強的強健性(Watanabe et al., 2018)。此外，隨著半監督學習 (Park et al., 2020) 預訓練等方法的發展，對於低資源語言的語音辨識又提供了更多的發展可能。本研究將以上述端到端語音辨識結合機器翻譯作爲研究方向，來將台語語音辨識任務轉換爲台語語音對應中文翻譯的任務。

2 研究方法

2.1 端到端語音翻譯模型

從語言性質來觀察，可以發現台語和中文的文法結構相近，發音規則也有許多相近之處，並且在中文和台文上書寫文字也相同。我們因此根據過往研究 (Bentivogli et al., 2021) 所評估的語言特性對於語音翻譯系統架構的選擇，決定採用端到端語音翻譯的架構來實作。

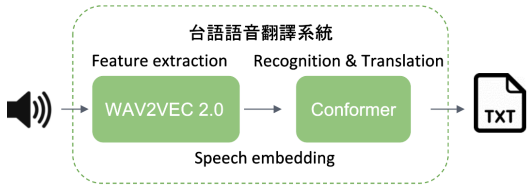

Figure 1: 端到端語音翻譯系統示意圖

圖1爲端到端台語語音翻譯系統示意圖，左側輸入爲語音，右側輸出爲中文翻譯文本。我們研究過往提出的架構，最後決定將整體架構分爲上下游兩個模型，上游特徵抽取使用 WAV2VEC 2.0 0 (Baevski et al., 2020) 預訓練語音模型，下游任務使用 Conformer (Gulati et al., 2020) 作爲端到端語音翻譯模型。其中對於 WAV2VEC 2.0，我們亦參考了前人研究 (Hsu et al., 2021)，使用 fairseq (Ott et al., 2019) 開源工具，來使用大量無標註台語語料作預訓練階段的微調。

2.2 半監督疊代訓練

半監督疊代訓練的目標，是透過大量無標註台語語音進行訓練語料擴增，解決語料不足的問題。於本研究中，我們主要參考了 Noisy student training (Park et al., 2020) 的訓練流程，設計半監督式疊代訓練。其整體的流程如圖 2 所示，主要會分成兩個階段。第一階段是訓練教師模型，與過往研究不同的是，我們在此處使用了預訓練語音模型，其中預訓練語音模型不隨著下游任務一起訓練，所以模型參數是固定不動的。第二階段是機器標註生成，透過訓練好的語音翻譯模型對無標註語音進行翻譯，生成對應語音的機器標註。由生成的機器標註加入原先的標註語料後返回第一階段訓練學生模型，反覆在兩個階段來回疊代，逐步改善模型能力以及語料品質，便是整個半監督學習的流程。

對於機器標註的生成，我們會額外再訓練一個 Transformer 語言模型，讓語言模型與語音翻譯模型在生成標註時進行淺融合。這麼做的目的是修正語音翻譯模型輸出的語句，使其結果更加準確，提高文本品質；並且透過語言模型協助翻譯，將不同訓練語料的資料分布特性帶入機器標註中，能增添整體訓練資料的多樣性。

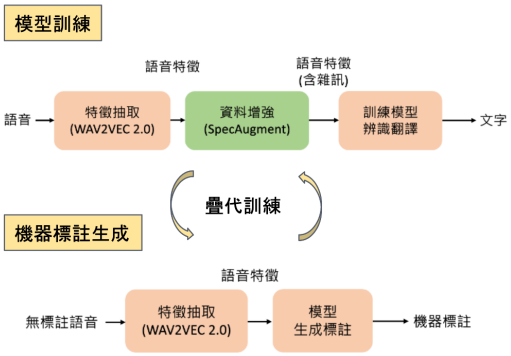

Figure 2: 半監督學習流程圖

2.3 語料清洗

由於大量語料的收集來自於網路之中，因此衍生出的問題是難以控制語料的品質，這些資料若是不經清理加入訓練，反而可能成爲訓練負樣本，導致模型學習錯誤的資訊。因此，我們針對這些大量收集的語料進行清洗處理，包含有標註的中品質語料和無標註的低品質語料。我們在方法上結合了自行訓練的語音翻譯模型進行清洗，並分爲基於翻譯出的文本的清洗，以及基於原始聲音訊號的清洗兩類方式。

2.3.1 文本處理

基於模型生成出的文本的清洗，我們進行了以下的處理：

1. 標註語料清洗：此目標是將有標註的中品質語料進行清洗，這類型的資料雖然格式與一般標註語料相同，但有兩個主要問題，一是無法確定文本內容和格式符合目標標註，例如雖然是目標是中文字幕但實際是收集到台文；二是字幕時間戳記準確度不佳，因爲對視眾來說語音和字幕不必完美對齊，但語音文本沒有對齊對於模型

學習會造成問題。所以我們提出了進行標註語料清洗，亦即利用訓練好的語音翻譯模型對標註語音進行翻譯，再利用機器文本與原標註進行 CER （character error rate，字錯誤率）計算的，當 CER 高於一定閥值，則以機器文本取代原本的標註。反之，則相信原標註並保留不變。這種方法優點在於可以用模型達到正規化訓練資料的效果，包含常見同義詞和台文標註的案例；也可以透過模型來進行初步的語音及文本的對齊，改善原本標註中不對齊的情況。

2. 語言模型過濾：此目標是將模型預測信心分數較低的標註去除掉。具體作法是對同一個語音進行兩次辨識，第一次只利用語音翻譯模型進行標註生成，第二次則結合 Transformer 語言模型進行淺融合得到修正的標註。我們設計過濾的方法參考了過往研究 (Chen et al., 2023)，透過比較模型使用淺融合前後的文本一致性決定文本的好壞；亦即 CER 較高時，代表模型並沒有很好的進行翻譯，被語言模型大量修改；反之則代表輸出結果較爲一致，是具有高可信度的翻譯結果。

3. 語速過濾：此目標是將語音片段中，可能爲音樂或歌聲的標註去除掉，具體方法是從語音和其被自動標註完成的文本進行長度統計，計算出每個片段的語速（即文本長度除以音檔長度），並依據對資料集的統計，來濾除語速太高（可能是解碼錯誤）或太低（可能是歌聲或無聲）的語句。

2.3.2 語音處理

基於原始的聲音訊號，我們則進行了以下的處理：

1. 語音活性偵測：此方法的目標是重新切割語音片段，達到去除片段中非人聲的部分，包括靜音、純音樂或背景噪音等等。我們採用的方法，是 RNNoise (Valin, 2018) 這項研究設計的 RNN （recurrent neural network）模型，來針對音檔輸入做人聲的偵測，並依此重新切割音檔。而我們除了使用原論文的預訓練模型以外，同時也利用少量台語資料串接合成出包含語音和安靜片段的混合音檔，來對原始模型進行微調。

2. 語言辨識：此方法目的是去除語音資料中非台語語音的片段。在收集大量的語音中，雖然我們能透過各種資訊從網路抓取台語語音，但還是無法避免收集到非台語的語音，尤其在許多戲劇或是新聞播報中，往往是中文和台語兩種語言夾雜，導致收集資料時有一部分比例實際是中文或英文等其他語言。這些資料可能有害於模型訓練，在機器標註時也可能出現錯誤。因此，我們透過訓練語言辨識模型，並根據模型辨識輸出的字詞類型，統計最高比例的字詞類型，來判斷一個輸入語音片段所屬的語言。

3 資料集介紹

本節將說明本研究所使用之資料集，包含將應用於語音翻譯模型本身的兩個台語資料集 TAT (Taiwanese across Taiwan) 以及 TAI YouTube，以及應用於語料清洗的英文 LibriSpeech 資料集，以及中文的 Common Voice Chinese 資料集。

3.1 台語資料集：**TAT (Taiwanese across Taiwan)**

TAT (Taiwanese across Taiwan)(Liao et al., 2022) 資料集是一個台語朗讀資料集，包含音訊和原生台文文本。爲了涵蓋台語發音的多樣性，收集來自台灣各地不同腔調的台語語音。每一次錄音同時以 6 種不同麥克風進行錄製，包括專業麥克風、IOS 裝置和 Android 裝置等等設備，並且在錄音後由人工進行二次校正文本和錄音對齊。TAT 資料集收集計畫由北科大師生發起，收集目的是爲了提供台語語音辨識研究和相關技術開發。收集時間爲 2019 至 2022 年，總共約 600 位語者參與錄製，每位語者錄製時間爲半小時，總共時長爲 300 小時，並切分 3 個資料集，分別爲 TAT-Vol1 50 小時、TAT-Vol2 50 小時和 TAT-MOE 200 小時。

因爲資料集標註皆爲台文文本，需要整理爲合適台語語音至中文文字的翻譯任務的格式，因此我們由 TAT-Vol1 資料集取出 4 小時共 2,452 句的語料，進行人工翻譯爲中文文本。

3.2 台語資料集：**TAI YouTube** 資料集

TAI YouTube 資料集是我們自行收集和整理的台語資料集，該資料集的收集目的有兩個方面，一是收集用於訓練台語語音翻譯模型的標註語料，其中的格式爲台語語音配對中文文本；二是收集大量無標註的純台語語音資料，作爲後續半監督資料擴充實驗的資料來源。

資料集的收集方法，爲利用網路影音平台收集與台語相關的大量影音內容，所收集的資料類型分爲具有字幕以及無字幕兩種。對於具有字幕的影片，我們利用字幕提供的時間戳記，

Table 1: TAI YouTube 資料集詳細資訊

資料集名稱	時長 (小時)	資料品質	標註類型
DaAi	40	中	YouTube CC 字幕
PTS	40	中	YouTube CC 字幕
Taiwan-mystery	40	中	YouTube CC 字幕
Unsupervised	2,000	低	N/A

將其按句切割成一份份語音至文本的配對資料。對於沒有字幕的影片，我們直接下載完整音檔，並通過基於能量規則的語音活動檢測 (Voice activity detection, VAD)(Pang, 2017)，將大音檔切割成許多至多 12 秒長度的音檔，以利於後續訓練預訓練語音模型和語音翻譯模型。

資料則主要來源於 YouTube 平台。我們從 YouTube 下載了各式各樣語音內容，包括大愛電視台 (DaAi)、公視電視台 (PTS) 以及民視電視台 (Taiwan-mystery)。這些頻道對應的主要內容類型分別是戲劇、新聞播報和介紹型節目，其中大愛和民視以中文字幕居多，而公視則是以台文字幕居多。此外，我們還廣泛收集了其他不特定領域的資料，總時長約爲 2,000 小時，細節整理如表 1。

3.3 英文資料集：LibriSpeech

LibriSpeech (Panayotov et al., 2015) 是一個常用的英文語音辨識資料集，它包含文本和語音，是一個有聲書閱讀的資料集。該資料集總共約有 1,000 小時的英語演講，聲音的取樣率爲 16 kHz。

該資料集的來源主要爲 LibriVox 專案，旨在提供有聲讀物的免費錄製。爲了建立 LibriSpeech 資料集，研究人員對這些有聲讀物進行了分項細分、整理合併的處理，最終切割和整理成每條約 10 秒左右的音訊檔案，並進行了文本標註。這樣的處理方式使得 LibriSpeech 成爲了一個常用的資料集，對於進行英文語音辨識任務非常有用。

3.4 中文資料集：Common Voice Chinese

Common Voice(Ardila et al., 2020) 是一個由 Mozilla 組織發起的開源計畫，旨在創建一個可由任何人使用的大規模多語言語音資料集。該資料集由全球志願者提供的語音樣本組成，主要用於訓練以及改善語音辨識相關任務系統。Common Voice 資料集包含了以下特色：

1. 多語言: 資料集包含來自世界各地不同語言的語音樣本，使得各地研究員能夠使用多國語言進行跨語言語音辨識研究。
2. 開源: 任何人都可以自由使用、分享和改進資料集。提供語音技術發展更廣泛參與和創新空間。
3. 多樣性與包容性: 資料集包含來自不同年齡、性別、口音和背景的人群樣本，有助於改善語音辨識系統對各種口音和多樣性的理解能力。
4. 數據驗證: 每個樣本經多人驗證，確保語音樣本的正確性和可靠度。
5. 數據蒐集平台: 爲方便資料的收集和貢獻，Common Voice 提供線上平台供使用者朗讀文本、錄製語音並上傳樣本，促進大眾群體共同貢獻與維護資料集。

本研究主要使用 Common Voice 資料集的中文部分，該部分亦經許多不同的整理維護，目前參與錄製語音樣本的語者超過 2,000 人，總語音時數約爲 120 小時。

4 實驗設定與結果

4.1 輸出正規化

正規化的目的是消除辨識結果與正確答案之間，因爲表示形式或格式差異，而引起的評量誤差。這種差異可能導致使用評量指標評估模型翻譯結果時，與實際翻譯效果有所落差。常見正規化處理的部分包括阿拉伯數字轉換，以及同義詞的處理等等，這些狀況可能導致計算評估指標時產生錯誤的估計，即本應評估爲正確結果的詞，因爲使用不同詞語，被評估爲錯誤翻譯。

爲了解決這些問題，我們需要採取正規化策略。首先，對於數字一律轉爲中文字表示，標點符號和空白等非中文字符號則一律移除，以確保計算指標只限於中文字翻譯結果。而對於同義詞的問題，我們建立同義詞表，範例如表 2，以將同義的詞彙映射到一個共同的標準詞彙。這樣可以確保在評估時同義詞校正到相同的結果，從而提高評估的準確性。

雖然同義詞中例如「口音」和「腔調」，根據談話內容和情境，可能不適合做替換，但本研究採用正規化的目的，主要目的是在不同系

Table 2: 同義詞表範例

正規化前	正規化後
人家	人
能	可以
小孩	孩子
口音	腔調

統測試下，對於設計的測試集能有同一基準的評量結果，不受同義詞的影響。

4.2 評估指標

本研究採用的評估指標是 CER。當 CER 的數值越低，代表語料中被辨識錯誤的字越少。計算 CER 時，首先將預測和參考序列對齊，並設定 S 爲被替換的字元數目，D 爲被刪除的字元數目，I 爲被插入的字元數目，N 代表參考序列的字元總數，並以如下公式計算：

$$CER = \frac{S + D + I}{N} \quad (1)$$

BLEU （bilingual evaluation understudy，雙語替換評測）(Papineni et al., 2002) 也是一個常見的評估指標，其結果是透過蒐集機器翻譯字句和參考翻譯字句配對的 n-gram 數量來計算，可以代表著機器翻譯和參考翻譯之間的相似度。我們曾預先對一小部分的測試集，觀察 BLEU 與 CER 的差異 (Lin, 2023)，發現兩者呈高度相關，如圖 3，而 CER 具有更直觀和易於解釋的特性，因此最終選擇使用 CER 作爲評量指標。

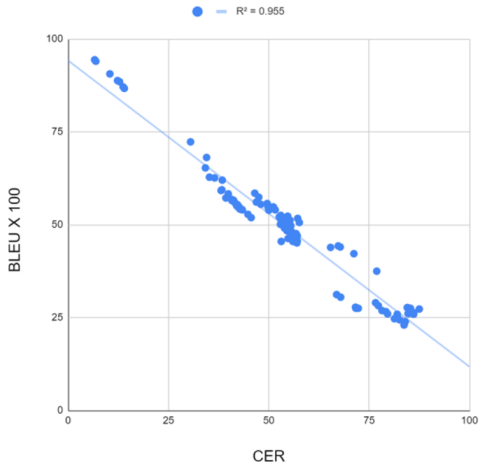

Figure 3: CER 與 BLEU 比較結果

4.3 實驗結果

4.3.1 下游模型與上游特徵

實驗首先固定使用傳統語音特徵 80 維 Fbank，替代爲上游模型輸出的特徵進行訓練，實驗結果如表 3所示，其中除了 Whisper (large) 爲零樣本學習以外，其於模型皆經過台語語料微調。從實驗結果可以發現，whisper (large) 的字錯率約爲 60.8%，以零樣本測試來看是目前最好的結果。而將 whisper (medium) 以台語語料進行微調訓練後可以改善到達 49.4%，可以觀察出進行語料微調的有效性以及必要性。而比較所有下游模型後，其結果以 Conformer 表現最佳，字錯率僅 40.0%；雖然我們也認爲 Whisper 較差的原因可能在於訓練資料量較少，測試也在較小的範圍領域內，但以本研究來說，我們仍根據以上結果，選擇 Conformer 作爲下游模型。

表 4的結果則爲不同上游特徵的比較，其中的下游模型皆爲 Conformer。我們可以觀察到，WAV2VEC 2.0 比起 Fbank 的效果較佳，且經過 TAI YouTube 無標註的 2,000 小時語料訓練 Taiwanese-WAV2VEC 2.0 後，在台語測試集上可以獲得更多改善，同時結果也說明預訓練語言模型最大的優勢在於訓練不需要像語音辨識任務，需要大量語音和文本的配對語料才能進行訓練。透過相對容易收集的純語音，不需要人工標註，便能進行自監督訓練，進而對目標任務有改善效果，發揮大資料量的優勢。然而，基於實作資源上的考量，我們在後續實驗中，仍然是以未經台語語料微調的 WAV2VEC 2.0 進行。

4.3.2 半監督疊代訓練

本實驗的目的是觀察半監督訓練方法的有效性，以及無標註語料的擴增量對於模型辨識度的影響。我們訓練每個模型基於同樣標註語料，分別加入無標註語料的擴增量爲 0、100、200 和 400 小時。

機器標註的生成是利用前項實驗最好的模型作爲教師模型，與語言模型進行淺融合。端到端台語語音辨識模型，則根據前面實驗結果及說明，使用 Chinese-WAV2VEC 2.0 作爲語音特徵抽取器，且其模型參數固定，不做下游任務訓練時的前向梯度傳遞。實驗結果如表 5發現，進行無標註語料擴增可以幫助改善辨識度，以 200 小時的擴增量來看能達到 2.7% 相對錯誤率改善。證明教師模型生成文本得到的機器標註語料有助於改善系統辨識度。然而當擴增語料量達到 400 小時，字錯誤率又有上升的趨勢，說明直接擴增無標註語料並不一定能持續改善辨識度，因此我們需要一套清洗語

Table 3: 上游特徵爲 Fbank 時的下游模型效能

下游模型	正規化前 CER	正規化後 CER
Whisper (large) (Radford et al., 2022)	65.3	60.8
RNN-LSTM (Graves et al., 2013)	58.9	50.8
Whisper (medium) (Radford et al., 2022)	56.5	49.4
Transformer (Zhang et al., 2020)	55.4	43.9
Conformer (Gulati et al., 2020)	52.0	40.0

Table 4: 不同上游特徵時的下游 Conformer 模型效能

上游特徵	正規化前 CER	正規化後 CER
Fbank	52.0	40.0
WAV2VEC 2.0	50.0	36.8
Taiwanese-WAV2VEC 2.0	49.8	35.5

料的方法來改善訓練語料，以確保加入訓練資料的品質。

Table 5: 不同擴增語料量的模型效能

擴增時數	正規化前 CER	正規化後 CER
0	50.0	36.8
100	50.6	36.0
200	49.6	34.1
400	49.4	34.9

4.3.3 語料清洗

Table 6: 不同清洗方式及擴增語料量的正規化後 CER

清洗方式\擴增時數	200	400
無清洗	34.1	34.9
標註清洗	33.9	32.9
LM	34.0	34.5
SR	34.0	34.4
LM+SR	33.9	34.1
VAD	30.7	30.6
LID	31.2	31.7
VAD+LID	30.7	30.0

進行各種不同語料清洗與否的結果，列於表 6。從實驗結果可以觀察到，標註清洗後對於台語語音翻譯的訓練有改善效果，代表透過標註的清洗可以修正原本語料中不一致的同義詞標註，達到接近正規化的效果。在語言模型 (LM) 過濾以及語速 (SR) 過濾的方面，我們可以發現兩者都比未處理有稍微好的表現，其中語速過濾器效果又比語言模型過濾器稍佳，但是都不如標註清洗的改善明顯，因此仍需要其他的清洗方法。

而在使用了語音活性偵測 (VAD) 以及語言辨識 (LID) 的清洗方式後，可以模型翻譯的效果，在使用 200 小時和 400 小時的擴增語料時，都有較顯著的改善，其中語音活性檢測的效果又比語言辨識過濾音檔好，且將兩種方法結合在 400 小時擴增量實驗下，與未處理相比有接近 5% 的相對字錯率改善。反映出語音活性偵測切割音檔對於語料有很大影響，不只能夠去除音檔中非人聲的無效語料，也提供模型訓練的音檔長度多樣性增加；語言辨識過濾也有近似效果，並且可以挑出在收集語料中不是目標台語語音的音檔，提升整體語音品質。

5 結論與未來工作

本研究設定由台語語音辨識翻譯中文輸出爲目標，以少量台語語音和中文文本的配對語料爲基礎，透過大量蒐集網路台語語音資料，設計語料清洗演算法。以語音預訓練模型結合端到端深度學習模型，訓練並改善台語語音翻譯系統。研究探討主要分爲端到端語音翻譯模型、預訓練語音模型特徵、半監督疊代訓練方法以及語料清洗四種改進方向。根據實驗結果，驗證上述方法皆能有效改善台語語音翻譯中文效果。

本研究的未來方向，除了將各種清洗方式進行整合性的測試以外，也將基於 Whisper 或其他模型對於多任務語音資訊的理解基礎，來提供整個系統能夠解決對應新詞彙的問題；此外，也可以研究使用 BERTScore(Zhang et al., 2019) 或 Sentence-transformer(Reimers and Gurevych, 2019) 作爲模型訓練的輔助目標函數，以在訓練時加入機器翻譯指標的學習目標，讓原本從單一語音對應單一翻譯的學習，轉爲能夠從語音學習到可能的翻譯結果，來提升模型理解的泛化。

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? *arXiv preprint arXiv:2106.01045*.

Yu Chen, Wen Ding, and Junjie Lai. 2023. Improving noisy student training on non-target domain data for automatic speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. 2021. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*.

Denis Ivanko, Dmitry Ryumin, Alexev Kashevnik, Alexandr Axyonov, and Alexey Karnov. 2022. Visual speech recognition in a driver assistance system. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1131–1135. IEEE.

Yuan-Fu Liao, Jane S Tsay, Peter Kang, Hui-Lu Khoo, Le-Kun Tan, Li-Chen Chang, Un-Gian Iunn, Huang-Lan Su, Tsun-Guan Thiann, Hak-Khiam Tiun, et al. 2022. Taiwanese Across Taiwan corpus and its applications. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–5. IEEE.

Yu-Chun Lin. 2023. Improving End-to-end Taiwanese-to-Chinese Speech Translation by Semi-supervised Learning. *Master Thesis, National Taiwan University*.

Abhinav Mathur, Tanya Saxena, and Rajalakshmi Krishnamurthi. 2015. Generating subtitles automatically using audio extraction and speech recognition. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 621–626. IEEE.

Mahnoosh Mehrabani, Srinivas Bangalore, and Benjamin Stern. 2015. Personalized speech recognition for internet of things. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pages 369–374. IEEE.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

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

Jing Pang. 2017. Spectrum energy based voice activity detection. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–5. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *2022 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. 2020. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jean-Marc Valin. 2018. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Shiyu Zhou, Shuang Xu, and Bo Xu. 2018. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. *arXiv preprint arXiv:1806.05059*.