

國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



用於端到端中文自動語音辨識的語境偏移

Contextual Biasing for End-to-End Chinese ASR

張開

Zhang Kai

指導教授：張智星 博士

Advisor: Jyh-Shing Roger Jang Ph.D.

中華民國 112 年 5 月

May, 2023

國立臺灣大學碩士學位論文  
口試委員會審定書



用於端到端中文自動語音辨識的語境偏移

Contextual Biasing for End-to-End Chinese ASR

本論文係張開君（學號 R10944061）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國一百一十二年五月二十日承下列考試委員審查通過及口試及格，特此證明。

口試委員：

張開君

(簽名)

(指導教授)

李宏毅

陳允呈

所長：

鄭卜壬





## 致謝

此論既成，先謝張智星師之教導，每週會議中賜指示，予我實驗室資源。論文書及口試備，亦多授方略。

又謝葉子、小龜二師兄之教導，使我學會書論及行研。特謝葉子師兄於消防局計劃中之指示，且不辭勞苦與我商議研方。

又謝父母之供養，使我學無憂，且得遊玩。謝父親購高性能電腦，此論多實驗於此電腦上成。謝母親支持我選擇，使我得從所好之方。

又謝實驗室同學之陪伴與交流。謝張秋霞、陳炫均、林育駿、黃一洋等同學與我同修課，共成小組作業及期末作業。

又謝語音組眾小夥伴之支持與分享。謝政鷹、毓慧、淑芬、彥綸四師兄姐於我學 Kaldi 軟體時之支持。每週二晚會議中，眾人皆能發表意見。

又謝育辰、雲行二學弟之支持，我們共建實驗室 nas 設備及升千兆網路。

又謝大學及研究生期間之友人陪伴。謝徐啟涵、林楠翔等自幼相識之友。及冠廷、建逢、偉恩等於台灣相識之友。

又謝大學師羅峻旗，高中師劉益琴於我計算機知識學習之道上之助與持。

最後謝 MIRLAB 實驗室眾同學，能入實驗室至碩士畢業是我一段珍貴時光。





## 摘要

端到端語音辨識方法相較於傳統方法，魯棒性較高，能在多種情境下提高辨識準確率。但因缺乏獨立語言模型，無法辨識訓練資料外的詞彙，影響部分專有名詞的辨識效果。要適應不同場合，必須對特定領域進行偏移。本研究以CATSLU資料集為基礎，建構了兩項中文語境偏移的任務，分別針對專有名詞和混合領域的語句。並且探討了四種在語音辨識流程的不同階段進行語境偏移的方法，包括辨識前、模型、解碼和後處理四個階段。實驗結果顯示，所有的偏移方法都在一定程度上提升了語音辨識模型在特定領域的辨識效果。

**關鍵字：**自動語音辨識，語境偏移，端到端語音辨識，自監督訓練模型，意圖辨識，意圖分類，熱詞





# Abstract

Compared with traditional methods, end-to-end speech recognition methods have higher robustness and can improve recognition accuracy in various contexts. However, due to the lack of an independent language model, they are unable to recognize vocabulary outside of the training data, which affects the recognition effect of some specialized terms. Specific domain adaptation is necessary to adapt to different situations. Based on the CATSLU dataset, this study constructed two Chinese context biasing tasks targeting specialized terms and mixed-domain sentences, respectively. Four different methods of context biasing were explored at different stages of the speech recognition process, including preprocessing, model refinement, decoding strategy, and postprocessing. The experimental results showed that all context biasing methods improved the recognition effect of the speech recognition model in specific domains to some extent.

**Keywords:** Automatic Speech Recognition, Contextual Bias, Context Biasing, End-to-end Speech Recognition, CATSLU, Intent Classification, Hot Word





# 目錄

|                                | Page       |
|--------------------------------|------------|
| 口試委員審定書                        | i          |
| 致謝                             | iii        |
| 摘要                             | v          |
| <b>Abstract</b>                | <b>vii</b> |
| 目錄                             | ix         |
| 圖目錄                            | xiii       |
| 表目錄                            | xv         |
| <b>第一章 緒論</b>                  | <b>1</b>   |
| 1.1 研究動機 . . . . .             | 1          |
| 1.2 研究貢獻 . . . . .             | 3          |
| 1.3 章節概述 . . . . .             | 3          |
| <b>第二章 相關文獻</b>                | <b>5</b>   |
| 2.1 傳統語音辨識 . . . . .           | 5          |
| 2.1.1 聲學模型 . . . . .           | 6          |
| 2.1.1.1 GMM-HMM 聲學模型 . . . . . | 6          |
| 2.1.1.2 DNN-HMM 聲學模型 . . . . . | 8          |
| 2.1.2 發音詞典 . . . . .           | 9          |
| 2.1.3 語言模型 . . . . .           | 9          |



|            |                                   |           |
|------------|-----------------------------------|-----------|
| 2.1.3.1    | N-gram 語言模型 . . . . .             | 9         |
| 2.1.3.2    | 神經網絡語言模型 . . . . .                | 10        |
| 2.1.3.3    | 評估指標 . . . . .                    | 11        |
| 2.2        | 端到端語音辨識 . . . . .                 | 11        |
| 2.2.1      | CTC 自動對齊 . . . . .                | 12        |
| 2.2.2      | 序列到序列模型 . . . . .                 | 12        |
| 2.2.3      | Conformer 端到端語音辨識模型 . . . . .     | 13        |
| 2.3        | 聲學特徵 . . . . .                    | 14        |
| 2.3.1      | 梅爾頻率倒譜系數 (MFCC) . . . . .         | 15        |
| 2.3.2      | FilterBank 語音特徵 (Fbank) . . . . . | 15        |
| 2.3.3      | 使用自監督預訓練模型提取語音特徵 . . . . .        | 16        |
| 2.4        | 語境偏移 . . . . .                    | 16        |
| 2.4.1      | 清洗訓練資料 . . . . .                  | 17        |
| 2.4.2      | 淺融合 . . . . .                     | 18        |
| 2.4.3      | 替代拼寫預測 . . . . .                  | 19        |
| 2.4.4      | CLAS 基於注意力機制的語境偏移 . . . . .       | 19        |
| 2.4.5      | Whisper 的語境偏移 . . . . .           | 20        |
| 2.4.6      | 語境偏移相關任務 . . . . .                | 21        |
| <b>第三章</b> | <b>資料集和任務介紹</b>                   | <b>23</b> |
| 3.1        | 資料集介紹 . . . . .                   | 23        |
| 3.1.1      | CATSLU 資料集 . . . . .              | 23        |
| 3.1.2      | WenetSpeech 中文語音辨識資料集 . . . . .   | 24        |
| 3.1.3      | AISHELL-1 . . . . .               | 25        |



|  |           |
|--|-----------|
| 3.1.4 AISHELL-2                              | 26        |
| 3.2 任務介紹                                     | 27        |
| 3.2.1 CATSLU 專有名詞語境偏移                        | 27        |
| 3.2.2 CATSLU 多領域混合語料語境偏移                     | 28        |
| 3.2.3 評量方法                                   | 29        |
| 3.2.3.1 字錯誤率                                 | 29        |
| 3.2.3.2 關鍵詞錯誤率                               | 29        |
| <b>第四章 研究方法</b>                              | <b>31</b> |
| 4.1 端到端意圖識別                                  | 31        |
| 4.1.1 整體流程                                   | 32        |
| 4.1.2 HuBERT 進行特徵提取                          | 33        |
| 4.1.3 意圖辨識下游模型                               | 33        |
| 4.2 微調預訓練模型                                  | 34        |
| 4.3 中文淺融合                                    | 35        |
| 4.4 備選詞預測 (Alternative Word Prediction, AWP) | 36        |
| 4.4.1 訓練資料的產生方法                              | 36        |
| 4.4.2 備選詞預測模型架構                              | 38        |
| 4.4.3 備選詞表的生成和使用                             | 38        |
| 4.4.4 常見詞表                                   | 39        |
| <b>第五章 實驗設計和結果討論</b>                         | <b>41</b> |
| 5.1 實驗流程及設定                                  | 41        |
| 5.1.1 實驗流程                                   | 41        |
| 5.1.2 實驗環境                                   | 42        |

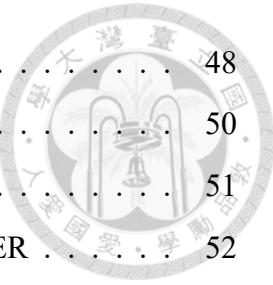


|             |                       |           |
|-------------|-----------------------|-----------|
| 5.2         | 使用端到端意圖識模型進行自動語境檢測    | 42        |
| 5.2.1       | 不同下游模型對意圖辨識的影響        | 42        |
| 5.2.2       | 不同方法的算力和正確率對比         | 43        |
| 5.2.3       | 在語境偏移前應用語境檢測方法        | 45        |
| 5.3         | 少量語料微調模型實現語境偏移        | 46        |
| 5.4         | 解碼和後處理階段語境偏移          | 47        |
| 5.4.1       | 偏移分數對中文淺融合的影響         | 47        |
| 5.4.2       | AWP 模型選擇              | 49        |
| 5.4.3       | 常見詞列表                 | 50        |
| 5.4.4       | 多領域任務中應用中文淺融合和 AWP 方法 | 52        |
| <b>第六章</b>  | <b>結論和未來工作</b>        | <b>55</b> |
| 6.1         | 結論                    | 55        |
| 6.2         | 未來工作                  | 56        |
| <b>參考文獻</b> |                       | <b>59</b> |



# 圖目錄

|     |                             |    |
|-----|-----------------------------|----|
| 1.1 | 傳統語音辨識流程和端到端語音辨識流程          | 2  |
| 2.1 | 語音辨識系統整體框架圖                 | 5  |
| 2.2 | 語音辨識 HMM 示意圖                | 8  |
| 2.3 | RNN 語言模型示意圖                 | 10 |
| 2.4 | RNN 端到端語音辨識模型示意圖 [7]        | 13 |
| 2.5 | Conformer 模型示意圖             | 14 |
| 2.6 | 清洗訓練資料的方法示意圖                | 17 |
| 2.7 | CLAS 偏移關鍵詞數量對於 WER 的影響 [44] | 20 |
| 3.1 | CATSLU 專有名詞語境偏移各類別語句數量      | 28 |
| 4.1 | 端到端意圖辨識流程示意圖                | 32 |
| 4.2 | 線性下游模型示意圖                   | 34 |
| 4.3 | 中文淺融合 WFST 示意圖              | 35 |
| 4.4 | 輸入輸出文字 jieba 斷詞後範例          | 37 |
| 4.5 | 完成對齊後句子範例                   | 37 |
| 4.6 | 備選詞預測模型架構圖                  | 38 |
| 4.7 | 備選詞表範例圖                     | 39 |
| 4.8 | 常見詞計數方法示意圖                  | 40 |
| 5.1 | 使用不同下游模型的正確率對比實驗結果          | 43 |
| 5.2 | 不同方法的正確率對比實驗結果              | 44 |
| 5.3 | 不同方法的效率對比實驗結果               | 44 |
| 5.4 | 語境檢測應用在語境偏移中實驗結果            | 45 |
| 5.5 | 微調模型實驗結果                    | 47 |



|      |  |    |
|------|--|----|
| 5.6  | 偏移分數調整實驗結果 . . . . .                       | 48 |
| 5.7  | 不同的 AWP 模型架構實驗結果 . . . . .                 | 50 |
| 5.8  | 常見詞列表對照實驗結果 . . . . .                      | 51 |
| 5.9  | CATSLU 專有名詞語境偏移全部子任務實驗語句整體 CER . . . . .   | 52 |
| 5.10 | CATSLU 專有名詞語境偏移全部子任務實驗關鍵詞錯誤率 KER . . . . . | 53 |



# 表目錄

|     |                                    |    |
|-----|------------------------------------|----|
| 3.1 | CATSLU 資料集各領域詳細資訊 . . . . .        | 24 |
| 3.2 | WenetSpeech 各領域語句時長分佈 . . . . .    | 25 |
| 3.3 | CATSLU 多領域混合語料語境偏移數據集劃分 . . . . .  | 28 |
| 4.1 | 備選詞預測模型訓練資料範例 . . . . .            | 36 |
| 4.2 | ASR 模型輸出和 groundtruth 範例 . . . . . | 36 |
| 5.1 | 實驗環境 . . . . .                     | 42 |





# 第一章 緒論

本章節通過生活中語音辨識存在的問題和引出研究動機，並說明本文後續章節的內容和研究貢獻。

## 1.1 研究動機

在日常生活當中，使用語音辨識的場景往往是多種多樣的。例如在車載語音助手 [21] 中，語音辨識常見詞以目的地名稱、地名、導航指令為主；在點歌語音助手中，語音辨識的常見詞則會是歌曲名、歌手名、歌曲風格等內容；在專業學術會議中 [18]，語音辨識則會碰到學術上的專有名詞。在其中一些語境中，往往不存在相關領域的語音辨識訓練語料，或者語料時長無法獨立訓練一個具有較好魯棒性的端到端語音辨識系統，我們需要使用通用的語音辨識語料訓練端到端語音辨識模型。因為在通用語料中不存在相關語境的專有詞彙，或常見詞頻率與相關語境不同。所以在這些語境中，使用通用的語料訓練的語音辨識模型不能很好的辨識正確常見詞和領域專有名詞。

以 kaldi [42] 為代表的語音辨識方法，語音辨識領域較為成熟的方法。其擁有一個獨立的聲學模型、發音詞典和語言模型。使用 GMM-HMM [51] 或 DNN-HMM [11] 聲學模型進行辨識音素任務後，可以使用 n-gram 語言模型或是其他神經網絡語言模型輸出文字。在訓練語言模型時，根據不同語境的需求，可以通過調整

語言模型訓練資料中語境常見詞的權重來提升常見詞的辨識率。也可以通過修改 WFST 的方法，將設定專有名詞賦予一定的初始分數，以提高領域專有名詞的辨識率 [31]。



現在端到端語音辨識模型相較於傳統的混合模型，可以提供更高的辨識率和更強的環境魯棒性 [55]。但因為其將聲學模型、發音詞典和語言模型全部整合在一起，沒有獨立的語言模型，無法通過訓練和修改語言模型的方法提升語境常見詞和專有名詞的辨識準確度。如何在只有相關語境或專業領域的文字語料或專有名詞的情況下，提升相關語境語句辨識的準確度，是端到端商用級的語音辨識的一個很重要的功能。

在實際的使用場景中，說話人的語句往往具有領域的不確定性 [29]。例如使用智能音箱時，使用者可能需要點播影片，或是想要詢問天氣，而這兩個領域的常見詞和專有名詞往往是不同的。如何對使用者語句的語境進行判斷，也是語境偏移任務非常有必要的一環。有時還需要以通過判斷語境來防止不同領域的相似讀音的專有名詞的干擾。例如點播歌曲時，歌手的名字叫王麟，而點播電影時，電影的名字叫亡靈。如果可以在語音辨識前先確定語句的意圖或是屬於哪個領域，就可以通過對已知的此領域的專有名詞和常見詞進行偏移，達到更好的辨識效果。

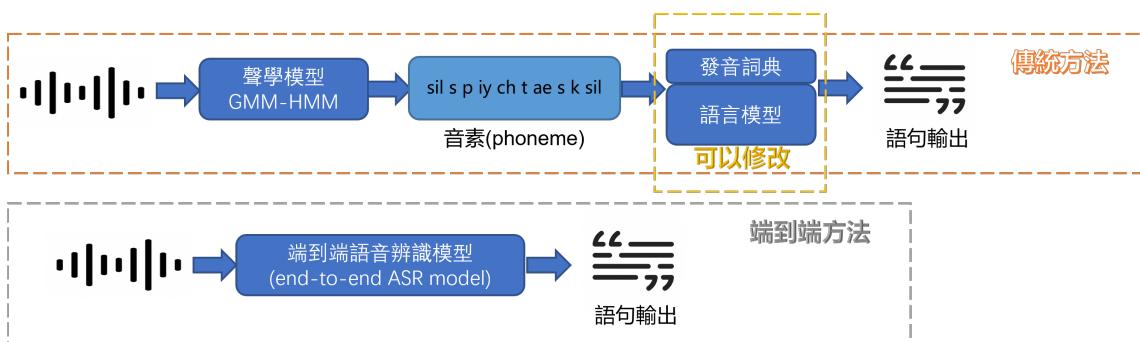


Figure 1.1: 傳統語音辨識流程和端到端語音辨識流程



## 1.2 研究貢獻

本篇文章的貢獻如下：

1. 參照先前的研究在英文上的語境偏移任務創建方法 [15]，基於開源資料集 CATSLU [62] 提出了兩個中文的開源語境偏移任務。
2. 通過預訓練的自監督學習模型進行語境檢測，並通過實驗證明，在算力需求和準確度兩個方面上，使用端到端的意圖辨識方法，要優於傳統的語境檢測的方法。並證明了使用端到端語境檢測方法有效降低了錯誤偏移的發生。
3. 提出了中文備選詞預測模型，為中文不常見專有名詞生成讀音相似易錯詞，並應用於語句後處理階段，通過實驗證明其有效性。
4. 分別在提出的兩個語境偏移任務上，復現了先前的研究在模型訓練、語句解碼兩個環節上語境偏移的方法，並通過實驗證明其有效性。

## 1.3 章節概述

本論文分為六個章節：

1. 第一章為緒論，介紹這篇論文的研究動機和內容大綱。
2. 第二章為文獻探討，通過之前的研究介紹語境偏移這個研究領域的主要目標和現況，探討了端到端語音辨識語境偏移研究的各個方向，介紹經典論文的語境偏移方法。
3. 第三章為資料集和任務介紹，介紹本論文基於中文開源資料集 CATSLU [62] 提出的語境偏移開源任務，以及後續章節中使用的資料集。



4. 第四章為研究方法，介紹本論文在語音辨識的四個階段分別進行語境偏移的方法。
5. 第五章為實驗設計與結果，詳細介紹了本論文在四個方向上設計的實驗，其中包括實驗目的、模型設計、參數設置、資料集選擇，並對比和分析實驗結果。
6. 第六章為結論和未來工作，根據本論文實驗結果總結結論，並提出未來可能的研究方向。



## 第二章 相關文獻

本章節回顧本研究所涉及的語音辨識技術，和語境偏移技術的發展。主要內容包括傳統語音辨識方法、端到端語音辨識方法、語音辨識前對聲音特徵處理的影響、語境偏移先前的研究、語境偏移評量方法。

### 2.1 傳統語音辨識

要了解端到端語音辨識模型為什麼需要語境偏移，就需要從傳統的語音辨識技術開始講起。

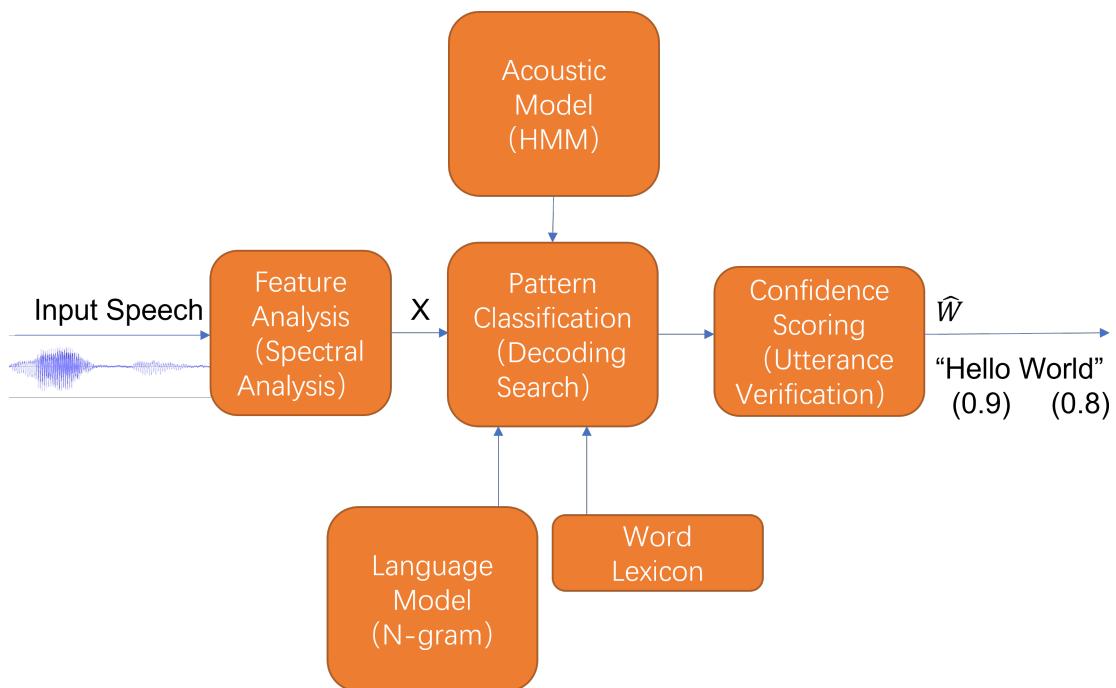


Figure 2.1: 語音辨識系統整體框架圖

如圖2.1所示 [45]。語音辨識技術流程需要經過聲學特徵提取、樣本分類、信度評分三個階段。傳統的語音辨識中，樣本分類階段需要聲學模型(使用 HMM)、語言模型(使用 n-gram)和發音詞典的參與。整個語音辨識系統的輸入是一段語音，輸出是這段語音對應的文字資訊，其中每個詞都有對應的信度。

## 2.1.1 聲學模型

在傳統的語音辨識系統中，聲學模型將輸入的語音特徵向量序列辨識成音素序列。聲學模型主要分為 GMM-HMM 模型和神經網路參與的混合模型，在這篇文章中混合模型將以 DNN-HMM 為例。

### 2.1.1.1 GMM-HMM 聲學模型

在深度學習技術快速發展之前，通過 GMM-HMM 模型作為語音辨識的聲學模型一直是業界的主流方法。在完成特徵提取之後，一段語音信號就變成了一個特徵向量的序列，語音辨識系統需要將特徵向量序列識別輸出一個詞序列，HMM 模型就是為了解決這種序列標註問題而設計的。[17] 在傳統的語音辨識系統中，聲學特徵提取通常為 39 維的 MFCC [36]。它是一個連續的向量，但如果需要輸入後續的 HMM 模型 [14]，我們需要一個離散的觀察序列。我們需要將連續的特徵向量離散分佈，之前的研究有用到矢量量化的方法 [24]，但是這個方法只是用類似聚類的方法來以離散的方式表示連續的資訊分佈，效果並不好。

而後應用到了高斯分佈 [19] 的高斯混合模型 [47] 被用在這項任務中。我們假設均值為  $\mu$ ，標準差為  $\sigma$ ， $O_t$  是輸入的連續變量，則我們可以用如下的高斯分佈來建模  $b_j(o_t)$ ：

$$b_j(o_t) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(o_t - \mu_j)^2}{2\sigma_j^2}} \quad (2.1)$$



高斯混合模型的每個分量的機率密度函數是高斯分佈函數，我們假設有 M 個分量，其中單個分量為 m，則得到以下公式：

$$b_j(o_t) = \sum_{m=1}^M c_{jm} \mathcal{N}(x | \mu_{jm}, \Sigma_{jm}) \quad (2.2)$$

其中  $c_{jm}$  和  $\mu_{jm}$  和  $\Sigma_{jm}$  是 GMM 模型需要學習的參數，我們可以使用已知的數據，用 EM 算法 [41] 進行學習。

在完成對特徵向量的離散化後，就需要一個序列對序列的模型將輸入的離散特徵向量辨識為音素序列，HMM 模型被提出並廣泛應用。一個 HMM 模型包含以下部分：

1. N 個狀態的集合  $Q = \{Q_1, Q_2, \dots, Q_N\}$
2. 狀態轉移矩陣  $A = \{a_{11}, a_{12}, \dots, a_{1N}, \dots, a_{N1}, a_{N2}, \dots, a_{NN}\}$
3. 觀察序列  $O = o_1 o_2 \dots o_T$
4. 觀察概率，表示在第 i 個狀態看到  $o_t$  的概率  $B = b_i(o_t)$
5. 初始狀態  $q_0$ ，結束狀態  $q_F$

一階的 HMM 假設了當前狀態依賴上一個狀態，但與更早的狀態無關，可以表示為：

$$P(q_i | q_1, \dots, q_{i-1}) = P(q_i | q_{i-1}) \quad (2.3)$$

還有一個假設是當前的觀察值只依賴當前的狀態，與之前的狀態和其他觀察無關，可以表示為：

$$P(o_i | q_1, \dots, q_T, o_1, \dots, o_T) = P(o_i | q_i) \quad (2.4)$$



在進行語音辨識時，我們會使用如圖2.2所示的 HMM 模型，我們用 HMM 來對單詞進行建模，可以將發音分為三個階段，開始、中間和結尾。發音只能從開始狀態向下一個狀態進行跳轉，不可逆。HMM 有三個基本問題，第一個問題是

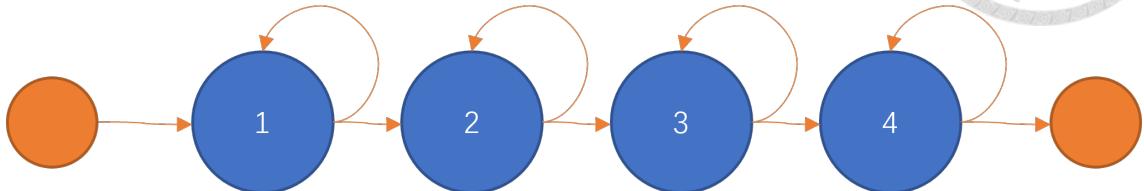


Figure 2.2: 語音辨識 HMM 示意圖

Likelihood，給定 HMM 和觀察序列，計算機率。第二個是 Decoding，給定 HMM 和觀察序列，尋找最優狀態序列。第三個是 Learning，給定觀察序列，尋找最優的 HMM 模型參數。[\[14\]](#) 通過語料對 HMM 模型進行訓練，而後可以使用 HMM 模型對輸入的語音求解最優輸出序列。

#### 2.1.1.2 DNN-HMM 聲學模型

在神經網絡的研究逐漸出現之後，混合模型被提出。其中被廣泛使用並在 kaldi 語音工具包 [\[42\]](#) 中實現的是 DNN-HMM 混合模型 [\[35\]](#)。在 kaldi 中的 Karel [\[54\]](#) 和 Dan 等人實作並開源了此模型 [\[43\]](#)，成為了往後一段時間語音辨識方法的主流選擇。

在這個方法中使用了 DNN 神經網絡模型來代替 GMM 進行特徵的離散化，與 HMM 模型組成了一個混合模型。這個模型的輸入是特徵向量，輸出是上下文相關的 senone，在 kaldi 中實作的是 tri-phone 這種標籤。

為了將訓練資料對齊，並生成對應的 senone 標籤。準備訓練資料的時候需要用已有的語音辨識系統對訓練資料進行對齊，在這裡一般使用 GMM-HMM 作為對齊的系統。準備好訓練資料之後，此任務就成為一個分類任務，可以直接使用 Cross-entropy [\[8\]](#) 損失函數來進行訓練。



## 2.1.2 發音詞典

在語音辨識系統中，發音詞典的作用是將目標語言的字或詞轉換成對應的音素。在中文中，常以詞作為文本的最小單元 [26]。但在聲學模型的輸出中，需要以最小發音單元音素為一個單位，一個詞可能會存在多個音素。所以在準備訓練資料時，會將訓練資料的中文詞轉換為音素的表達形式。而在進行解碼時，發音詞典將和語言模型一起將聲學模型輸出的音素序列轉換為對應的詞。

## 2.1.3 語言模型

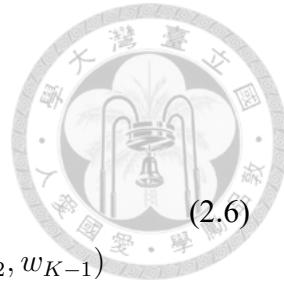
語言模型分為規則模型和機率模型，目前主流使用的是基於機率的模型 [63]。現在應用在語音辨識上的機率模型主要分為 n-gram 語言模型和神經網絡語言模型。語言模型的主要作用是根據聲學模型輸出預測的音素序列，通過音素序列上下文等特徵資訊，通過之前學習的先驗知識，結合發音詞典，產生對應的文字。

### 2.1.3.1 N-gram 語言模型

N-gram 語言模型採用了馬可夫假設 [13]，就是第  $i$  個詞出現的機率，和先出現前  $i-1$  個詞先出現再出現第  $i$  個詞的機率相關。使用我們只要統計前  $i-1$  個字的出現後出現第  $i$  個字的次數即可，這就是最大似然估計 [48]。

$$P(w_i | w_1, w_2, \dots, w_{i-1}) = \frac{\text{count}(w_1, w_2, \dots, w_i)}{\text{count}(w_1, w_2, \dots, w_{i-1})} \quad (2.5)$$

在實際的應用中，因為可能存在非常長的句子，我們不能無限度的向前看來統計機率，我們通常會設置一個閾值來限制向前看的步數。通常會選擇向前看三個詞，



這就是 tri-gram 語言模型。

$$\begin{aligned}
 P(W) &= P(w_1) P(w_2 | w_1) \dots P(w_K | w_1, w_2, \dots, w_{K-1}) \\
 &= P(w_1) P(w_2 | w_1) P(w_3 | w_1, w_2) \dots P(w_K | w_{K-2}, w_{K-1})
 \end{aligned} \tag{2.6}$$

而現在受計算機性能進步的影響，在很多應用中會選擇 4-gram 甚至更多的向前看詞數。

### 2.1.3.2 神經網絡語言模型

神經網絡語言模型 [3] 解決了 n-gram 語言模型存在的兩個問題。第一個問題是詞的泛化能力，無法辨識相似詞義的詞，比如在訓練資料中只存在「小明喜歡吃紅鯉魚」的情況下，比較難解碼出「小紅喜歡吃草魚」這樣的句子。第二個問題是在 n-gram 中解碼單詞的機率只取決於前 n 個詞，n 通常為 3 或者 4，但是神經網絡語言模型可以根據更多之前生成的詞去推斷此詞的內容，或者根據之後產生的內容去修正之前的內容。

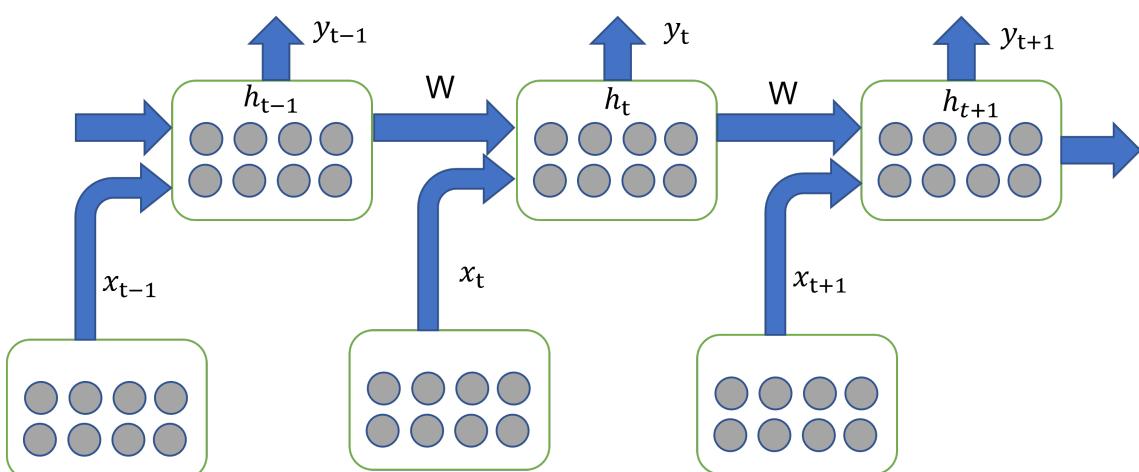


Figure 2.3: RNN 語言模型示意圖

現在在語言辨識上應用最多的神經網絡語言模型是 RNN 語言模型和 Transformer 語言模型。



RNN 源自前饋神經網絡 [50]，特點是使得某些節點的輸出會影響後續節點的輸入，是一個序列到序列的神經網絡模型。如圖2.3所示的 RNN 結構，其中  $t$  是指某一時刻， $x$  指此時刻的輸入， $h$  指隱藏層， $y$  代表輸出， $W$  代表權重矩陣。前一個時刻的隱藏層通過先和權重矩陣相乘，再和輸入相乘得到這一時刻的隱藏層，如下所示：

$$h_t = \sigma(W^{(hh)}h_{t-1} + W^{(hx)}x_{[t]}) \hat{y} = \text{softmax}(W^{(S)}h_t) \quad (2.7)$$

通過輸出的  $y$  來和詞庫中的詞計算 word embedding，就可以得到需要輸出的詞。

[39]

Transformer 語言模型 [30] 也被廣泛使用。Transformer 用自注意力模塊 [53] 進行建模，解決了 RNN 模型不能並行運作的問題，同時通過注意力模塊可以注意到輸入序列中任意兩個位置的匹配與差異，帶來了語言模型更好的表現。transformer 語言模型採用多層編碼器和解碼器架構，在訓練資料充足的情況下，表現優於 RNN 語言模型。[63]

#### 2.1.3.3 評估指標

評價語言模型的優劣的指標主要是困惑度 (PPL)。困惑度是一個句子中所有組成句子的詞概率的平均導數，假定一個句子為  $w_1 \dots w_n$ ，計算公式如下：

$$\sqrt[n]{\frac{1}{P(w_1 \dots w_n)}} = P(w_1 \dots w_n)^{-\frac{1}{n}} \quad (2.8)$$

## 2.2 端到端語音辨識

這一節講述了端到端語音辨識的發展，在發展過程中遇到的問題和改進方法。



### 2.2.1 CTC 自動對齊

這個研究提出了一種訓練方法，使得在模型訓練時可以將輸入序列和輸出序列自動對齊，直接從未標記位置的輸入序列中輸出結果，是端到端語音辨識系統的重要組成部分。這項研究在原先輸出的 label 集合之外又定義了一個 unit，通過輸出  $L$  個 units 來定義在某時刻觀察到某 label 的機率。每個 label 的機率可以看做是不同對齊可能性下的所有機率的累加 [22]。分類器的輸出就是輸入序列最可能的 label 序列，設輸入為  $\mathbf{x}$ ，長度為  $T$ ，字符集  $L$ 。可以用以下公式定義：

$$h(\mathbf{x}) = \arg \max_{\mathbf{l} \in L^{\leq T}} p(\mathbf{l} \mid \mathbf{x}). \quad (2.9)$$

同時，這項研究也提出了兩種解碼方法，分別是最佳路徑解碼和前綴搜索解碼 [22]。在多數端到端語音辨識系統進行解碼時就使用了 CTC 的解碼方法。

### 2.2.2 序列到序列模型

端到端語音辨識系統最早被提出是將序列到序列的方法應用在語音辨識上 [7]。這項研究參考了在端到端機器翻譯 [2] 中提出的模型，提出了一個由 RNN 組成的編碼器-解碼器。其中編碼器將語音信號轉換為特徵向量序列，解碼器通過給定的特徵向量和之前的文字輸出確定下一個字符的機率分佈，將特徵向量轉換為句子。模型架構如圖2.4所示。

端到端的學習方式摒棄了繁雜的發音詞典和語言模型的訓練解碼過程，也簡化了聲學模型的建模過程。使得訓練過程不再需要先訓練一個傳統的聲學模型對訓練資料進行對齊，與語言模型分開訓練。解碼時也不需要兩步驟進行。但是在這個端到端模型上的準確度表現，是要差於傳統的混合神經網絡聲學模型 + 語言

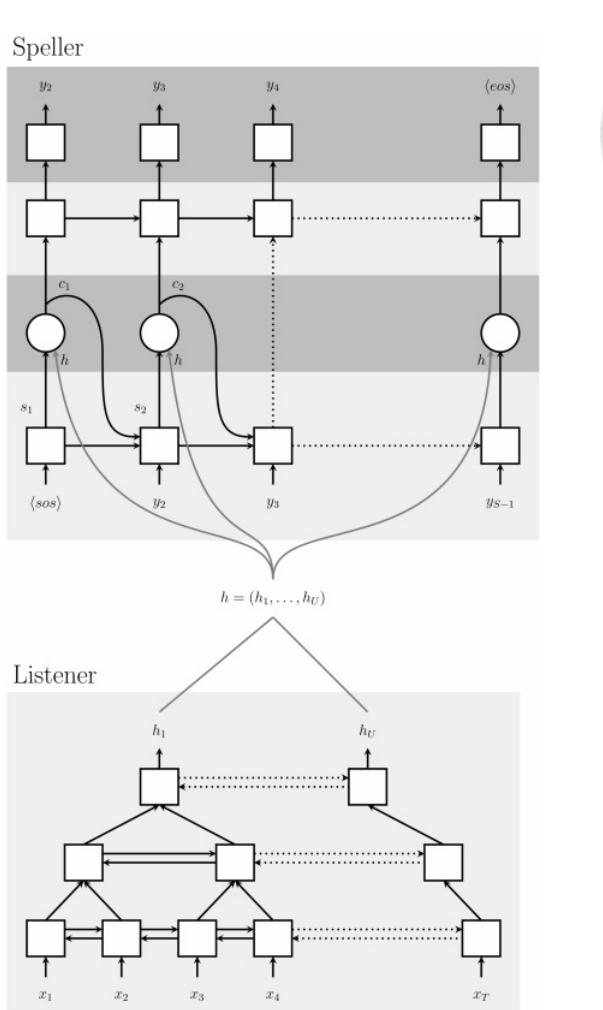


Figure 2.4: RNN 端到端語音辨識模型示意圖 [7]

模型的。在之後對於端到端語音辨識的研究當中，進行改進以提高模型準備度。

### 2.2.3 Conformer 端到端語音辨識模型

在先前的序列到序列的語音辨識模型提出之後，如何提高模型的準確度成為了後續研究的目標。Transformer 因為其可以對全局特徵的關聯性進行捕捉，所以基於 Transformer 的端到端語音辨識技術被提出。通過使用 Transformer 模塊來替代之前的 RNN 模型，實現了超過傳統混合模型的準確度 [60]。

但是 Transformer 只關注全局特徵的關聯性，而無法基於特定的局部特徵更高的權重。CNN 可以關注局部特徵，但做不到全局特徵的關注。在之後的研究中，

Conformer 被提出並用於語音辨識 [25]。通過在 Transformer 中加入了卷積模塊，如圖2.5，將 CNN 的特性和 Transformer 的特性相融合，在語音辨識的準確度上取得了更好的表現。且更少參數量的小型端到端模型也具有很高的準確度 [25]。

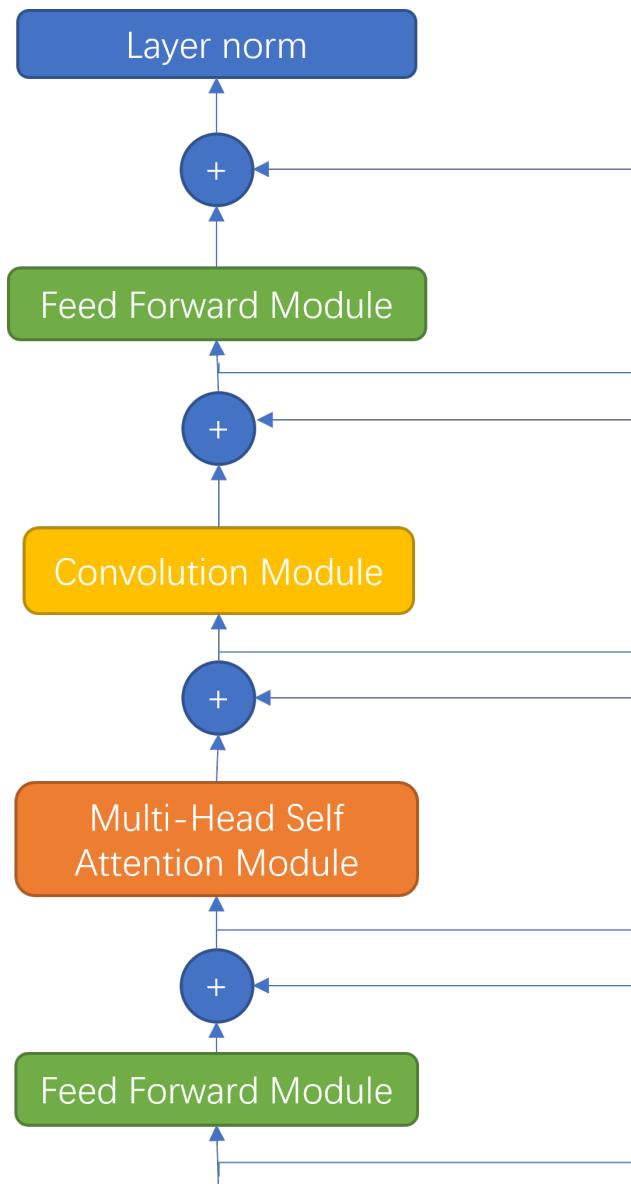


Figure 2.5: Conformer 模型示意圖

## 2.3 聲學特徵

本節內容主要介紹語音辨識中常用的聲學特徵及其提取方式



### 2.3.1 梅爾頻率倒譜系數 (MFCC)

在傳統語音辨識方法中，MFCC 特徵是使用最普遍的。MFCC 語音特徵是模擬人耳對聲音的感知能力，提取 MFCC 特徵需要通過以下幾個步驟 [36]：

1. 將語音信號分解為多個幀
2. 採用週期圖法估算功率譜
3. 進行傅立葉變換 [5]，將訊號轉換到頻域上
4. 採用 Mel 濾波器組對功率譜進行濾波，計算每個濾波器的能量
5. 取得每個濾波器的能量對數值
6. 執行離散餘弦轉換 (DCT)
7. 保留第二至第十三個 DCT 係數，捨棄其餘係數

將每一幀的能量作為一個特徵並加上 12 個 DCT 係數，構成了 13 維的 MFCC 語音特徵。若需要使用 39 綴的 MFCC，則是計算 13 綴的 MFCC 的一次微分和二次微分。MFCC 特徵降低了輸入信號的維度，提高了信號對高頻干擾的魯棒性。

### 2.3.2 FilterBank 語音特徵 (Fbank)

在部分端到端的語音辨識任務或其他語音任務中，需要輸入的特徵保留更多的資訊，Fbank 被提出並使用。Fbank 特徵只是省略了 MFCC 特徵提取的最後一步，不進行 DCT 變換，保留了 80 綴的特徵向量。在實際的應用中，Fbank 特徵被廣泛應用於各種語音任務中，包括語者辨識、情緒辨識等。



### 2.3.3 使用自監督預訓練模型提取語音特徵

在語音辨識任務中，除了使用一定規則來提取聲學特徵，還可以使用自監督學習的預訓練模型提取語音特徵。目前語音領域的開源中文自監督預訓練模型主要有 HuBERT [27] 和 wav2vec2.0 [1]。在使用相同的下游模型 Conformer 的情況下，使用預訓練的 HuBERT 或 wav2vec2.0 替代 Fbank 聲學特徵作為模型訓練的輸入，可以得到更好的效果 [52]。

在 SUPERB [56] 的研究中，提出了一套評量自監督訓練模型優劣的方法。使用最新提出的自監督訓練模型進行的實驗表明，在幾乎所有的語音任務中，使用自監督預訓練模型進行語音特徵抽取，比使用 Fbank 聲學特徵具有更好的效果。在關於意圖辨識的輸入特徵選擇的研究中 [4] 提出，使用自監督預訓練模型抽取的語音特徵作為意圖辨識的模型輸入，與使用 ASR 模型相比有更好的效果。

但目前使用自監督訓練模型進行特徵抽取需要花費的算力資源過大，很難在沒有顯示卡的終端設備進行，所以其應用還具有局限性。但是在雲端算力充足的環境中，可以最大化的提升模型的性能。

## 2.4 語境偏移

因為在端到端語音辨識中，沒有獨立的語言模型，無法通過單獨調整語言模型的方式使語音辨識向某個領域偏移。但一個成熟的商用 ASR 系統需要具有根據不同的語境去調整不同詞彙出現機率的能力，所以在端到端語音辨識出現後，有較多的研究嘗試從不同的方向上對語音辨識系統進行語境偏移。

這一節會講到目前應用在端到端語音辨識系統上的語境偏移方法，其中包括在語音辨識的多個階段上的方法，同時簡述了各個方法的發展歷程，最後根據文

獻總結評量方法。



### 2.4.1 清洗訓練資料

在 google 進行的一項研究中 [28]，提到可以進行資料清洗來進行語境偏移，將文字搜索的文本訓練資料來用在語音搜索的語音辨識系統上。這篇論文提出了三種簡單的數據選擇策略，用於減少語言模型訓練所需的語料庫的大小，並提高稀有詞彙的識別質量，而不損害整體性能。這些策略分別是：

1. 使用軟對數函數逐漸降採樣高頻句子，以減輕語料庫的重頭現象
2. 明確地過濾聲學數據中包含稀有單詞的句子，以增加它們在訓練中的表示
3. 使用困惑度為基礎的對比數據選擇，只過濾與目標領域（語音搜索）最相似的查詢

如圖2.6所示。作者在一部分匿名化的谷歌搜索流量上進行了數據選擇，並獲得了  
一個比原始語料庫小 53 倍的子集，同時在數據分布的頭部和尾部都改善了結果。  
通過淺融合，經過數據選擇的語言模型相比於在原始語料庫上訓練的語言模型，  
在稀有詞彙上實現了高達 24% 相對 WER 降低，同時保持了整體 WER 不變。

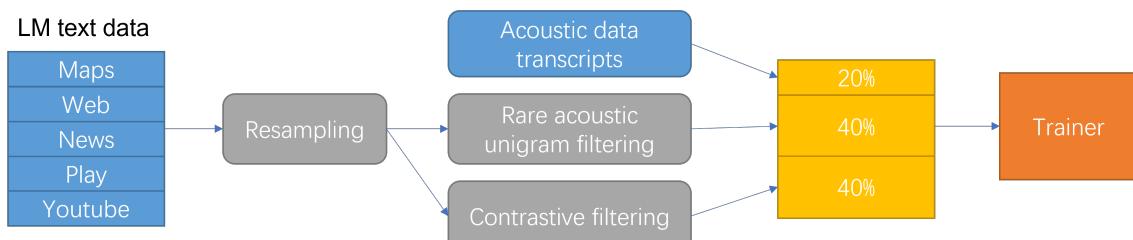


Figure 2.6: 清洗訓練資料的方法示意圖



## 2.4.2 淺融合

在端到端的語音辨識中，在解碼階段，使用語言模型參與打分的方法稱為淺融合。最早由 2018 年的一項研究提出 [32]，這篇論文提出了一種方法，即淺融合技術，在每個搜索步驟中使用對數線性插值來結合外部語言模型和內部語言模型（即解碼器）的概率分佈。這樣可以增強輸出序列的流暢性和準確性。相比於深度融合技術，淺融合技術有以下優點：

1. 不需要修改原始的序列到序列模型結構或重新訓練它
2. 可以靈活地調整外部語言模型和內部語言模型之間的權重
3. 可以方便地更換不同類型或來源的外部語言模型。

這項研究比較了在不同類型的語言模型（神經網絡和 n-gram）、不同的解碼單元（詞、字符和詞片）和不同的任務（Google Voice Search 和 Switchboard）上，淺融合技術的效果。在 Google Voice Search 上，使用神經網絡語言模型和詞片作為外部語言模型，透過淺融合技術，相較於基準序列到序列模型，實現了 9.1% 的相對詞錯誤率降低 (WERR)，而不需要進行第二遍重評分 (rescoring)。但是這項研究提出的方法是在 word 單元進行淺融合的，如果部分詞彙在訓練資料中沒有出現，就會在 beam search 前被提前剪枝，無法對其進行偏移。

在 2019 年的一項研究中 [61]，通過在 subword 單元上執行淺融合，並在剪枝前執行來避免罕見詞在 beam search 時被剪枝。同時，為了減少對不需要執行偏移的文本的影響，這項研究嘗試使用前綴詞觸發的方法（如「call」、「text」），出現前綴詞才執行偏移。這項改進過後的淺融合技術，在之後的端到端語音辨識系統中被廣泛使用，例如開源工具 wenet2.0[59] 中就有對其進行整合。



### 2.4.3 替代拼寫預測

在完成解碼後的後修正階段，也有進行語境偏移的方法。可以通過收集 ASR 輸出的錯誤詞彙，訓練一個替代拼寫預測模型，用來生成目標偏移詞的替代拼寫 [15]。

這項研究的方法是使用一個基於 CTC 和注意力機制的端到端語音辨識模型，結合兩種不同的解碼算法，分別是 CTC 前綴束搜索和 WFST 解碼。在解碼過程中，作者使用淺層融合的方式，將用戶定義的上下文偏置詞列表編譯成 WFST 表示，並與語音辨識模型進行組合，以增加偏置詞的權重。

為了改善稀有詞和詞表外詞的識別效果，作者提出了一種替代拼寫預測模型，利用語音辨識模型和訓練數據生成偏置詞的替代拼寫，並將它們添加到偏置 WFST 中。這種方法不需要額外的發音字典或語音合成系統。這項研究的方法顯著的改善了罕見詞和 OOV 詞的準確率，為語境偏移的方法提供了一個新的思路。

### 2.4.4 CLAS 基於注意力機制的語境偏移

在一項名為 CLAS [44] 的研究中，提出了一種基於注意力的語境偏移方法。在推理期間，CLAS 系統可以處理可能包含訓練期間未見過的 OOV 術語的上下文短語。該方法不需要在訓練時提供特定的上下文信息，並且不需要仔細調整重新評分權重，同時仍然能夠包含 OOV 術語。在多項任務中，CLAS 系統相對於基線方法的 WER 相對提高了 68%，表明聯合優化比單獨訓練組件更具優勢。該方法的優點是可以動態地將上下文信息納入識別過程中，而不需要在推理期間使用外部獨立訓練的 LM 進行 OTF 重新評分。

根據論文中的實驗結果來看，這種方法的缺點是在處理大量的偏移詞時，如



圖5.4，其偏移效果會顯著降低。而在實際的應用當中，是需要對大量可能存在的專有名詞進行偏移的。

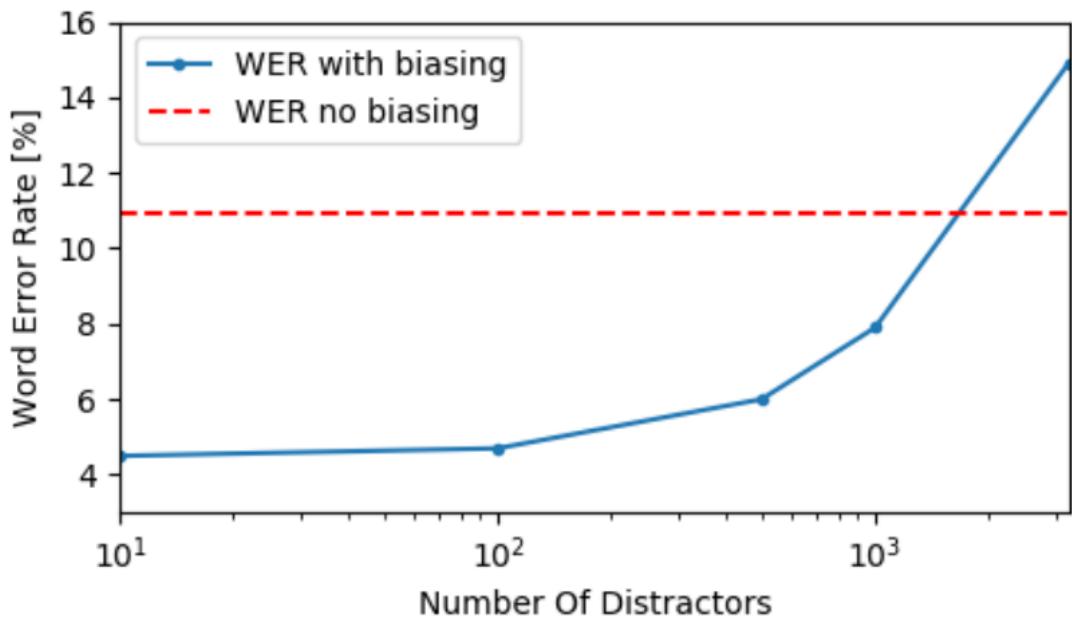


Figure 2.7: CLAS 偏移關鍵詞數量對於 WER 的影響 [44]

## 2.4.5 Whisper 的語境偏移

在目前魯棒性最強的開源模型 Whisper [46] 上，也具有語境偏移的方法。該方法仍然屬於基於注意力的方法。我們可以給即將輸入的聲音創造上文，作為 Prompt 輸入 Whisper 模型中。模型在進行 decode 的時候，就會參考上文中出現的專有名詞、語境、語言、語法結構等內容。這樣的方法如果可以提前知道輸入的語音屬於哪個領域，就可以輸入此領域可能的上文，以增加模型解碼的正確性。

但是這種方法具有局限性，我們無法給定大量的專有名詞令其偏移，同時如何去設計 Prompt 也是一個可能需人工去完成的內容，Prompt 的內容可能會很大程度的影響後文的語境偏移效果。



#### 2.4.6 語境偏移相關任務

進行語境偏移的方向和方法有很多，如何通過統一的標準對其進行評價，在各個研究中大都按照自己的方式進行。例如在亞馬遜的研究中，採用的都為內部資料集應用不同的方法進行比較 [20]。而在 google 的研究中，也是選擇自己的文字搜索資料集和語音搜索資料集 [28]。

這樣的非開源資料無法用來在其他研究中復現這一結果，不是我們想要的方法。而如何在一個開源資料集上設計語境偏移任務，有一項研究給出了很好的思路 [15]。研究採用了包含大量高管名字、公司名稱的上市公司電話會議的資料集 Earnings21 [9]，並使用 spaCy [21] 對其中的專有名詞進行標記，作為偏移的目標關鍵詞。並加入了干擾項，也就是測試集中不出現但是存在目標關鍵詞列表中的詞。

通過使用開源的 ASR 模型，並應用不同的語境偏移方法，對測試集進行解碼。並將解碼結果分為四類：單詞、短語、稀有詞、OOV 詞，根據四類詞的解碼結果的表現來評價語境偏移方法的優劣。在我們後續的研究中，設計中文的開源語境偏移任務就參考了這種做法，保證我們的結果可以被後續的研究人員復現出來。





## 第三章 資料集和任務介紹

### 3.1 資料集介紹

本章節會介紹本研究所使用的資料集的詳細資訊。

#### 3.1.1 CATSLU 資料集

CATSLU 資料集是一個口語理解資料集，包含音頻和文本訊息，涵蓋了四個對話領域：地圖導航、音樂搜索、天氣預報和視頻（電影和電視）搜索。這些資料來自真實對話系統，包括不同用戶的語音信號、自動語音辨識文本和手動轉錄文本。每個用戶的語句都有相應的語義標註，使用一組  $act(slot=value)$  來表示。該資料集旨在探索如何利用多模態信息和領域知識來提高口語理解的性能和魯棒性。該資料集分為兩個場景：單領域口語理解和跨領域適應口語理解。[\[62\]](#)

CATSLU 資料集包括了 521 個使用者、4 萬多條對話記錄和超過 10 萬個槽位值。該資料集的總長度約為 30 小時，平均每條記錄長度為 2.7 秒。CATSLU 資料集是第一個公開的中文口語理解資料集，也是第一個同時包含音頻和文本特徵的口語理解資料集。詳細資訊見表3.1。



Table 3.1: CATSLU 資料集各領域詳細資訊

| 域 (Domain) | 說話人數量 (人) | 訓練集 (句) | 驗證集 (句) | 測試集 (句) |
|------------|-----------|---------|---------|---------|
| 地圖導航       | 1788      | 5093    | 921     | 1578    |
| 音樂搜索       | 268       | 2189    | 381     | 676     |
| 天氣預報       | 276       | 341     | 378     | 2660    |
| 影片搜索       | 227       | 205     | 195     | 1641    |

### 3.1.2 WenetSpeech 中文語音辨識資料集

WenetSpeech 是一個大規模的多領域中文語音辨識資料集，包含以下部分 [58]：

- 超過 10000 小時的高質量標註語音資料，涵蓋 10 個領域，包括有聲書、評論、紀錄片、戲劇、訪談、朗讀、談話、綜藝和其他。這些資料來自 YouTube 和 Podcast，具有多樣的說話風格、場景、主題和噪聲條件。
- 超過 2400 小時的弱標註語音資料，主要來自 Podcast，使用高質量的自動語音辨識系統進行自動轉寫，再使用一個端到端的標籤錯誤檢測方法進行篩選和過濾。
- 超過 10000 小時的無標註語音資料，可用於半監督或無監督的語音表示學習。
- 三個人工標註的高質量測試集，用於評估語音辨識系統的性能。其中 Dev 是用於訓練過程中交叉驗證的測試集；Test Net 是從網絡收集的匹配測試集；Test Meeting 是從真實會議錄製的更具挑戰性的不匹配測試集。

WenetSpeech 資料集是目前最大的開源中文語音辨識資料集之一，利用光學字符識別和自動語音辨識技術生成候選的音頻/文本對，並提出一種新穎的端到端標籤錯誤檢測方法來進一步驗證和過濾候選對。該方法使用預訓練的 wav2vec 2.0 模型作為特徵提取器，並在其上構建了一個基於 CTC 損失函數和編輯距離閾值策



略的二分類器。該方法能夠有效檢測低質量或不匹配的候選對，並保留高質量或匹配的候選對。

WenetSpeech 資料集旨在為研究社群提供一個具有領域多樣性和規模優勢的中文語音辨識資源，以開發更通用和魯棒的自動語音辨識系統。該資料集已在 CC-BY 4.0 協議下開源，並提供了基於 Kaldi、ESPnet 和 WeNet 三個流行工具包構建的基準系統。這些基準系統在三個測試集上都取得了令人滿意或超越現有開源中文資料集上最佳結果。

Table 3.2: WenetSpeech 各領域語句時長分佈

| Domain      | Youtube | Podcast | Total  |
|-------------|---------|---------|--------|
| audiobook   | 0       | 250.9   | 250.9  |
| commentary  | 112.6   | 135.7   | 248.3  |
| documentary | 386.7   | 90.5    | 477.2  |
| drama       | 4338.2  | 0       | 4338.2 |
| interview   | 324.2   | 614     | 938.2  |
| news        | 0       | 868     | 868    |
| reading     | 0       | 1110.2  | 1110.2 |
| talk        | 204     | 90.7    | 294.7  |
| variety     | 603.3   | 224.5   | 827.8  |
| others      | 144     | 507.5   | 651.5  |
| Total       | 6113    | 3892    | 10005  |

### 3.1.3 AISHELL-1

AISHELL-1 是一個由北京貝殼科技有限公司釋出的開源中文普通話語音語料庫，主要用於語音辨識研究和系統構建。

AISHELL-1 包含了來自中國不同口音地區（包括普通話、粵語、上海話等）的 400 名人員在安靜室內環境下使用高保真麥克風錄制的約 178 小時（約 150 萬字）不同主題內容（如智能家居、無人駕駛、工業生產等 11 個領域）及其相應手工轉寫與標注信息。數據採樣率為 16kHz，轉寫準確率為 97% 以上。數據集分為訓練集、開發集和測試集三部分，分別包含 120098、14326 和 7176 個句子。數據



集還提供了與轉寫對應的詞典文件。[6]

AISHELL-1 為 AISHELL 系列資料集中第一個發布的版本，也是目前最廣泛使用的版本之一。它已被多篇論文引用，並在多個評測任務中作為基準或對比資料集使用。例如，在 2018 年舉辦的第十二屆國際漢語自動語音辨識評測（CCLASR 2018）中，AISHELL-1 作為訓練資料之一提供給參賽者；在 2020 年舉辦的第十四屆國際漢語自動語音辨識評測（CCLASR 2020）中，AISHELL-1 作為測試資料之一用於評估參賽系統。

### 3.1.4 AISHELL-2

Aishell-2 資料集是一個開源的大規模語音語料庫，專為普通話語音辨識研究而設計，由希爾貝殼基金會發布。此資料集包含 1000 小時的高品質普通話語音資料，經由 iOS 裝置錄音而成。這些數據免費提供給學術研究社群使用。

本研究錄製的語音來自 1991 名說話者，其中包括 845 名男性和 1146 名女性，年齡範圍在 11 歲至 40 歲以上之間。所有說話者理論上應以標準普通話讀出所有要錄製的內容，但有些微小的口音差異存在。根據口音特徵分類，1293 名說話者使用北方口音，678 名說話者使用南方口音，20 名說話者使用其他口音。在參與者中，有 1347 名說話者在錄音棚內進行錄製，其餘則在自然混響的客廳中進行。錄音內容包含 8 個主要主題，分別是語音命令（如物聯網設備控制和數字順序輸入）、旅遊景點、娛樂、金融、科技、體育、英文拼寫和沒有特定主題的自由發言。[12]



## 3.2 任務介紹

按照先前的研究 [15] 在英文資料集 Earnings21 上提出的語境偏移任務評量標準，我使用中文資料集 CATSLU 作為語境偏移的任務資料集，因為其具備各類專有名詞、且具備關鍵詞列表，同時有 16khz 的音檔和文字轉錄檔。同時根據不同的應用場景，我提出了以下兩個語境偏移任務。

### 3.2.1 CATSLU 專有名詞語境偏移

在這個任務中，只包含特定專有名詞出現的語句，根據專有名詞的不同，分為三個子任務。

1. 含影片名語句語境偏移
2. 含歌手名語句語境偏移
3. 含城市名語句語境偏移

其中每個任務都有自己獨立的關鍵詞列表，關鍵詞列表是已知的，作為語境偏移的目標關鍵詞。但並不是所有在列表中的關鍵詞都會在語句中出現，我們把沒有出現的關鍵詞稱為干擾項。三個子任務的語句和關鍵詞數量如圖 3.1 所示。

這個任務為零樣本學習任務，不提供訓練資料，需要使用統一的開源模型進行測試。開源模型使用的是在 Wenet [57] 上的 Conformer 默認配置，使用資料集 WenetSpeech 中 10000 小時強標記語料訓練而來。設計這個任務的目的是測試語境偏移方法針對專有名詞的偏移能力。

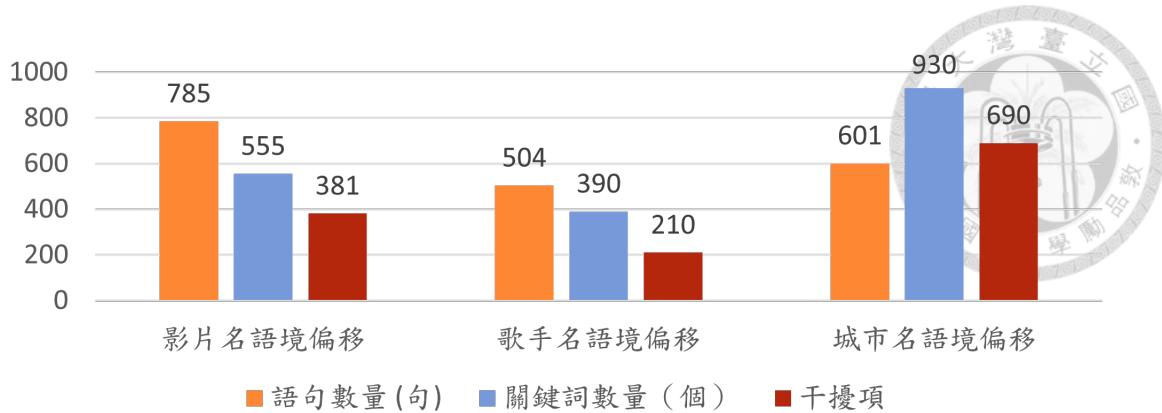


Figure 3.1: CATSLU 專有名詞語境偏移各類別語句數量

### 3.2.2 CATSLU 多領域混合語料語境偏移

在這個任務中，包含影片搜索、音樂搜索、天氣詢問三個領域的所有語句，並採用 class balanced 的方法將其劃分為訓練集和測試集，其中訓練集佔比 80%，測試集佔比 20%。詳細資訊如表 3.3。

Table 3.3: CATSLU 多領域混合語料語境偏移數據集劃分

| /        | 訓練集  | 測試集  |
|----------|------|------|
| 語句數量 (句) | 4191 | 1050 |
| 時長 (小時)  | 3.6  | 0.9  |

在這個任務中，在使用 10000 小時 WenetSpeech 語料訓練的開源模型的基礎上，可以使用訓練集的資料和關鍵詞列表進行語境偏移。關鍵詞列表為領域常見詞，例如「播放」、「導航」，但並不一定是專有名詞。在測試時，測試資料集是混合的，並不會知道屬於哪個領域。這個任務更接近現實中碰到的使用需求，目的是測試語境檢測方法對語境偏移的影響，同時檢測方法的領域相關語境偏移性能。



### 3.2.3 評量方法

#### 3.2.3.1 字錯誤率

通過計算所有句子解碼完成之後，和標準答案比對的 CER 可以評估語境偏移方法對整體的影響，是否會因為偏移過度導致關鍵詞之外的詞錯誤率過高。CER 數值越低，說明在語料中錯誤的字越少。我們設 S 表示替換的字符數目，D 表示刪除的字符數目，I 表示插入的字符數目，N 表示參考序列中字符總數。CER 的計算公式如下：

$$CER = \frac{S + D + I}{N} \quad (3.1)$$

#### 3.2.3.2 關鍵詞錯誤率

為了評估不同的語境偏移方法在語音辨識任務上的優劣，我們計算了語料中所有出現的關鍵詞在識別結果中的錯誤率。關鍵詞錯誤率是指原先正確語句中所有存在的關鍵詞被錯誤預測的比例，也就是 1-召回率 (Recall)。





## 第四章 研究方法

在這個研究中，我們試圖透過在語音辨識的四個階段進行語境偏移來降低目標關鍵詞的錯誤率 (KER, keyword error rate)，並且降低特定領域的語句字錯誤率 (character error rate, CER)。我們的語境偏移分成四個階段，分別為語音辨識開始前、辨識模型、解碼時和解碼後修正。

在語音辨識開始前，我們通過端到端意圖辨識的方法進行語境檢測，保證偏移方向的正確。在語音辨識模型上，我們選擇微調預訓練模型的方法，以適應領域特徵和環境特徵。在解碼時，我們使用中文淺融合的方法，給予目標關鍵詞更高的獎勵分數。在解碼完成後，我們使用訓練的模型生成目標關鍵詞的備選詞表，並進行後修正的語境偏移。

在這個章節中，我們會詳細介紹這四種方法，並在之後的章節通過實驗證明其有效性。

### 4.1 端到端意圖識別

因為語音辨識系統的應用場景豐富多樣，輸入的語音領域可能具有不確定性，在未知語音領域的情況下，我們不能確定偏移的方法。我們嘗試用意圖識別的方法，進行語境檢測。目的是確定輸入的語音屬於哪個領域，為後續的偏移方



法指引正確的方向。

#### 4.1.1 整體流程

我們採用端到端的方式進行意圖辨識，這和傳統的方法是不一樣的。傳統的方法往往會先使用 ASR 系統進行語音辨識，再將辨識輸出的文字輸入到語言模型中進行意圖辨識。但是有研究證明，通過端到端的方法對訓練資料的需求量更少，且具有更高的準確度 [4]。因為語境偏移問題往往都缺乏相關語境的訓練資料，使用一般的 ASR 模型不能給出很好的解碼效果。所以我們採用端到端的方法進行意圖辨識，以確定需要偏移的語境。

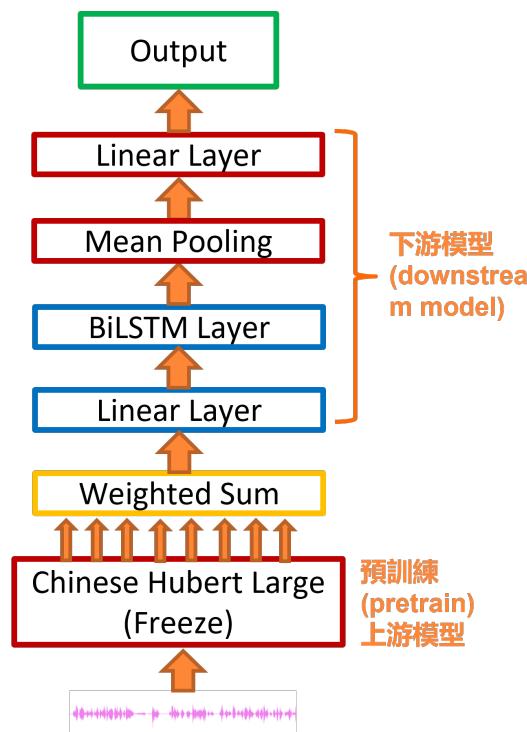


Figure 4.1: 端到端意圖辨識流程示意圖

如圖4.1所示。首先，長度超過 5 秒的音檔，取輸入音檔的前 5 秒輸入預訓練上游模型 HuBERT [27] 提取語音特徵。然後，將 HuBERT 輸出的每一層的特徵向量以 weighted sum 的形式輸入下游模型，其中 weight 值與下游模型一起學習。下游模型是一個由 Bi-LSTM [23] 組成的模型，最後由下游模型直接輸出意圖。



#### 4.1.2 HuBERT 進行特徵提取

在自監督語音模型中，HuBERT [27] 是一種使用聚類方法對音訊資料進行分類和標記的自監督學習模型。中文版本的 HuBERT 使用了 WenetSpeech [58] 進行訓練，使其能夠對輸入的語句進行有效的特徵提取。與傳統的深度學習模型不同，自監督學習模型不需要大量的標注資料，這使得它能夠更好地處理語音語料中的噪聲和變異。

HuBERT 使用類似 BERT 的遮罩方式進行訓練，將部分標籤作為需要預測的對象，從而使模型得到訓練。不同的是，HuBERT 使用了聚類方法對音訊資料進行分類和標記，從而使模型能夠自動學習到音訊的特徵和模式。與傳統的深度學習模型相比，自監督語音模型的訓練過程更加自然和有效。

為了更好地利用 HuBERT 提取的特徵，本研究參考 SUPERB[56] 的研究，將 HuBERT 每一層的輸出進行加權求和後再進行下游模型的訓練。這種方法能夠最大程度地利用 HuBERT 提取的特徵，提高下游模型的性能。在加權求和的過程中，權重值與下游模型一起訓練，使得模型可以自動選擇要提取 HuBERT 模型中特定層的資訊。這種方法能夠有效地解決深度神經網絡中存在的過度擬合和資訊傳遞不良等問題，提高下游模型的泛化能力。

#### 4.1.3 意圖辨識下游模型

意圖分類下游模型的功能，是判斷給定的特徵向量屬於哪一類意圖。輸入是由 HuBERT 上游模型抽取的音檔的特徵向量，輸出是每個類別的機率。在進行意圖分類的時候，我們可以根據任務的複雜程度選擇合適的下游模型架構。

當資料量較少時，可以考慮選擇 Linear model 作為下游模型的架構，通過

mean pooling 層對特徵進行提取後，直接輸入 Linear 進行分類預測，如圖4.2所示。

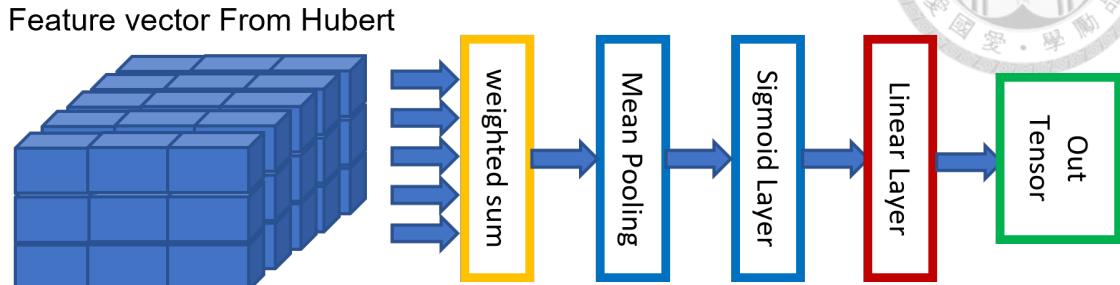


Figure 4.2: 線性下游模型示意圖

但在資料量充足時，選擇 Bi-LSTM 架構的下游模型可以得到更好的效果。為了減少參數量，我們通過 Linear Layer 將最後一維特徵數降到 256。為了考慮上下文的連續特徵，我們將線性模型的輸出輸入一層雙向長短記憶網路，隱藏層大小為 256。為了進一步提取意圖特徵，再進行核大小為 (3,3) 的均值池化，最後經過線性層後輸出。

這樣做的優點是可以直接從輸入的語句中提取意圖資訊，不需要進行 ASR 轉文字，因為轉文字時可能會受到 ASR 模型辨識結果的影響。

## 4.2 微調預訓練模型

如果有少量目標領域的語料，微調原始的 ASR 模型是一個好方法。

少量的語料無法訓練一個完整的 ASR 系統。所以使用已經訓練好的，且具有一定魯棒性的端到端 ASR 模型作為模型的初始參數，然後通過調整訓練的步數和學習率，可以決定領域相關的語料對原始模型的影響的程度。一般為了最大化領域偏移效果，通過設定驗證集的方法，反復迭代訓練，選擇在驗證集上表現最好的模型作為微調後使用的模型。

訓練資料較多時，往往需要足夠訓練步數才能達到最優的效果。而訓練資料較少時，只需要少量的步數，就可以完成對原始模型的微調。



### 4.3 中文淺融合

很多語境偏移任務中並不包含訓練語料，只有一些需要偏移的關鍵詞，且不能對每個小領域都訓練一個專有的語音辨識模型。所以我們選擇在解碼階段應用淺融合的方法。

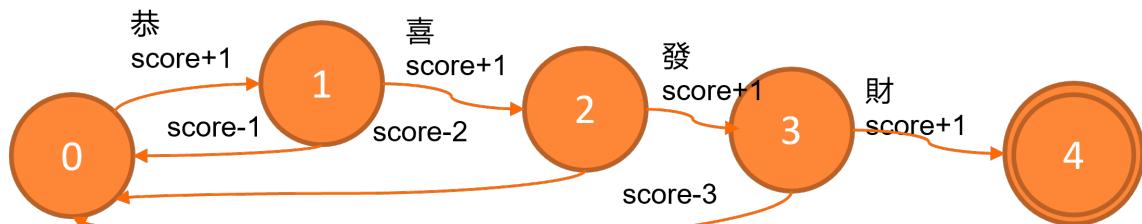


Figure 4.3: 中文淺融合 WFST 示意圖

在中文上，語音辨識的最小單元是字，正好對應到英文上的音素。所以我們使用 Wenet [59] 工具，將需要進行偏移的目標關鍵詞組建成 WFST [40]。如圖4.3所示，恭喜發財被編譯成 WFST，每按順序解碼出一個字時，都會獲得對應的獎勵分數，直到完全解碼出整個詞為止。

在解碼時，會同時搜索目標關鍵詞的 WFST，使得選定候選字時會同時考慮 CTC loss 和關鍵詞給予的加分。若無法正確解碼目標關鍵詞，則會扣除已添加的分數。給定的獎勵分數可以自定義，獎勵分數越高，偏移權重就越大，獎勵分數的設定可以根據不同資料集和實際需求設定。



## 4.4 備選詞預測 (Alternative Word Prediction, AWP)

為了在後修正階段進一步提升專有名詞的辨識正確率，我們探索了在中文上的備選詞預測方法。這個方法是通過訓練一個 Transformer，用於生成專有名詞的常見錯詞。通過使用訓練好的 Transformer 生成一個備選詞表，表上內容是專有名詞的  $n$  個常見錯詞， $n$  的大小可以自行設定。在解碼完成後，進行常見錯詞搜索，將常見錯詞替換回可能的專有名詞。

### 4.4.1 訓練資料的產生方法

Table 4.1: 備選詞預測模型訓練資料範例

|    |    |
|----|----|
| 阿思 | 阿斯 |
| 二成 | 二城 |
| 变形 | 辯行 |

備選詞預測模型的一筆訓練資料由一對詞語組成，其中一個詞是語料的 groundtruth，另一個詞是對應的解碼錯詞，如表4.1所示。

為了更準確的適配模型，我們需要用目標偏移模型來產生備選詞預測模型的訓練資料。通過準備大量語音語料，可以使用 AISHELL [12] 這樣的開源語料，如果可以選擇和目標偏移領域相關的語料則更好。我們將準備的語料輸入需要偏移的 ASR 模型，可以得到語音通過 ASR 模型解碼後的結果，如表4.2所示。

Table 4.2: ASR 模型輸出和 groundtruth 範例

|             |                |
|-------------|----------------|
| ASR Output  | 在那个面粉开始归国面包情况下 |
| Groundtruth | 在面粉开始贵过面包的情况下  |

為了獲取到訓練資料詞語對，我們需要找到解碼錯誤的詞語對。但是目前端到端 ASR 模型的輸出單位大多是字，所以我們需要先對輸入和輸出的語句進行分詞。我們選擇使用中文分詞工具 jieba [16]，對輸入和輸出的語句進行分詞。完成



分詞後，語句會如圖4.4所示。



Figure 4.4: 輸入輸出文字 jieba 斷詞後範例

我們還需要從完成分詞後的輸入和輸出文字中找到出現 substitution 的詞，才可以獲得訓練資料對。我們通過最短編輯距離的演算法將兩個句子中的詞語對齊，最短編輯距離的算法公式如下所示：

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} \end{cases} \quad (4.1)$$

完成對齊後可以得到如圖4.5所示的句子樣式



Figure 4.5: 完成對齊後句子範例

這樣，我們就可以找到發生了 substitution 的詞語對。因為 ASR 模型輸出的長度不確定性，會導致上個詞中的字落入下一個詞。針對因為 ASR 輸出錯誤導致的，對齊後詞長度不同的問題，我們刪除了詞長度不同的資料，只取用詞長度相同的資料作為訓練資料。



#### 4.4.2 備選詞預測模型架構

候選詞預測模型我們選擇了 Transformer，其中有兩層編碼器和兩層解碼器，隱藏層大小為 256，multi-head 設置為 8，feed-forward hidden size 設置為 2048。針對訓練資料的處理我們以中文字為最小單元，將每個詞視為由多個字組成的序列輸入。輸出內容的長度不一定與輸入序列相同。輸出詞的長度由 Transformer 決定，直到其產生結束符號為止。模型架構如圖4.6所示。

這個模型的具體實現我們參考了端到端機器翻譯的網路架構和資料處理方法，使用開源工具 OpenNMT [34] 實現。

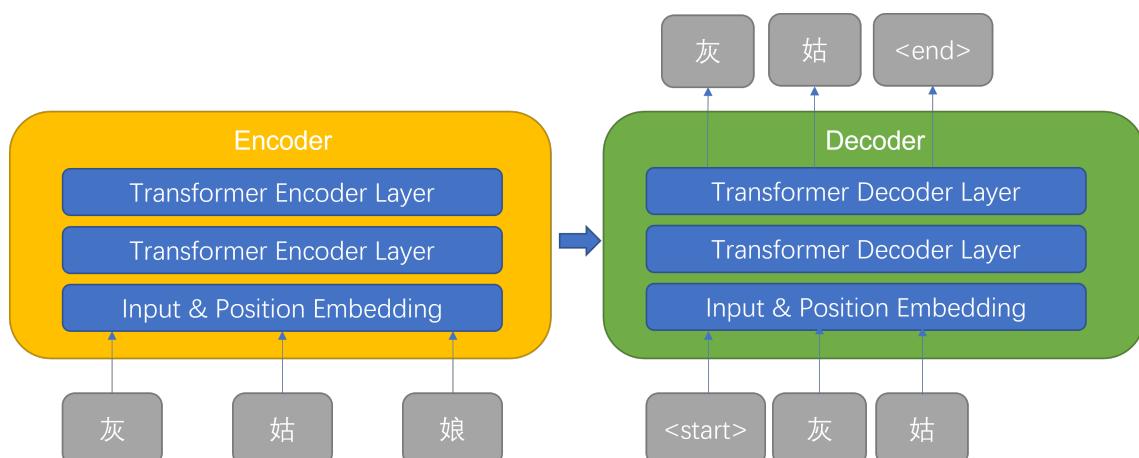


Figure 4.6: 備選詞預測模型架構圖

#### 4.4.3 備選詞表的生成和使用

通過收集的訓練資料完成模型訓練後，就可以使用模型生成專有名詞的備選詞。我們將需要進行偏移的目標關鍵詞輸入備選詞預測模型，模型可以設定輸出 Nbest 個備選詞，同時給出每個詞和輸入的相似度。通過設定 n 的大小，可以設定模型輸出備選詞的數量。通過設定相似度閾值，可以過濾相似度過低的備選詞，保證之後替換的合理性。通過這兩個參數，就可以控制備選詞表中每個關鍵詞的



備選詞的數量多少。

生成後的備選詞表如圖4.7所示。



Figure 4.7: 備選詞表範例圖

我們可以使用備選詞表，執行針對關鍵詞的後修正。我們會先對輸入的文本進行檢查，首先通過4.1節的方法，判斷其是否屬於可能存在專有名詞的語句。若可能存在專有名詞，才執行後修正的步驟。下一步對句子進行搜索，查看是否已經含有關鍵詞，若已經含有關鍵詞，就停止修正。若沒有，則開始搜索其中是否含有備選詞表中的備選詞，字數越多的備選詞優先級越高。若搜索到需要修正的句子中存在備選詞，則將備選詞替換回關鍵詞，並結束搜索。在實際的應用中，關鍵詞和備選詞的搜索可以用 WFST [40] 實現。

#### 4.4.4 常見詞表

為了防止生成的備選詞在日常對話中出現頻率過高，容易導致錯誤替換的發生，我們使用常見詞表來避免這種情況出現。我們對開源資料集的文字語料進行4-gram 計數。也就是說，對每個字的前後共四字組合的出現次數進行統計。當然，統計也包括三字、二字、一字。我這裡以中文字「恭喜發財」為例子，統計方法如圖4.8所示。

當一個文字組合出現次數超過  $n$  以上時，將這個文字組合納入常見詞表。 $n$



Figure 4.8: 常見詞計數方法示意圖

的大小可以自行設定，根據實際情況避免發生錯誤替換。

在常見詞表確定之後，可以直接於修正階段應用。後修正階段搜索到備選詞時，會額外檢查備選詞是否存在於常見詞表中。若存在則不進行替換。也可以將常見詞表於生成備選詞表階段進行應用，若生成的備選詞出現在常見詞表中，則不納入備選詞表。



## 第五章 實驗設計和結果討論

本章節針對語境偏移任務，採用四種方法進行實驗，並分為四個階段進行語境偏移實驗，以驗證其有效性。我們在5.1中介紹了實驗的流程和環境參數等基本信息。我們在每個階段設計了不同模型及參數的對比實驗，以證明我們方法的有效性。

### 5.1 實驗流程及設定

本小節介紹了實驗的流程和進行實驗的硬體環境。

#### 5.1.1 實驗流程

我們在5.2中設計語境檢測方法的相關實驗，其中包括三個實驗，分別對比了不同的下游模型、對比了不同的意圖辨識方法，並應用到語境偏移中觀察其對語境偏移整體效果的影響。

我們在5.3中進行訓練模型時偏移的方法實驗，通過對比是否使用相關領域語料微調模型的實驗結果，來證明微調模型方法的有效性。

我們在5.4中進行解碼和後處理階段的語境偏移實驗。其中包括四個子實驗，第一個實驗通過調整獎勵分數來觀察對中文淺融合方法的影響。第二個實驗對比

了不同的備選詞預測模型架構。第三個實驗嘗試通過常見詞列表來避免錯誤替換。第四個實驗將中文淺融合方法和備選詞預測方法應用於語境偏移任務中，並與僅使用淺融合方法的方式進行對比，以證明該方法的優勢。



### 5.1.2 實驗環境

本研究的實驗環境如表5.1所示

Table 5.1: 實驗環境

| 項目  | 參數                           |
|-----|------------------------------|
| CPU | Intel ® Core (TM) i7-12700KF |
| GPU | NVIDIA GeForce RTX 3080ti    |
| RAM | 64GB                         |
| OS  | Ubuntu 20.04 LTS             |

## 5.2 使用端到端意圖識模型進行自動語境檢測

在這組實驗中，我們使用 CATSLU 多領域混合語料語境偏移任務作為我們的語境偏移任務，目的是測試在面對不確定輸入語句的領域時，方法的有效性。

### 5.2.1 不同下游模型對意圖辨識的影響

在本實驗中，我們旨在比較不同下游模型的準確度優劣。為此，我們參考章節4.1的方法，利用上游模型對音頻數據進行特徵提取，然後使用不同的下游模型進行意圖辨識。我們使用訓練集語料對意圖辨識模型進行訓練，並在測試集上評估其效能。在下游模型的比較中，我們選擇了 Linear 模型和 Bi-LSTM 模型，設定 batch size 為 32，學習率設置為  $1e-4$ ，並使用 adam [33] 優化器，同時採用 warm up 策略，進行 10 個 epoch 的訓練。實驗結果如圖5.1所示。

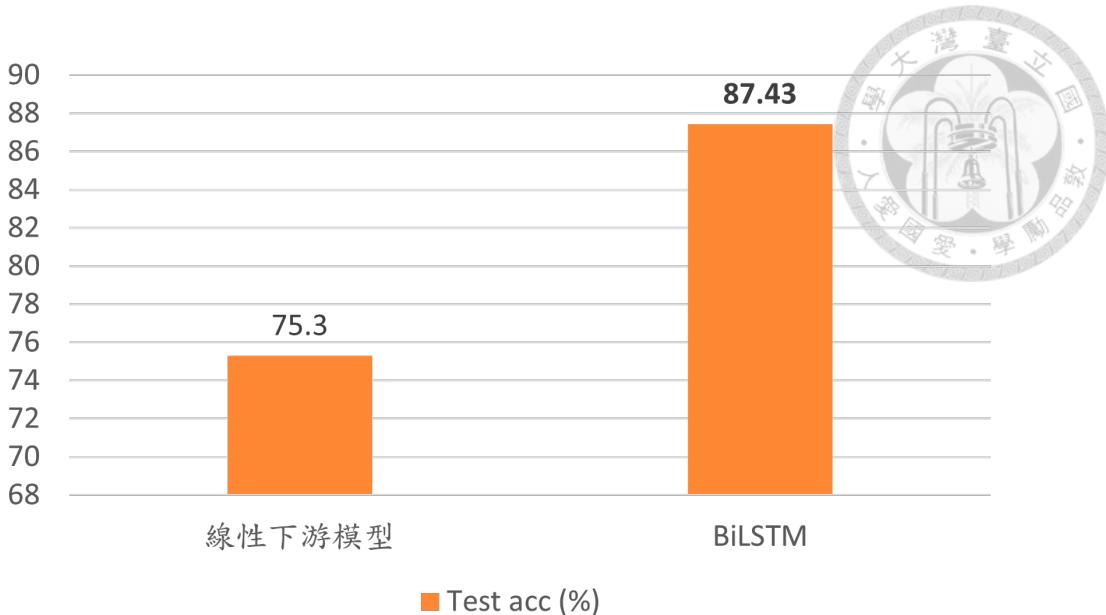


Figure 5.1: 使用不同下游模型的正確率對比實驗結果

從實驗結果可以看出，在這個任務中，我們改進的 Bi-LSTM 下游模型有更好的正確率表現。我們認為是在連續的特徵向量中，Bi-LSTM 模型更容易觀察到上下文的變化，對輸入語句意圖辨識的正確率可以達到 87.43%。也同時證明這個模型可以應用到實際的語境檢測中。

### 5.2.2 不同方法的算力和正確率對比

本實驗的目的是比較端到端意圖辨識方法和傳統意圖辨識方法在準確率和算力需求上的差異。為此，我們採用了兩種不同的方法來進行意圖辨識。一種是傳統的方法，即先使用語音辨識模型將語音轉換為文字，再使用語言模型對文字進行分析。另一種是端到端的方法，即直接使用一個模型對語音進行分析，不需要中間轉換。

在傳統的方法中，我們選擇了目前最先進的開源語音辨識模型 Whisper [46] 和廣泛應用的語言模型 BERT [10]。Whisper 模型我們選取目前辨識效果最好的 Large 版本，BERT 模型則採用適用於中文的 Chinese 版本。在訓練過程中，我們

固定 Whisper 模型不變，只訓練 BERT 模型的下游分類層，即一層 Linear 層。我們將 batch size 設為 16，learning rate 設為 5e-4，使用 AdamW [37] 優化器，並設置 10 個 epoch 作為訓練結束條件。



在端到端的方法中，我們選擇了前一個實驗中表現較好的 Bi-LSTM 模型作為下游模型，並保持與傳統方法相同的訓練參數。我們在測試集上比較了兩種方法在意圖辨識正確率上的表現，實驗結果如圖5.2所示。

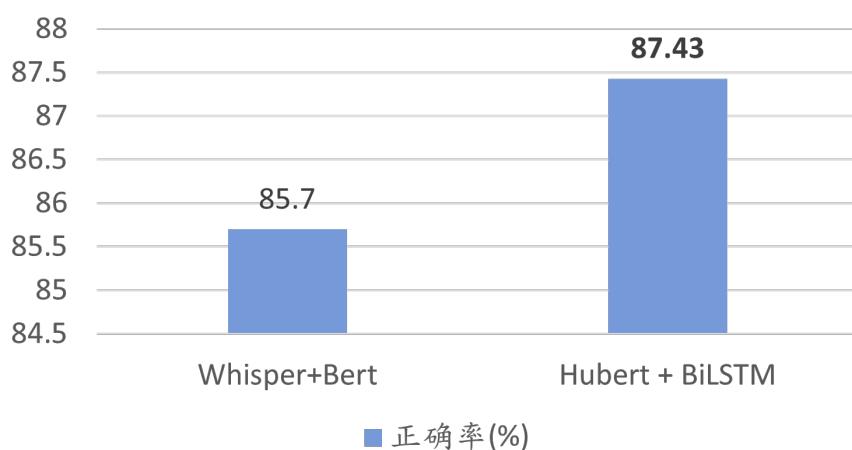


Figure 5.2: 不同方法的正確率對比實驗結果

此外，我們還對兩種方法在5.1.2中所述的實驗環境下進行了效率測試。我們分別計算了傳統意圖辨識方法和端到端意圖辨識方法對一小時語料進行推理的所需時間，並將結果繪製成圖表，如圖5.3所示。

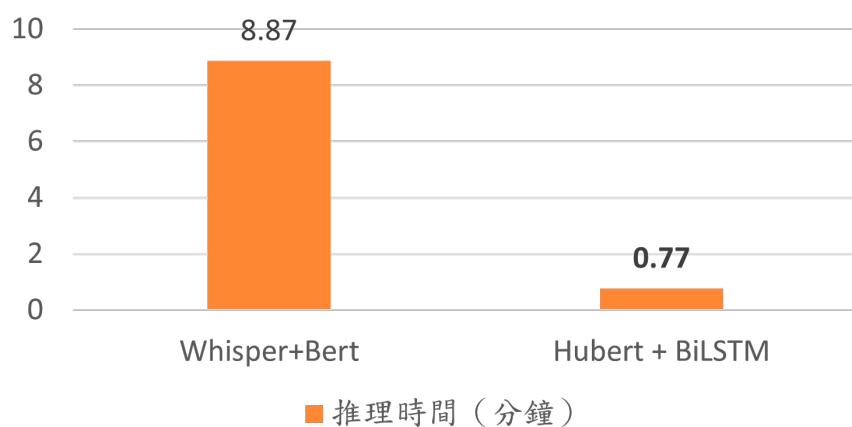


Figure 5.3: 不同方法的效率對比實驗結果

從實驗結果中，我們可以明顯地發現，相比於傳統的意圖辨識方法，端到端的意圖辨識方法在算力需求上更低，且在準確率上更高。



### 5.2.3 在語境偏移前應用語境檢測方法

本實驗的目的是測試端到端意圖辨識方法在面對未知領域語句的語境偏移上的表現。我們首先使用端到端意圖辨識方法對語料進行語境檢測，並根據檢測出的語境來進行相應的語境偏移。我們使用了中文淺融合的方法來實現語境偏移，並將偏移分數設置為 3.0。我們比較了三種不同的方法在準確率上的表現，分別是不進行任何偏移、對所有關鍵詞進行中文淺融合、根據語境檢測結果進行中文淺融合。同時我們設置了語境檢測結果為全對的組別作為語境檢測的對照組。實驗結果如圖所示。

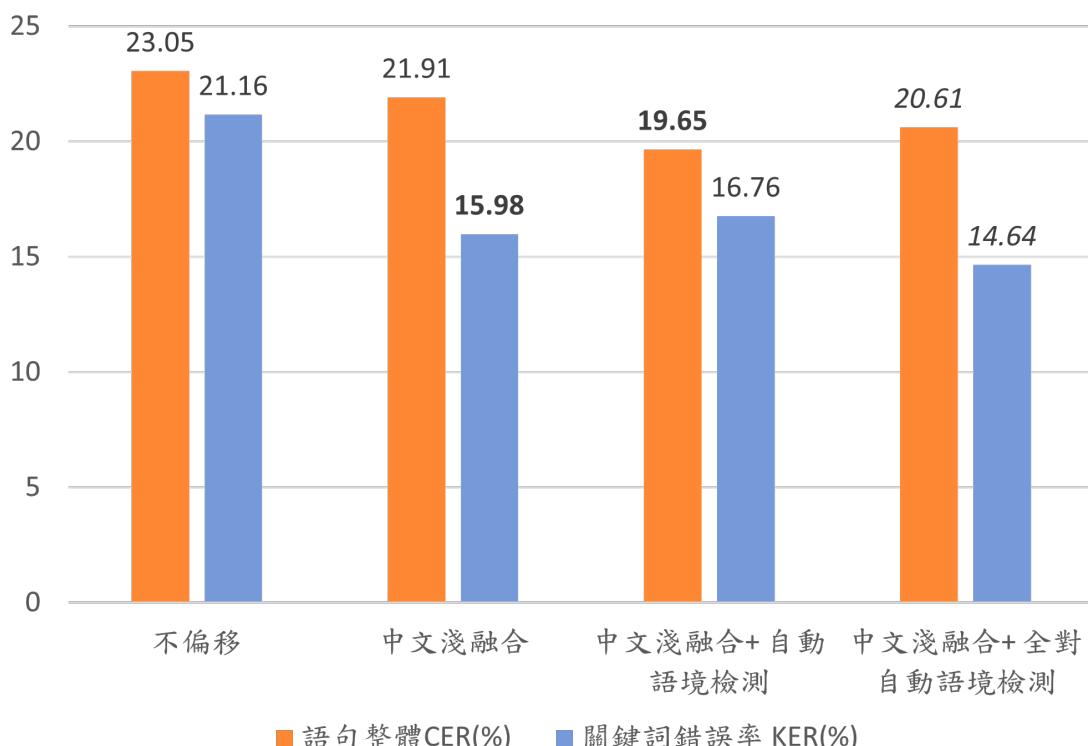


Figure 5.4: 語境檢測應用在語境偏移中實驗結果

圖表中的數據顯示，使用語境檢測的方法能夠有效地降低語句整體的 CER，



這表明語境檢測方法能夠修正一些錯誤的偏移方向。然而，我們也發現關鍵詞的錯誤率有所上升，我們推測這可能是由於語境檢測本身存在一定的不準確性造成的。此外，由於目前測試資料中各個分類關鍵詞數量較少，因此使用語境檢測的方法在關鍵詞準確率上不如不使用偏移的方法。但是，在實際應用中，如果有多个領域和大量關鍵詞，我們預期使用語境檢測會帶來更好的效果。

我們使用原始的任務標籤模擬語境檢測結果全對時的情況，我們發現其可以做到最優的語境偏移效果。但是與我們推測不同的是，全部正確的自動語境檢測的 CER，要高於有錯誤的自動語境檢測的 CER。關於這個問題我們推測可能是部分端到端意圖辨識分類錯誤的內容，其本身的非專有名詞的部分是具有混淆狀況的，而自動語境檢測很好的將其區分出來並分入可能性較大的類別，使其非專有名詞部分被正確辨識的機會更大。這間接證明自動語境檢測的優勢，可以幫助語意不清的語句找到最適合它的類別，並執行語境偏移。

### 5.3 少量語料微調模型實現語境偏移

本研究旨在驗證預訓練模型在少量語料微調後能否有效地適應不同的語境偏移場景。

為此，我們設計了一項使用少量語料微調模型的實驗，並以 CATSLU 混合語料語境偏移任務作為評估指標。該任務提供了一部分訓練集資料，可供我們利用來微調模型。我們選用了基於開源 WenetSpeech [58] 語料庫（約 10000 小時）訓練出來的 Conformer 模型作為預訓練模型，並以其參數作為微調的初始值。在微調過程中，我們採用了 80 維的 FBank 特徵作為輸入，並使用梯度累加 16、batch size 64、Adam 優化器和 2e-3 的學習率的超參數設置，共進行了 640 個 epoch。

圖5.5展示了微調後模型與原始模型在性能上的對比結果。證明如果使用領域

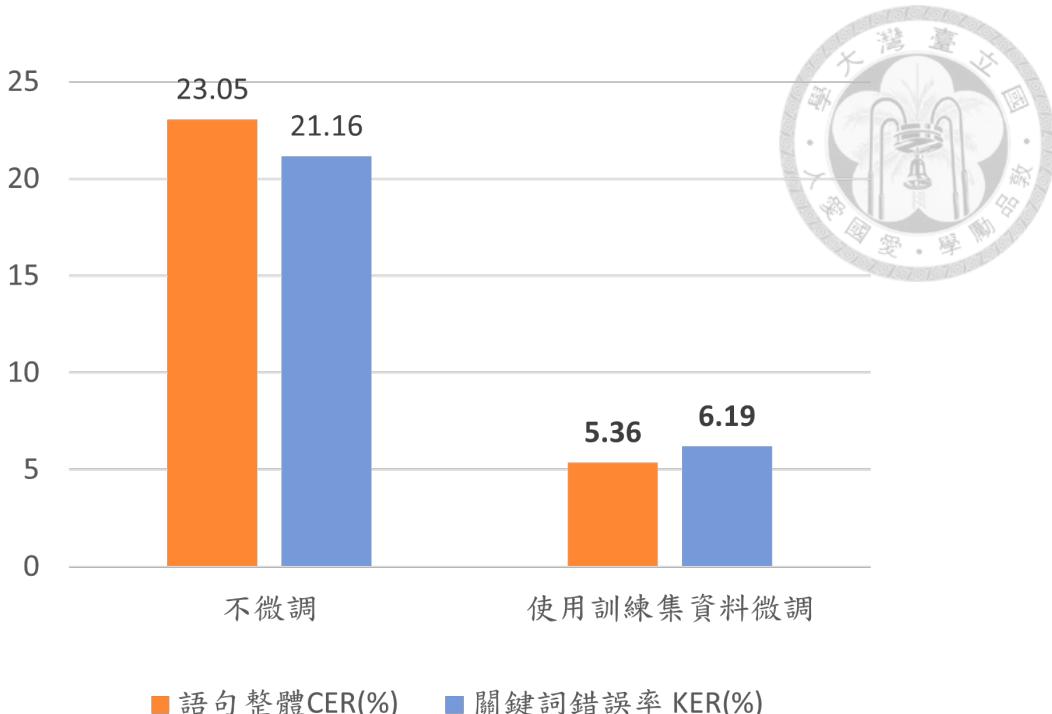


Figure 5.5: 微調模型實驗結果

相關的資料對原先的 ASR 模型進行微調，可以得到很優秀的效果。這個方法可以針對不同領域訓練出不同的 ASR 模型，並和語境檢測方法同時使用，可以達到較為優秀的效果。但使用這個方法的前提是具備相關領域的語音語料。

## 5.4 解碼和後處理階段語境偏移

在這組實驗中，我們選用 CATSLU 專有名詞語境偏移任務，測試解碼階段和後處理階段的語境偏移方法對於專有名詞的偏移效果。

### 5.4.1 偏移分數對中文淺融合的影響

本實驗旨在探討不同的偏移獎勵分數對於語境偏移效果的影響，並以 CER 和 KER 兩個指標作為評估標準，以期找出最適合的偏移分數設定。為此，我們選擇了 CATSLU 專有名詞語境偏移任務中的一個具有挑戰性的子任務，即影片名辨



識，作為實驗場景。

在實驗設計上，我們使用了一致的開源模型，即基於開源 WenetSpeech<sup>[58]</sup> 語料庫（約 10000 小時）訓練出來的 Conformer 模型。在解碼階段，我們採用了中文淺融合方法來增強模型對於影片名稱的識別能力，並嘗試了不同的淺融合獎勵分數從 0 到 9 之間的整數值，其中 0 表示不使用淺融合方法。我們將實驗結果以折線圖的形式呈現出來，如圖 5.6 所示。

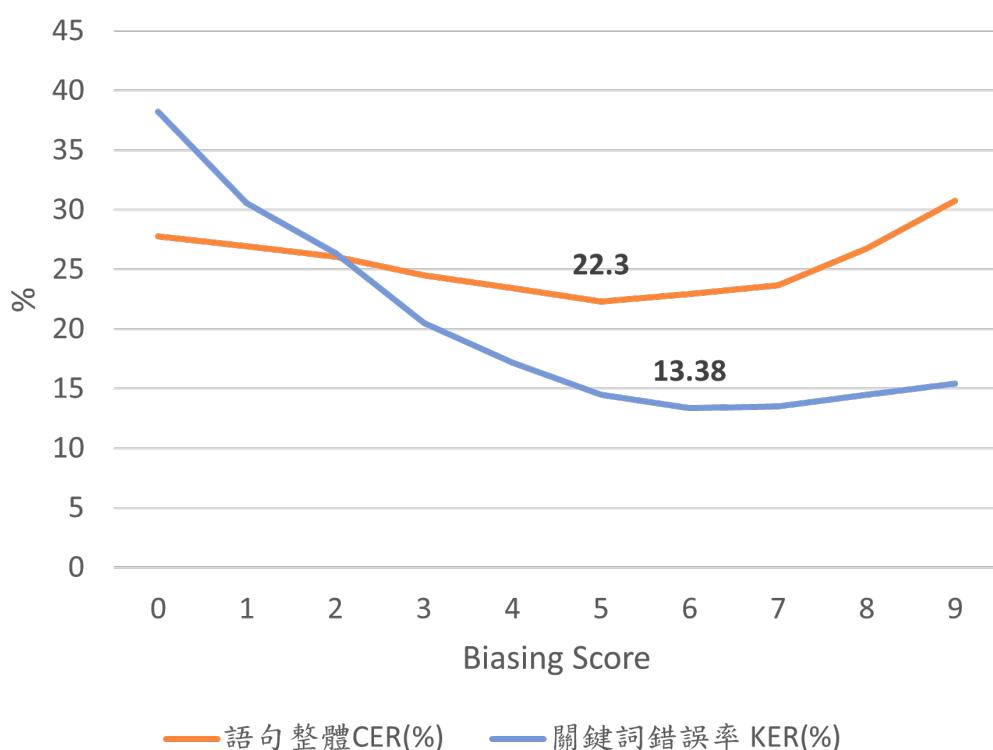


Figure 5.6: 偏移分數調整實驗結果

從實驗結果中，我們可以發現一些符合預期的規律。

首先，當獎勵分數較低時，適當地提高獎勵分數能夠有效地提升文章整體和關鍵詞辨識率，這表明淺融合方法對於增強模型的偏移能力是有幫助的。其次，當獎勵分數超過一定閾值時，則會導致整體語句的 CER 升高，這可能是因為過高的獎勵分數使得模型過度偏向於關鍵詞而忽略了其他部分。最後，當獎勵分數達到極高時，甚至會對 KER 產生負面影響，這可能是因為過高的獎勵分數使得模型

產生了錯誤的關鍵詞或者重複的關鍵詞。因此，在實際應用中，我們需要根據不同的使用場景和需求，通過實驗來確定最佳的獎勵分數數值。



### 5.4.2 AWP 模型選擇

本實驗旨在探討不同類型的 AWP 模型架構如何影響 AWP 模型在預測 ASR 錯誤上的表現。

本實驗採用正確率作為評估指標。正確率指當 groundtruth 作為輸入時，AWP 模型能夠產生與目標偏移 ASR 模型相同的錯誤輸出。一個表現良好的 AWP 模型能夠有效地預測 ASR 模型會如何將目標關鍵詞辨識成錯誤的詞彙。

本研究所採用的訓練資料集與測試資料集均來自於 AISHELL-2 語料，並且利用第4.4.1節所述的方法進行特徵提取。在測試資料集中，共有 53817 個字對，佔所有資料的約 10%；在訓練資料集中，共有 538169 個字對，佔所有資料的約 90%。

為了進行這項實驗，我們選擇了一個基於 Transformer 架構的模型，該模型由兩層自注意力編碼層和兩層自注意力解碼層組成。我們將其與一個預訓練的語言模型 mengzi-t5-base-mt 進行對比，該語言模型是基於 t5 模型 [38]，並使用大規模的中文語料進行了訓練。

在訓練 Transformer 模型時，我們將中文文本按字切分作為輸入單位，並採用 adam 優化器進行參數更新。我們設置 batch size 為 32，學習率為 2，並在前 8000 步進行學習率預熱，在後面的步驟中保持學習率不變，直到訓練達到 100000 步。在微調 T5 模型時，我們使用相同的輸入單位和 batch size，但改用 Adafactor [49] 優化器進行參數更新。我們設置學習率為 1e-3，並對整個訓練資料集進行 10 次迭代以確保模型收斂。

在完成訓練後，我們對兩種模型進行了實驗評估，並將結果展示在圖5.7中。

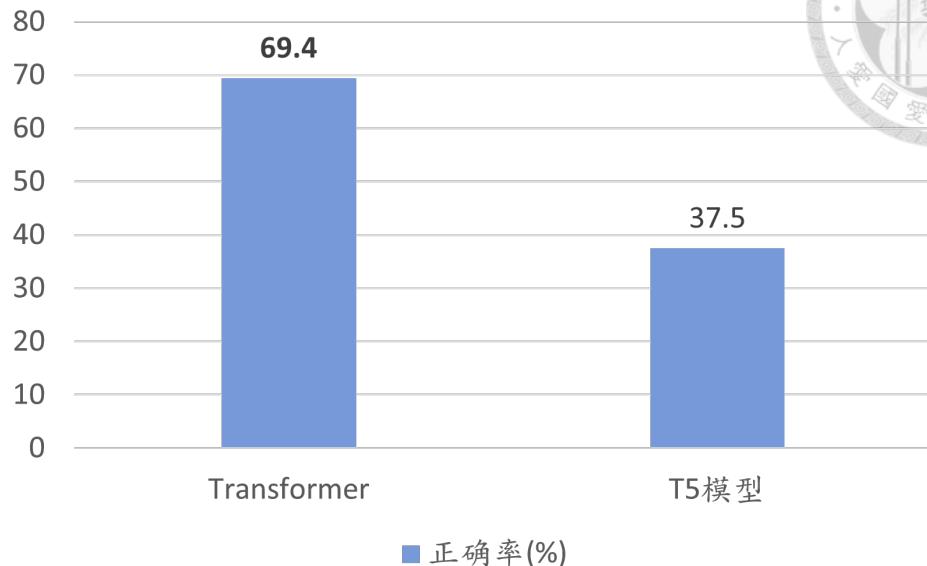


Figure 5.7: 不同的 AWP 模型架構實驗結果

Transformer 模型進行備選詞預測任務時，其在預測正確率上顯著優於使用預訓練的語言模型 T5。這表明 Transformer 模型能夠有效地生成合適的備選詞，而 t5 語言模型的預訓練對這個任務起不到作用。因此，在後續的實驗中，我們將採用 Transformer 架構作為 AWP 模型的基礎架構，並在此基礎上進行改進和優化。

### 5.4.3 常見詞列表

此實驗目的是檢驗使用常見詞列表方法對 AWP 方法的改進效果。我們選擇了 CATSLU 專有名詞語境偏移這個測試任務，並在其中選取了影片名語境偏移作為子任務進行實驗。進行解碼的 ASR 模型是使用 WenetSpeech（約 10000 小時）訓練的 Conformer 模型。

為了建立常見詞列表，我們對 AISHELL-2 這個大規模的中文語音資料集進行了 4-gram 統計，並將出現頻率超過 500 次的 4-gram 作為常見詞。在 AWP 方法中，我們在備選詞搜索階段忽略所有常見詞，以降低錯誤替換常見詞的可能性。

我們將使用常見詞列表方法和不使用常見詞列表方法的 AWP 方法進行比較，並與不使用 AWP 方法的基準方法進行對照。所有的方法都是解碼階段進行了中文淺融合技術後進行的後修正步驟。圖5.8中展示了各種方法在不同指標上的實驗結果。

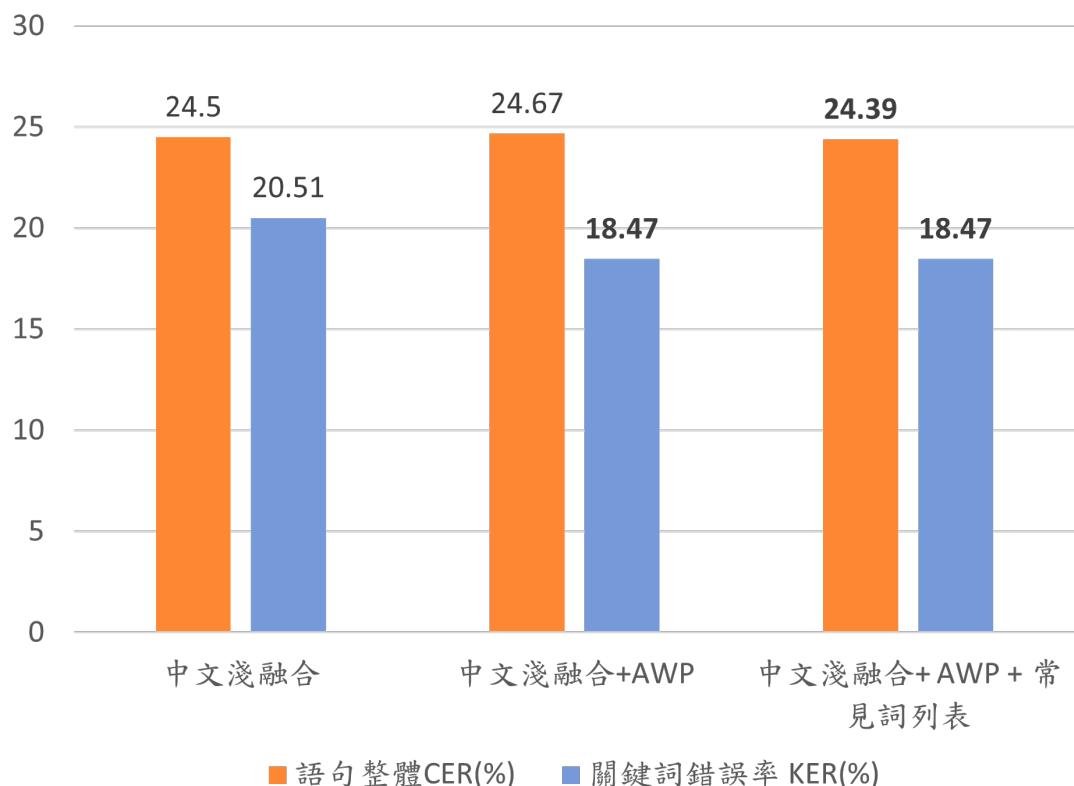


Figure 5.8: 常見詞列表對照實驗結果

根據圖表中展示的數據，我們可以發現，使用常見詞列表方法相比不使用常見詞列表方法的 AWP 方法，能夠顯著降低 CER 這一指標。這說明了使用常見詞列表方法能夠有效地避免錯誤替換備選詞的情況，從而提高了語音辨識的準確性。因此，在後續的實驗中，我們將採用常見詞列表方法作為 AWP 方法的一個重要組成部分，並在備選詞搜索階段對備選詞進行過濾。



#### 5.4.4 多領域任務中應用中文淺融合和 AWP 方法

為了驗證 AWP 方法在不同的測試任務上的通用性和有效性，我們在 CATSLU 專有名詞語境偏移這一測試任務的所有子任務上進行了實驗。

我們分別使用了不使用偏移、只使用中文淺融合方法和使用中文淺融合方法加上 AWP 方法這三種方案進行了解碼，並比較了它們在各項指標上的表現。我們使用的 ASR 模型是基於 WenetSpeech 這個大規模的中文語音資料集（約 10000 小時）訓練出來的 Conformer 模型，該模型具有良好的泛化能力和識別能力。圖 5.9 和圖 5.10 中展示了各種方案在不同子任務上的實驗結果。

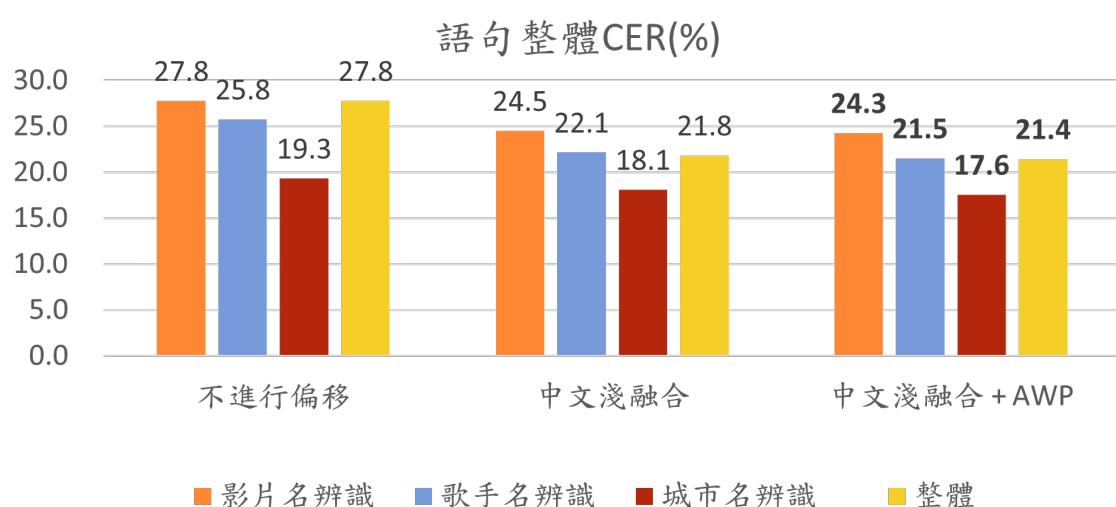


Figure 5.9: CATSLU 專有名詞語境偏移全部子任務實驗語句整體 CER

根據圖表中展示的數據，我們可以發現，在 CER 和 KER 這兩項指標上，使用中文淺融合加上 AWP 的方法在所有子任務上都顯著優於不使用偏移和只使用中文淺融合的方法。這說明了我們提出的中文備選詞預測方法能夠有效地糾正專有名詞的錯誤，在中文淺融合方法的基礎上進一步增強了專有名詞識別的能力。

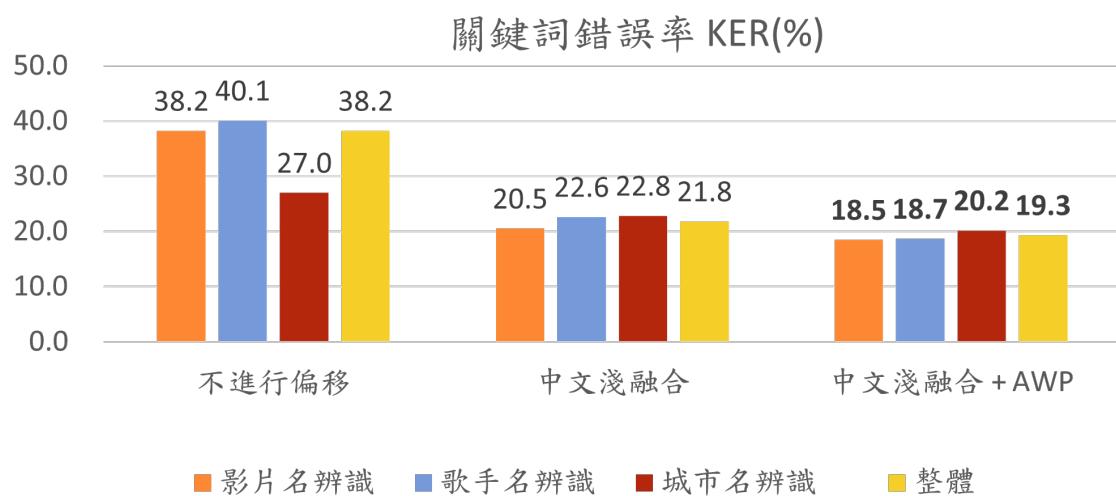


Figure 5.10: CATSLU 專有名詞語境偏移全部子任務實驗關鍵詞錯誤率 KER





## 第六章 結論和未來工作

本章節分為兩個小節分別介紹結論和未來工作。

### 6.1 結論

本文在開源資料集 CATSLU 的基礎上，定義了兩個中文語境偏移任務，分別針對專有名詞和多領域混合語料的語境偏移問題，並採用了四種不同階段的語境偏移方法，對端到端 ASR 模型進行語境偏移。在原有的淺融合方法之上，本文還提出了一種中文備選詞預測模型，用於生成備選詞表並進行後處理錯誤修正，從而進一步提高了語境偏移的效果。

本文通過學習先前的研究在英文開源資料集上定義語境偏移任務的方法，使用中文的 CATSLU 資料集定義了兩個中文語境偏移任務。分別是 CATSLU 專有名詞語境偏移任務和 CATSLU 多領域混合語料語境偏移任務。前者目的是評價針對專有名詞的語境偏移方法的效果優劣，後者則是貼近現實情況評價語境偏移方法在遇到來自不同領域的語句時如何通過語境偏移來提高端到端 ASR 模型在各個領域的表現。

本文接著探討了四種不同階段的語境偏移方法，分別涉及到 ASR 模型的辨識前、模型、解碼和後處理階段。在辨識前階段，本文使用了一種基於自監督預訓



練模型和雙向長短期記憶網絡端到端的語境辨識模型，用於自動檢測輸入語句所屬的領域類別。在模型階段，利用少量目標領域的數據對端到端模型進行微調訓練，以提升其在特定領域下的性能。在解碼階段，本文使用了一種基於關鍵詞獎勵機制的中文淺融合方法，用於針對特定領域下的關鍵詞進行優先解碼。在後處理階段，本文使用了一種基於 Transformer 的備選詞預測模型，用於生成備選詞表並進行後處理錯誤修正。

從實驗結果上看，使用語境檢測的方法 CER 的錯誤減少率 (error reduce rate, ERR) 為 10.31%。使用語料微調原始模型的 CER 錯誤減少率為 76.75%。使用中文淺融合的 CER 的錯誤減少率為 21.58%，KER 的錯誤減少率為 42.93%。在中文淺融合基礎上使用備選詞預測方法 CER 的錯誤減少率為 23.02%，KER 的錯誤減少率為 50.52%。可以看出，在四個方向上的偏移方法都有其有效性。

## 6.2 未來工作

本研究所面臨的困難在於，缺乏足夠規模的開源資料集以進行不同領域和場景下的語境偏移任務。由於開源資料集覆蓋範圍有限，無法對各種可能出現的語境偏移情況進行全面而深入地評估與分析。因此，未來工作將致力於利用大型開源資料集（如 WenetSpeech 等）建立具有代表性和挑戰性的語境偏移任務，並採用合理有效的篩選標準挑選出符合特定語境要求的語句，以期能夠對不同類型的語境偏移方法進行客觀而全面地評價與比較。

在備選詞預測方法方面，本研究仍有進一步改進和拓展的空間。例如，可以考慮將預測出的備選詞納入淺融合策略中偏移詞列表的生成過程中，以增加偏移詞列表的多樣性和豐富性。此外，也可以探討不同領域語料對備選詞模型訓練效果的影響，並比較不同領域模型在不同場景下的表現。



在備選詞搜索方法方面，本研究亦有改善和優化的可能。例如，可以備選詞表轉換成有限狀態轉換器（FST），並通過 FST 匹配句子中是否包含目標備選詞。由於實際場景下可能涉及大量關鍵詞和句子，因此需要尋找更高效率和準確度的搜索和替換算法。

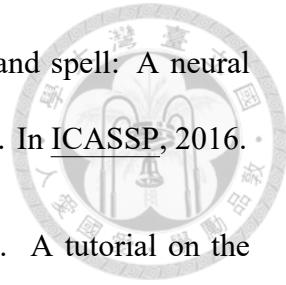
在於端到端意圖辨識配合的語境偏移方法上，可以嘗試搭配 whisper 的 Prompt 輸入進行，在輸入語音進行辨識的同時輸入其可能存在的語境上文，以幫助 whisper 模型更好的辨識。



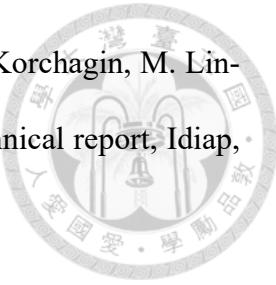


## 參考文獻

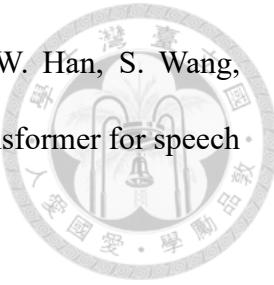
- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [3] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.
- [4] L. Borgholt, J. D. Havtorn, M. Abdou, J. Edin, L. Maaløe, A. Søgaard, and C. Igel. Do we still need automatic speech recognition for spoken language understanding? *arXiv preprint arXiv:2111.14842*, 2021.
- [5] R. N. Bracewell and R. N. Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.
- [6] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5, 2017.



- [7] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In ICASSP, 2016.
- [8] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. Annals of operations research, 134:19–67, 2005.
- [9] M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Zelasko, and M. Jetté. Earnings-21: a practical benchmark for asr in the wild. arXiv preprint arXiv:2104.11348, 2021.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [11] P. Dighe, A. Asaei, and H. Bourlard. On quantifying the quality of acoustic models in hybrid dnn-hmm asr. Speech Communication, 119:24–35, 2020.
- [12] J. Du, X. Na, X. Liu, and H. Bu. Aishell-2: Transforming mandarin asr research into industrial scale. arXiv preprint arXiv:1808.10583, 2018.
- [13] E. B. Dynkin and E. B. Dynkin. Markov processes. Springer, 1965.
- [14] S. R. Eddy. Hidden markov models. Current opinion in structural biology, 6(3):361–365, 1996.
- [15] J. D. Fox and N. Delworth. Improving contextual recognition of rare words with an alternate spelling prediction model. arXiv preprint arXiv:2209.01250, 2022.
- [16] fxsjy. jieba. <https://github.com/fxsjy/jieba>, 2020.
- [17] M. Gales, S. Young, et al. The application of hidden markov models in speech recognition. Foundations and Trends® in Signal Processing, 1(3):195–304, 2008.



- [18] P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiat, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang. Real-time asr from meetings. Technical report, Idiap, 2009.
- [19] N. R. Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of mathematical statistics*, 34(1):152–177, 1963.
- [20] A. Gourav, L. Liu, A. Gandhe, Y. Gu, G. Lan, X. Huang, S. Kalmane, G. Tiwari, D. Filimonov, A. Rastrow, et al. Personalization strategies for end-to-end speech recognition systems. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7348–7352. IEEE, 2021.
- [21] R. Graham, L. Aldridge, C. Carter, and T. C. Lansdown. The design of in-car speech recognition interface for usability and user acceptance. Engineering psychology and cognitive ergonomics, pages 313–320, 2020.
- [22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376, 2006.
- [23] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks, 18(5-6):602–610, 2005.
- [24] R. Gray. Vector quantization. IEEE Assp Magazine, 1(2):4–29, 1984.

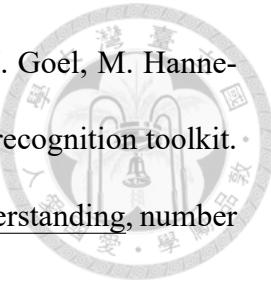


- [25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. [arXiv preprint arXiv:2005.08100](#), 2020.
- [26] T. Hori, S. Watanabe, Y. Zhang, and W. Chan. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. [arXiv preprint arXiv:1706.02737](#), 2017.
- [27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. [IEEE/ACM Transactions on Audio, Speech, and Language Processing](#), 29:3451–3460, 2021.
- [28] W. R. Huang, C. Peyser, T. N. Sainath, R. Pang, T. D. Strohman, and S. Kumar. Sentence-select: Large-scale language model data selection for rare-word speech recognition. 2022.
- [29] S. Husnjak, D. Perakovic, and I. Jovovic. Possibilities of using speech recognition systems of smart terminal devices in traffic environment. [Procedia Engineering](#), 69:778–787, 2014.
- [30] K. Irie, A. Zeyer, R. Schlüter, and H. Ney. Language modeling with deep transformers. [arXiv preprint arXiv:1905.04226](#), 2019.
- [31] N. Jung, G. Kim, and J. S. Chung. Spell my name: keyword boosted speech recognition. In [ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing \(ICASSP\)](#), pages 6642–6646. IEEE, 2022.
- [32] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar. An analysis of incorporating an external language model into a sequence-to-sequence

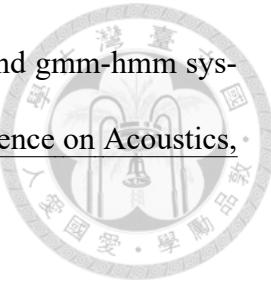
model. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5828. IEEE, 2018.



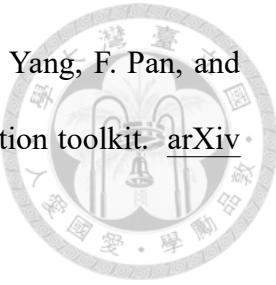
- [33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.
- [34] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810, 2017.
- [35] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In 2013 Humaine association conference on affective computing and intelligent interaction, pages 312–317. IEEE, 2013.
- [36] B. Logan et al. Mel frequency cepstral coefficients for music modeling. In Ismir, volume 270, page 11. Plymouth, MA, 2000.
- [37] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.
- [38] A. Mastropaoletti, S. Scalabrino, N. Cooper, D. N. Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota. Studying the usage of text-to-text transfer transformer to support code-related tasks, 2021.
- [39] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In Interspeech, volume 2, pages 1045–1048. Makuhari, 2010.
- [40] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. Computer Speech & Language, 16(1):69–88, 2002.
- [41] T. K. Moon. The expectation-maximization algorithm. IEEE Signal processing magazine, 13(6):47–60, 1996.



- [42] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding, number CONF. IEEE Signal Processing Society, 2011.
- [43] D. Povey, X. Zhang, and S. Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. arXiv preprint arXiv:1410.7455, 2014.
- [44] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao. Deep context: end-to-end contextual speech recognition. In 2018 IEEE spoken language technology workshop (SLT), pages 418–425. IEEE, 2018.
- [45] L. R. Rabiner, R. W. Schafer, et al. Introduction to digital speech processing. Foundations and Trends® in Signal Processing, 1(1–2):1–194, 2007.
- [46] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [47] D. A. Reynolds et al. Gaussian mixture models. Encyclopedia of biometrics, 741(659-663), 2009.
- [48] F. S. Richards. A method of maximum-likelihood estimation. Journal of the Royal Statistical Society: Series B (Methodological), 23(2):469–475, 1961.
- [49] N. Shazeer and M. Stern. Adafactor: Adaptive learning rates with sublinear memory cost, 2018.
- [50] D. Svozil, V. Kvasnicka, and J. Pospichal. Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems, 39(1):43–62, 1997.



- [51] P. Swietojanski, A. Ghoshal, and S. Renals. Revisiting hybrid and gmm-hmm system combination techniques. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6744–6748. IEEE, 2013.
- [52] TencentGameMate. chinese speech pretrain. [https://github.com/TencentGameMate/chinese\\_speech\\_pretrain](https://github.com/TencentGameMate/chinese_speech_pretrain), 2022.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- [54] K. Vesely, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In Interspeech, volume 2013, pages 2345–2349, 2013.
- [55] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, et al. Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015, 2018.
- [56] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, et al. Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051, 2021.
- [57] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. arXiv preprint arXiv:2102.01547, 2021.
- [58] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6182–6186. IEEE, 2022.



- [59] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu. Wenet 2.0: More productive end-to-end speech recognition toolkit. [arXiv preprint arXiv:2203.15455](#), 2022.
- [60] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar. 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, 2020.
- [61] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang. Shallow-fusion end-to-end contextual biasing. In [Interspeech](#), pages 1418–1422, 2019.
- [62] S. Zhu, Z. Zhao, T. Zhao, C. Zong, and K. Yu. Catslu: The 1st chinese audio-textual spoken language understanding challenge. In [2019 International Conference on Multimodal Interaction](#), pages 521–525, 2019.
- [63] 蔡淑芬. [基於語言模型的集成和重新排序以改善 ASR](#). 碩士論文, 國立臺灣大學資訊工程學研究所, 2022.