國立清華大學 電機工程學系
實作專題研究成果摘要

Development of a Traditional Chinese Automatic Speech Recognition System

台灣中文語音辨識系統開發

專題領域：資工領域

組　　別： B329
指導教授：李祈均教授
組員姓名：林冠宏
研究期間：2023 年 2 月至 2023 年 11 月止，共 10 個月

摘要

中文語音辨識技術在台灣的應用範圍日益擴大，包括智能語音助手、自動語音識別系統以及許多其他領域。這項技術的開發和應用需要克服語音多樣性和語音中的干擾，這是一個具有挑戰性的課題。

在開始語音辨識系統的模型訓練前，我們需要大量的語音錄音檔（.wav）以及其對應的文字檔（.txt），彙整成一個語料庫，並依照規定流程（去除標點符號、不屬於繁體中文的Unicode字符，以及簡體中文）對文字檔進行處理後，建立出屬於此語料庫的字典（lexicon，中文詞語對應注音符號）。資料庫建立完成且符合指定格式後，即可以開始使用Kaldi語音辨識模型針對此語料庫進行訓練。

本專題使用的資料庫來自NER Manual Transcription 語料庫第七集，為臺北科技大學和國立教育廣播電台合作產製之語料庫，語料庫節目內容大部分是談話性節目，多為自發性語音，只有少部分是新聞報導的朗讀式語音。語料庫包含以下節目：創青宅急便、教育開講、從心歸零，共計約116小時的語音檔。

1. Introduction

語音辨識在生活中廣泛地被運用，手機裡的語音辨識功能即為一個很常見的應用，能即時地將接受到的語音訊號，轉換成其對應的文字。這樣的功能可以解決許多問題，在不適合使用打字功能的情況下，仍可以透過語音辨識功能，來輸出想要的文字內容。如此實用的系統讓我對其背後運作的原理產生了興趣，因此希望在透過這個專題研究，實際地從語料庫建立開始，學習建立一個語音辨識系統的流程。

本專題中使用 Kaldi 語音辨識模型作為主要骨架，Kaldi 是一個開源的語音辨識工具包，被廣泛用於語音辨識和語音處理的研究和開發中。這個工具包提供了一整套用於建立語音辨識系統的工具和函式庫，涵蓋了從特徵提取、聲學模型訓練到解碼等整個語音辨識流程。

Kaldi 的優勢之一是其強大的工具組和靈活性，使得研究者和開發者能夠根據不同的需求進行定制和擴展。它被廣泛應用於學術界和工業界，在此語音辨識專題研究中和相關領域的研究和應用提供了很大的幫助。

2. Research Methodology

2-1. Workflow

1. 利用 Python 清理 NER 語料庫中的文字檔，去除標點符號、不屬於繁體中文的 Unicode 字符，以及簡體中文。
2. 第一步處理完的文字檔內容會變成一整句無分段的中文詞句，利用 CkipTagger（Github 開源工具）中的斷詞功能，將該中文詞句分割成有獨立意義的單詞。
3. 將第二步分割完的句子中的每個單詞，使用 Python Pinyin（Github 開源工具）生成其對應的注音符號。
4. 合併所有單詞及其對應的注音符號，刪除重複出現的單詞，即建立出屬於此資料庫的辭典（lexicon）。
5. 將建立好的 lexicon 加入語料庫，將其設定為 Kaldi 語音辨識模型的訓練資料庫。

3. Experimental Results

Fig. 1. 語料庫中原始文字檔（單份txt檔內容）

Fig. 2. 清理後的文字檔（僅留下繁體中文）

Fig. 3. 使用CkipTagger切詞後的文字檔

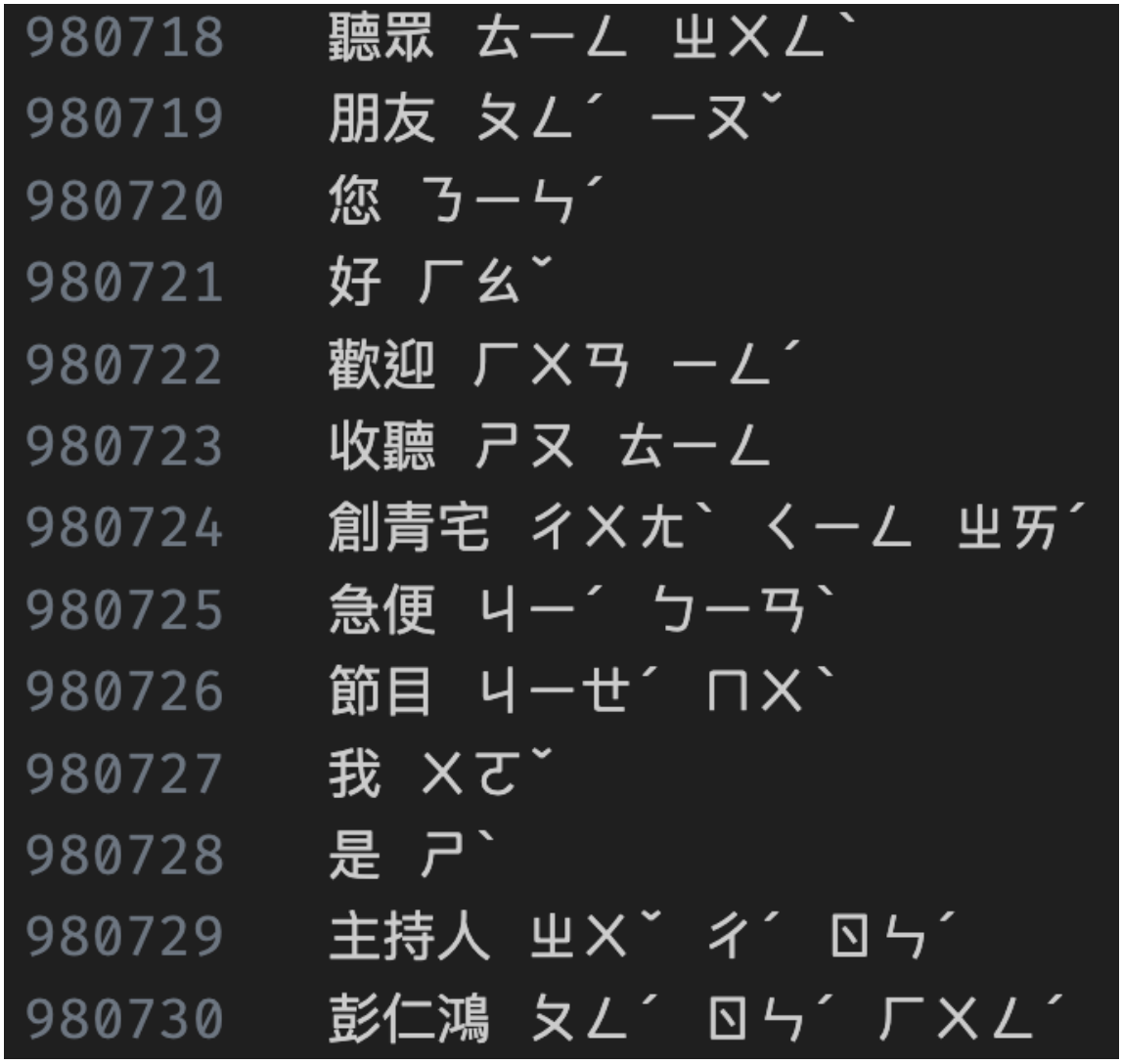

Fig. 4. 使用Python Pinyin產生對應注音符號的lexicon（僅擷取部分畫面）

4. Conclusion

從語料庫最初的資料整理，到建立此語料庫專屬的lexicon的過程相當繁雜，需要學習去使用許多Github上的開源工具（CkipTagger、Python Pinyin），但也是因為有這些開源的工具可以直接使用，可以協助我們更有效率地處理資料，以及執行預期所需的功能，不用全部靠自己從零開始編寫程式，就能依照各種不同規格的語料庫，順利生成專屬於該語料庫的lexicon。

5. Review and reflections

這次的專題實驗是透過實驗室server上的docker，進行語料庫的資料處理及模型的訓練，因為是初次接觸docker，所以在熟悉環境的建置下就花費了大量的時間與精力。常常遇到因為缺少引入的套件，產生許多程式的錯誤，起初不太理解程式無法運行的原因，以及無法理解其提供的錯誤訊息。但是在上網查詢相關資料以及其他人使用相同模型時所遇到的問題後，大部分的問題都能迎刃而解。大大地提升我在解決程式環境建構上處理問題的能力。

在實際深入探討語音辨識背後的整體流程後，發現了許多更為複雜的組成內容，如：特徵提取的方式（Kaldi中使用梅爾頻率倒譜系數（MFCC））、聲學模型（ Kaldi 使用類神經網絡（DNN）和隱馬爾可夫模型（HMM）等技術構建聲學模型）。雖然在此專題研究中，並沒有去改良Kaldi內部結構，但是已對其各個組成部分有初步的瞭解。促進我在未來學習到更多相關語音模型相關知識後，對於整個語音模型的架構，以及語料庫處理的方式做更完善的優化。

非常感謝李祈均教授以及實驗室的蘇柏豪、吳亞澤及林蔭澤學長，大力的幫忙及即時的解惑，讓我有機會在語音辨識的領域裡進行研究。學習如何自己整理語料庫，建立一個自己的語音辨識模型。