Computational Linguistics and Chinese Language Processing
Vol. 28, No.1, June 2023, pp. 1-18

基於語音自監督模型的複雜環境長語句意圖偵測與辨識

Intent Detection and Recognition of Long Sentences in Complex Environment Based on Speech Self-supervised Model

張開 *、葉子雋 †、王崇喆 #、張秋霞#、藍偉任†、張智星#

Kai Zhang, Tzu-Chun Yeh, Chung-Che Wang, Qiuxia Zhang, WeiRen Lan, and Jyh-Shing Roger Jang

摘要

本論文針對消防指揮記錄中心無綫電語料中，語句長、噪聲多的特點，提出了一種在長語句中辨識意圖的方法。此方法首先使用自監督學習 (self supervised learning) 模型來進行語音的特徵提取，然後再使用兩個下游模型分別對語音中的意圖進行偵測和辨識。此方法在無線電語料長語音意圖辨識任務上與 Whis- per+BERT 的方法相比，錯誤減少率 (error reduction rate, ERR) 為 33.2%。在關鍵詞發現(keyword spotting) 任務上與區域提案網絡(region proposal network, RPN) 方法相比，在每小時誤報次數(false alarm per hour, FAH) 相近的情況下，錯誤拒絕率 (false rejection rate, FRR) 的 ERR 為 73.2%。在短語句分類任務上和 Whisper+BERT 方法相比，ERR 為 4.3%。同時與

* 國立台灣大學資訊網路與多媒體研究所
Graduate Institute of Networking and Multimedia, National Taiwan University
Email: r10944061@ntu.edu.tw
† 迪威智能股份有限公司
DeepWave
Email: {kenshincs, welly}@dwave.cc
國立台灣大學資訊工程學系
Department of Information Engineering, National Taiwan University
Email: geniusturtle6174@gmail.com; r10922164@ntu.edu.tw; jang@csie.ntu.edu.tw

Whisper+BERT 方法相比，推理算力需求下降了 91.4%。我們所提出的方法，可以廣泛地應用在從長語音（電話或無線電對話等）中提取關鍵資訊。

關鍵詞：意圖分類，長語句，自監督學習，關鍵詞發現，語句分類

Abstract

According to the characteristics of long sentences and lots of noise in the radio corpus of the fire command record center, this paper proposes a method to identify intent in long sentences. This method first using a self-supervised learning model for speech feature extraction, and then using two downstream models to detect and recognize intent in speech respectively. Compared with the Whisper+BERT method, this method has an error reduction rate (ERR) of 33.2% in the long speech intent recognition task of radio corpus. Compared with the region proposal network (RPN) method on the keyword spotting task, the false alarm per hour (FAH) is similar, and the false rejection rate (false rejection rate, FRR) ERR is 73.2%. Compared with the Whisper+BERT method on the short sentence classification task, the ERR is 4.3%. At the same time, compared with the Whisper+BERT method, the inference computing power requirement has dropped by 91.4%. This method can be widely used in the fields of extracting key information from long speech, recording key information of telephone or radio communication and so on.

Keywords: Intent Classification, Long Speech Sentence, Self-supervised Learning, Keyword Spot-ting, Speech Sentence Classification

1. 緒論 (Introduction)

在消防員和指揮中心的溝通中，如何從消防員回報的語音中提取出求助者的訴求，是指揮中心碰到的實際問題。在無線電通訊的特殊環境下，有說話人語速較快且現場環境嘈雜的情況，通用的自動語音識別 (automatic speech recog- nition, ASR) 模型的錯誤率很高。在其基礎上再使用口語理解 (spoken language understanding) 的方法對語音轉譯的文字進行資訊提取，效果並不理想。使用通用的 ASR 模型，往往會因為缺少專業領域的詞彙而不能正確轉譯相關語音 (Jung, Kim, & Chung, 2022)。在先前的研究中有指出，先轉換為文字再進行口語理解會丟失說話人語音中所含的部分資訊，會對口語理解的效果造成影響(Zhu, Zhao,Zhao, Zong, & Yu, 2019)。

有一項研究，其在 Fluent Speech Commands (Lugosch, Ravanelli, Ignoto, Tomar, & Bengio, 2019)上進行意圖辨識實驗結果表明，在有限的資料條件下，使用端到端的意圖辨識方法，與從人工轉譯或 ASR 轉譯的文本進行意圖辨識相比，有更高的正確率(Borgholt *et al*., 2021)。我們在中文的語音任務上對其進行了驗證。我們受到 SUPERB(Yang *et al*., 2021)利用語音的自監督學習模型(self supervised model)進行多種語

音任務的啟發。我們選擇了一個中文的 HuBERT (Hsu *et al*., 2021) 模型，其在 10000 小時的中文語料 (Zhang *et al*., 2022)上進行了預訓練，用來對音檔資訊進行特徵提取。我們設計了兩個下游模型和一個組合策略，利用上游模型抽取的音檔資訊，對長語音中的意圖進行偵測和分類。因為目前開源的中文資料集缺少在長語音中進行意圖偵測和辨識的任務，所以我們選擇在兩個基於開源資料集的語音任務，以及一個內部資料集上與現有方法進行對比以證明其有效性。

在語音模型的應用中，如何減少推理時對於算力的需求是一個很重要的問題，我們在新北市消防局無線電語料長語音意圖辨識任務中，對其算力需求與使用 Whisper+BERT 方法的算力需求進行了評估。

此文章的貢獻如下：

1. 我們提出了一種方法，基於語音的自監督學習模型，在長語句中偵測意圖出現並進行分類。此方法由 HuBERT 上游特徵提取模型和兩個下游模型構成。我們通過三個資料集上的實驗證明其有效性。
2. 我們的方法進行推理時相比於 Whisper+BERT 的方法算力需求更小，更適合在嵌入式設備或終端設備中運作。
3. 我們提出了一種對特定位置包含意圖的長語句進行註記和資料整理的方法，可以分離出訓練兩個模型所需要的資料，並有效的訓練模型。
4. 我們提出了一種雙模型的應用策略，和原始的方法流程進行對比，可以提高整體正確率。
5. 我們介紹了一種中文短語音的分類任務，它是基於開源資料集 CATSLU (Zhu *et al*., 2019) 進行的。

2. 任務介紹 (Tasks)

2.1 新北市消防局無線電語料長語音意圖辨識任務 (New Taipei City Fire Department's Radio Corpus Long Speech Intent Recognition Task)

新北市消防局無線電語料長語音意圖辨識任務的語料來源是新北市消防局執行任務的消防員和指揮中心溝通所用的無線電裝置，主要內容為醫療救護任務。其中包含消防員回報指揮中心傷者的身體狀況、車輛資訊、事故地點、原因、患者主訴等內容。任務是在長語句中找到患者主訴的內容，並進行意圖的分類。在無線電語料中，包含大量的環境雜訊、無線電傳輸雜訊，且說話人語速較快，發音不標準。

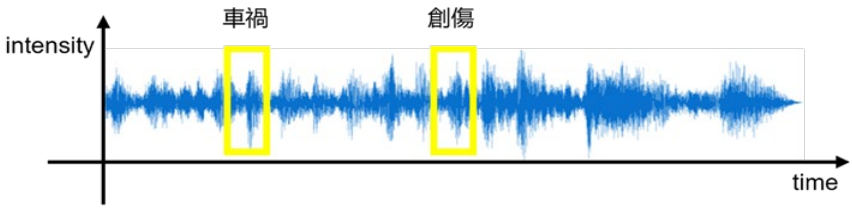

圖 1. 消防局無線電語料標記方法
[Figure 1. Fire Department Radio Corpus Marking Method]

2.1.1 評估標準 (Evaluation Criteria)

單意圖正確率是衡量模型對單意圖語句的預測能力的指標，其定義為模型輸出信心最高的一個意圖，為語句中的意圖的比例。若語句中含有多個意圖，則只要模型輸出信心最高的意圖，在語句中的任何一個意圖當中，即視為該語句被正確判斷。

三意圖正確率是衡量模型預測多個意圖時的效能的指標，此時模型會輸出信心最高的前三個意圖，而三意圖正確率的定義則為測試集中，每個語句包含的每個意圖被模型正確辨識的比例。由於資料集中，語句中出現的意圖數量最多為三個，因此採用三意圖正確率能更好的反映模型對多意圖語句的預測能力。

圖 2 為計算方法示意圖。三意圖正確率是每個語句中所有存在的意圖被正確預測的比例，在圖中可以表示為 (1 + 1 + 2)/7 = 57.1%。單意圖正確率為模型輸出信心度最高的意圖在語句的任意一個意圖中的比例，在圖中可以表示為(0 + 1 + 0)/3 = 33.3%。

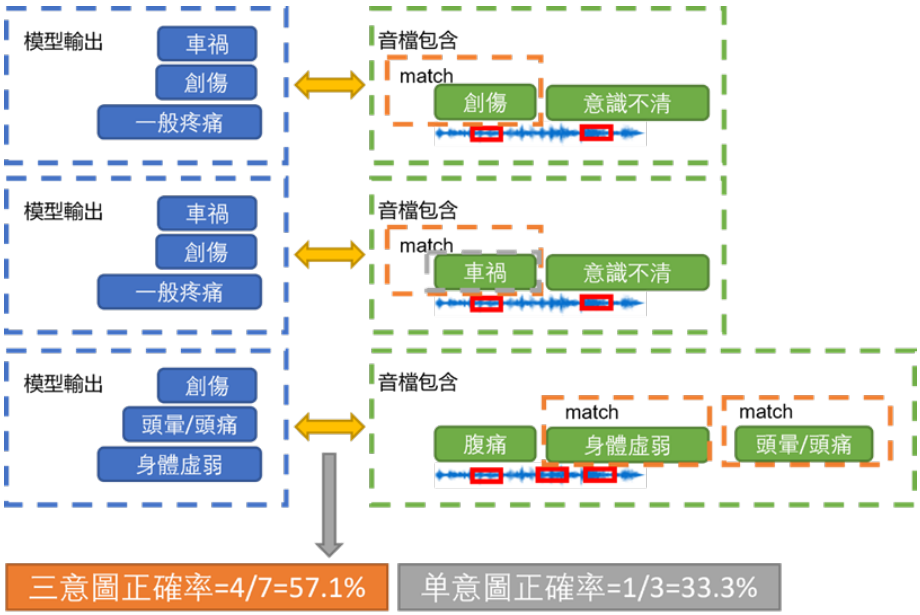

圖 2. 三意圖正確率計算方法示意圖
[Figure 2. Schematic Diagram of the Calculation Method of the Accuracy Rate of Three Intentions]

2.2 Mobvoi Hotwords中文關鍵詞發現任務 (Mobvoi Hotwords Chinese Keyword Spotting Task)

Mobvoi Hotwords (Hou, Shi, Ostendorf, Hwang, & Xie, 2019)是由出門問問公司釋出的語料，語料分為包含關鍵詞的語句和不包含關鍵詞的語句。其中包含的關鍵詞為「嗨小問」和「你好問問」。兩個關鍵詞各有約 36,000 段語句，來自 788 個說話人，年齡在 3 到 65 歲之間，錄音裝置為智慧音箱。這些智慧音箱會被放置在說話人不同距離的位置進行聲音採集，說話人會以不同的聲音強度進行說話。在此任務中，我們按照原文章的實驗要求，將關鍵詞「你好問問」視為目標關鍵詞，將「嗨小問」和無關鍵詞的語句視為非關鍵詞，通過比較模型的 FRR 和 FAH 來判斷模型的優劣。

2.3 CATSLU 短語句分類任務 (CATSLU Short Sentence Classification Task)

CATSLU 是由 21 屆 ACM International Conference on Multimodal Interaction 會議的第一屆 Chinese Audio-Textual Spoken Language Understanding Challenge 釋出的資料集，其中包含四個類別的詢問語句分別是天氣、音樂、影片、地圖。是通過智慧音箱採集的用戶命令語音，大多數為短音檔，但是也有部分語句為雜訊，並不存在意圖。

我們按照資料集的四種類型詢問語句進行分類，其中每個類型的語句中都包含有此類型的子意圖，並按照每句語句中包含的子意圖數量去確定此語句的有效性，以防止此語句存在不明確的意圖。我們將有效意圖為兩個以上的語句定為此類別的有效語句，意圖數量為零的語句為雜訊。此外我們將原先訓練集、驗證集、測試集資料重新進行劃分，比例為 70%、10%、20%。

通過這樣的方法劃分的語句則被分為五類，其中一類為雜訊，四類為各個類別意圖的語句。此資料集可以用來檢測中文短語句分類任務模型的性能。我們通過計算模型對于短語句分類的正確率來比較模型的優劣。

3. 方法設計 (Method Design)

3.1 方法流程 (Method flow)

我們在執行長語句的意圖辨識任務時，會遇到的問題是長語句中有很多與意圖無關的訊息，首先我們要對意圖出現的位置進行偵測。我們會先對長語句進行切割，為了防止意圖存在的位置被切割成兩段，影響後續模型的判斷，我們將切割語音的長度和步進長度分開設定，使得切割後的每段語音都會和前段語音有重疊的部分。如圖 3。

我們經過多次對比實驗確定在單段語音長度為 5 秒時，經過 HuBERT 進行聲音特徵抽取後，更有利於下游模型學習到音檔的特徵。因為意圖出現的平均時長都小於 1 秒，所以我們將其按照 5 秒為一段，步進長度 2.5 秒切割成多段 5 秒的語音。然後將切分好的每一個語音輸入 HuBERT 進行特徵的抽取得到每一段音檔的特徵向量。

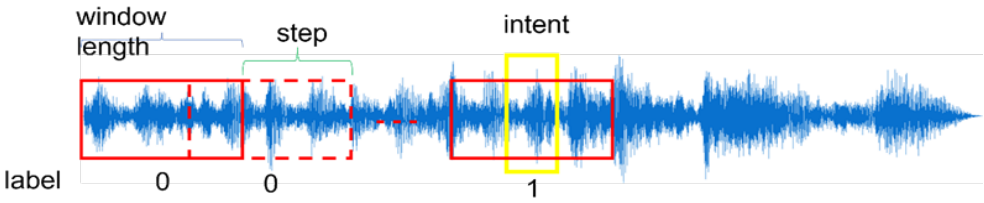

圖 3. 長語音切割方法
[Figure 3. Long Speech Segmentation Method]

我們可以根據特徵利用意圖偵測下游模型在這些語音中找到最有可能存在意圖的語音區間，確定了具有意圖的語音區間後，再將具有意圖的語音區間對應的 HuBERT 特徵向量輸入意圖分類下游模型，最後輸出三個可能存在的意圖。如圖 4。

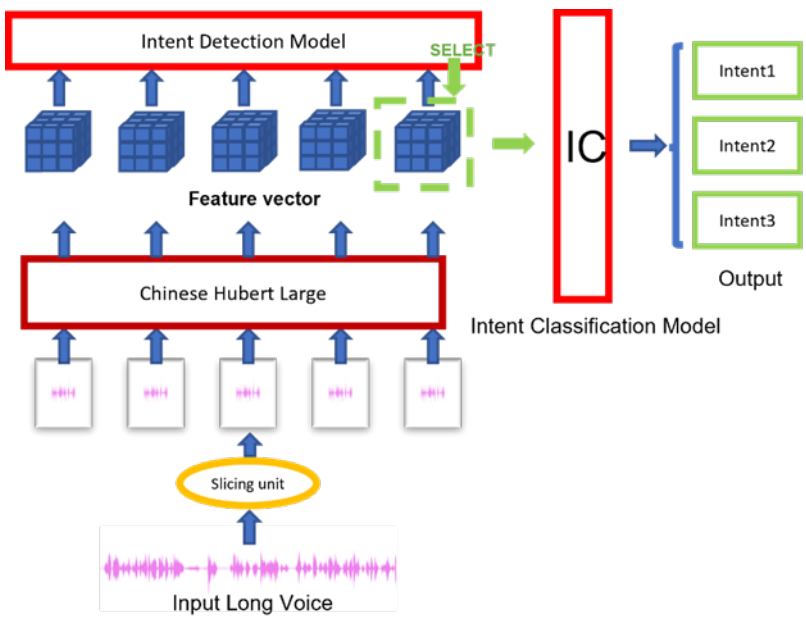

圖 4. 方法流程圖
[Figure 4. Method Flow Chart]

後續我們對方法流程進行了優化，會在雙模型組合策略中詳細進行說明。

3.2 HuBERT進行特徵提取 (HuBERT for Feature Extraction)

我們選用在 WenetSpeech (Zhang *et al.*, 2022)的 10,000 小時的中文語料上進行訓練的 HuBERT (Hsu *et al.*, 2021)對輸入的語句進行特徵提取。HuBERT 是語音的自監督學習模型的其中一個，使用聚類的方式從音檔數據中對音檔數據進行分類和標記，再使用類似 BERT 的 mask 方式將其中一部分標籤作為需要預測的對象，從而使模型得到訓練，這使得 HuBERT 在處理語音相關問題上面具有通用性。

我們為了最大程度的利用 HuBERT 提取的特徵，基於 SUPERB (Yang *et al*., 2021) 的研究中發現 HuBERT 的每一層輸出都有有價值的信息，所以我們取其每一層的輸出進行加權(weighted sum)後再進行下游模型的訓練，其中的權重(weight) 值會和下游模型一起訓練，這樣使得我們在訓練下游模型的過程中，模型可以自動選擇要提取 HuBERT 模型中特定層的資訊。

3.3 意圖偵測下游模型 (Intent Detection Downstream Model)

意圖偵測下游模型的功能，是判斷給定的特徵向量中是否包含意圖。模型輸入由 HuBERT 上游模型抽取的音檔的特徵向量，輸出存在意圖的機率。在意圖偵測模型的設計中，我們會傾向於選擇一種簡單的架構，因為可能意圖無關的語句會比較多，訓練資料不足時很容易造成模型過擬合 (overfitting) 的狀況。在解決消防局無線電長語音中的意圖偵測時，我們選擇了一種結構較為簡單的下游模型，我們是參考 SUPERB (Yang *et al*., 2021) 中解決關鍵詞發現 (keyword spotting) 任務的下游模型設計的。模型架構如圖 5。

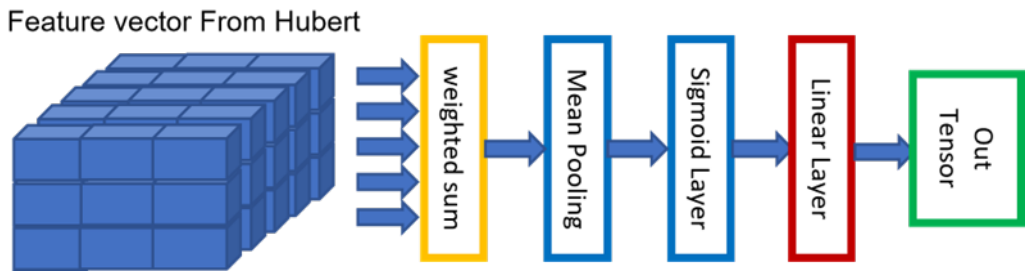

圖 5. 下游模型架構
[Figure 5. Downstream Model Structure]

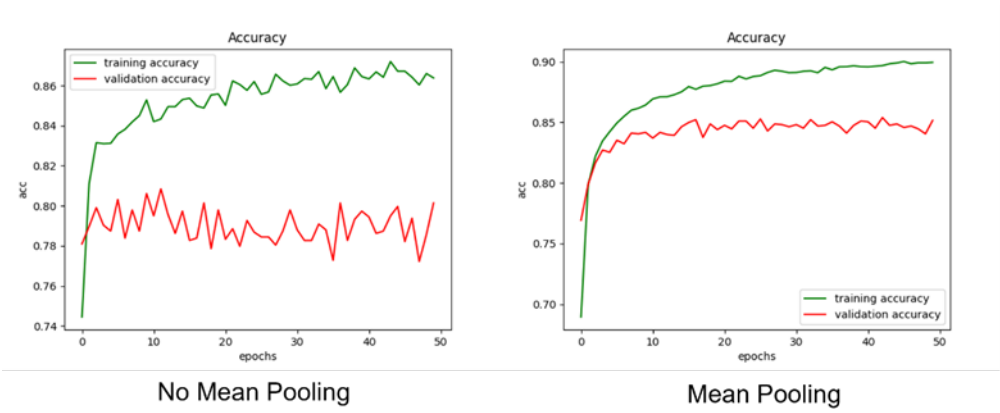

圖 6. 是否做均值池化之實驗結果對比
[Figure 6. Comparison of Experimental Results Between Whether to Do Mean Pool-ing]

我們通過 weighted sum 的方式使用 HuBERT 的每一層的輸出，對 HuBERT 的每一層進行提取和利用可以加快下游模型的訓練速度和得到更好的最終結果。均值池化 (mean pooling) 層的作用是從 weighted sum 後的資訊中提取特徵。

根據內部資料的實驗結果（圖 6），均值池化層可以降低輸入參數的複雜度，從而增強模型對意圖相關特徵的提取能力。我們採用了 4*4 的內核大小 (kernel size)， 在針對複雜度高的任務，可以減少內核大小以從特徵向量中獲取更多的資訊。在進行均值池化 (mean pooling) 後，採用一個激活 (sigmoid) 層和一層線性 (linear) 層就可以完成預測，進行輸出。

3.4 意圖分類下游模型 (Intent Classification Downstream Model)

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。我們在後續的實驗中，有與意圖偵測下游模型相同的較為簡單的下游模型。也有在 (Borgholt *et al*., 2021) 中提出，由兩個 linear 層和 Bidirectional LSTM (Graves & Schmidhuber, 2005) 構成的下游模型。

分類問題往往會遇到樣本不均衡的情況，我們採用 class-balanced loss (Cui, Jia, Lin, Song, & Belongie, 2019) 進行訓練，通過前期內部資料集的實驗證明（圖 7）可以有效的減少各個類別之間正確率的差距。

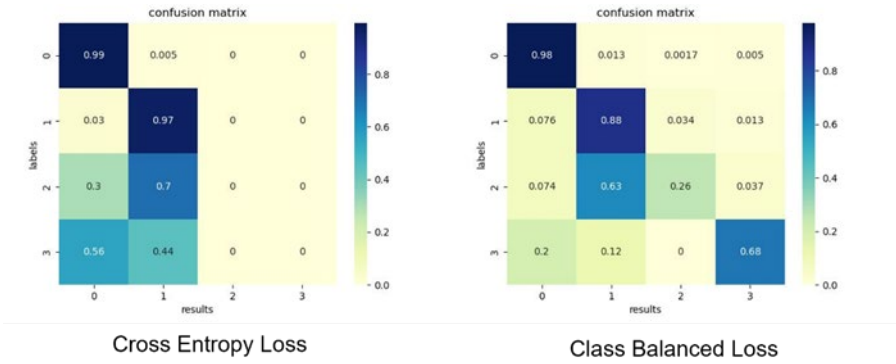

圖 7. 不同損失函數的比較
[Figure 7. Comparison of Different Loss Functions]

3.5 雙模型組合策略 (Bi-model Combination Strategy)

我們的模型在進行辨識時，若提供的音檔過長，模型無法學習到資料的特徵。我們解決此問題的方法是，將其按照固定長度和步進剪輯後分段送入偵測模型進行預測，會產生多段輸入的短音檔。

在之前的流程中，意圖偵測模型選取存在意圖機率最高的特徵向量，輸入意圖分類模型進行意圖的分類。我們發現如果意圖偵測模型產生錯誤的預測，則會導致任務整體的錯誤，意圖偵測模型的預測結果缺少容錯性。所以我們調整了雙模型的組合策略，將意圖偵測模型認為可能存在意圖的特徵向量全部送入分類模型進行預測。將意圖偵測模型給定音檔存在意圖的機率乘上意圖分類模型給出的存在前三大類別分別的機率，並將所有特徵向量的結果在每一個類別上進行加總，就可以算得該類別的意圖存在機率。組合策略圖見圖 8，其中 ID 表示意圖偵測模型 (intent detection model)，IC 表示意圖分類模型 (intent classification model)。意圖存在機率計算公式如下：

$$pk = \frac{\sum_{a=1}^{n} p_{a1} * p_{ka2}}{n}$$

其中 k 是指第 k 類的意圖，a 指切割後的第 a 段音檔，n 指切割後的音檔總數量，pa1 指意圖偵測模型給定 a 段音檔存在意圖的機率，pka2 指意圖分類模型給定 a 音檔存在 k 意圖類別的機率。通過此方法可以算得存在某個類別意圖的機率，我們取存在機率最高的前三大類別意圖進行輸出。

我們在新北市消防局無線電語料長語音意圖辨識任務上進行了實驗證明其有效性。

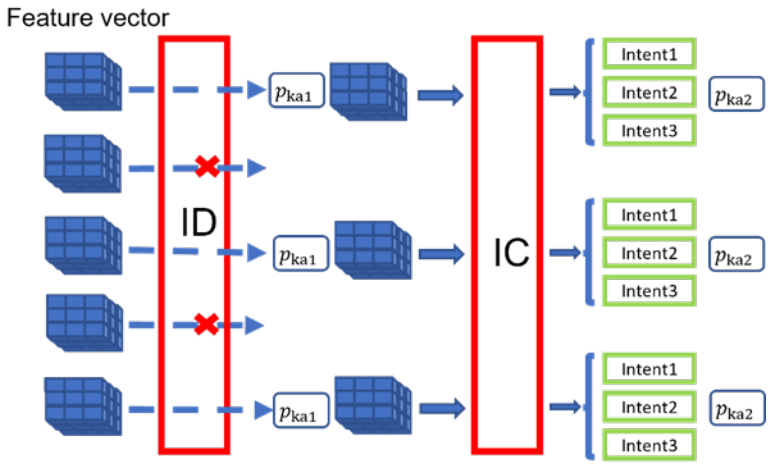

圖 8. 整合策略流程圖
[Figure 8. Integration Strategy Flowchart]

4. 實驗 (Experiments)

4.1 新北市消防局無線電語料長語音意圖辨識實驗(Experiments of Long Speech Intent Recognition in Radio Corpus of New Taipei City Fire Department)

4.1.1 資料整理 (Data Arrangement)

新北市消防局提供給我們總共一年的無線電錄音，其中包含大量無效音檔。我們對其進行清洗，刪除了無人聲、過短無意圖的音檔後再進行人工註記。我們註記時同時標記意圖出現的位置和意圖的類型，意圖類型根據消防局使用需求和資料的分佈狀況共分為 16 類，每類數量都不均勻，數量最多的有 2,891 個，最少只有 2 個。每個類別的語句數量如圖 9 所示。我們將資料每類均衡的劃分為訓練集、驗證集和測試集，劃分後的資料如表 1 所示。

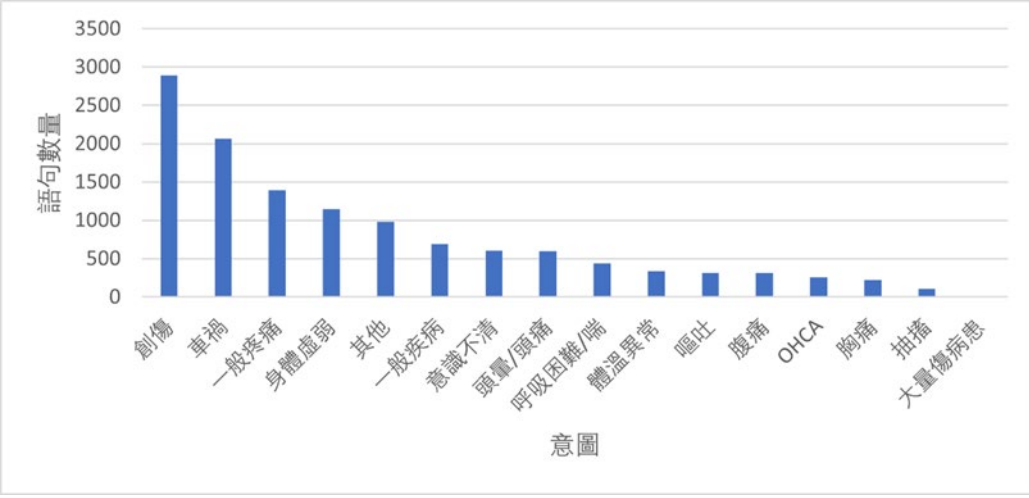

圖 9. 消防局無線電語料意圖類別分佈圖
[Figure 9. Intention Category Distribution Map of Fire Department Radio Corpus]

表 1. 新北市消防局無線電語料長語音資料集劃分
[Table 1. Division of Radio Corpus Long Voice Data Set of New Taipei City Fire Department]

資料集	語句數量	語句時長（小時）
訓練集	11372	57.80
驗證集	403	2.02
測試集	588	3.17

因為消防局無線電的語音皆為長語音，在整理意圖偵測模型的訓練資料時，我們將每一段訓練資料按照 5 秒一段 2.5 秒的步進 (step) 長度切成了若干段，將包含意圖的段落的標記 (label) 設置為 1，沒有包含意圖的段落設置為 0。使得我們的資料中有若干標記為 1 的資料和若干標記為 0 的資料。

整理意圖分類模型的資料時，因為原本音檔資料過長，有很多無關的資料，我們按照任務中標記的意圖出現的位置，以意圖出現的位置為中心，剪取 5 秒長度的音檔作為訓練資料，對應的標記則為這段音檔資料對應的意圖。

4.1.2 模型訓練 (Model Training)

因為新北市消防局無線電語料的意圖類別分佈不均勻，所以我們在訓練時採用 class-balanced loss (Cui *et al*., 2019)。同時在此資料集上達到更好的效果，我們參考 ASR 的資料增強 (data augmentation) 方法 (Ko, Peddinti, Povey, & Khudanpur, 2015) 採用資料變速的資料增強方法，分別將語句變速 0.9 和 1.1 倍。意圖偵測和意圖分類模型我們都採用 linear model。我們設定批次大小 (batch size) 為 32，使用 ADAM (Kingma & Ba, 2014) 優化器，學習率 (leanrning rate) 設定為 1e-4，同時引入有 warm up 的 scheduler 幫助模型快速收斂，訓練的 epoch 數為 30，模型已幾乎完全收斂。

4.1.3 實驗設計 (Experimental Design)

此任務的實驗設計的目的是選出一個較好的方法，提供解決此任務的 Base- line，因為資料的稀少和可能存在的錯誤，所以此實驗的正確率數值並不會如開源資料集上的實驗那麼高。

本實驗將我們提出的長語句意圖辨識方法與先前文獻中的端到端意圖辨識方法以及傳統的 ASR+BERT 的方法進行了對比實驗，並以正確率作為評估指標， 比較各方法的性能優劣。

其中的 ASR 方法我們使用 OpenAI 開源的 Whisper (Radford *et al*., 2022) 多語言 ASR 模型，是目前在多個開源模型上進行對比選擇的對複雜環境魯棒性最好的模型。因為無線電語音主要內容為中文，所以我們的 BERT 模型採用由 Hugging Face 訓練的 BERT-Base-Chinese 模型(Wolf *et al*., 2019)，我們用 BERT 進行短文本分類任務的方法參考了(Sun, Qiu, Xu, & Huang, 2019)。

之後會進行優化後的雙模型組合策略和不優化的對比實驗，通過調整預測流程和剪輯音檔時的步進長度來使得整體的三意圖正確率得到更好的數值，使用三意圖正確率作為評估標準的原因是這樣更接近真實用途。

為保證公平，在進行實驗時，所有自監督學習模型都是 Freeze 的，包括 Whisper，不會進行微調 (fine-tuning)。

4.1.4 實驗結果 (Experimental Results)

本實驗首先將單模型方法，即利用一個下游模型進行短語音意圖分類的方法，與本研究方法進行比較，此方法在表格中標示為 HuBERT+ 單下游模型。其次，本研究與傳統的兩步式意圖辨識方法，即先用 Whisper 進行語音辨識，再用 BERT 進行意圖辨識的方法，進行比較，此方法在表格中標示為 Whisper+BERT。本研究採用兩個下游模型分別執行

意圖偵測和意圖辨識的任務，此方法在表格中標示為 HuBERT+ 雙下游模型。評估標準是單意圖正確率 (acc)。結果如表 2。

表 2. 消防局資料集不同方法對比實驗結果
[Table 2. Comparison of Experimental Results of Different Methods in Fire Station DataSet]

方法	單意圖正確率 (%)
Whisper+BERT	28.9
HuBERT+ 單下游模型	4.2
HuBERT+ 雙下游模型	**38.5**

從實驗結果可以看出，我們的方法在環境嘈雜的情況下，在長語句中對意圖偵測和分類的能力與 Whisper+BERT 方法相比 ERR (error reduction rate) 为 33.2%。由於單一個下游模型無法同時定位和辨識長語音中的意圖，因此在處理長語句的意圖分類問題時，其效能最低。

同時我們對我們的方法和 Whisper+BERT 方法進行推理時單位小時的資料所需算力進行了評估。我們對模型進行算力評估時，所用的電腦硬體環境為 I7-12700 的處理器，64GB 記憶體，一張 RTX3080ti 顯示卡。我們通過模型在此環境中進行一個小時的語料推理需要運行的時間來評估其所需算力。結果如表 3。可以看出我們的方法所用算力相比於使用 Whisper+BERT 方法減少了 91.4%。

表 3. Whisper+BERT 方法與 HuBERT+ 雙下游模型所需算力對比
[Table 3. Comparing the Computing Power Required by the Whisper+BERT Method and the HuBERT + Bi-downstream Model]

方法	推理時間（分鐘）
Whisper+BERT	8.87
HuBERT+ 雙下游模型	0.77

本研究在原有實驗的基礎上，採用改進的雙模型組合策略進行實驗。透過調整流程方法和切割音檔時的步進長度，提升模型的整體正確率。原始的流程 s1 是將偵測模型選出最可能含有意圖的特徵向量，交由分類模型預測並輸出機率最高的三個類別。改進的流程 s2 是將所有可能含有意圖的特徵向量都送入分類模型預測，計算每個類別的意圖存在機率並輸出最高的三個類別。

我們通過調整模型運作流程和剪輯窗口的步進長度，得到了三意圖正確率最高的實驗結果。如表 4。

表 4. 策略調整實驗結果
[Table 4. Experiment Results of Strategy Adjustment]

流程	步進長度（秒）	三意圖正確率（%）
s1	2.5	47.8
s1	1	48.2
s2	2.5	47.8
s2	1	56.2

實驗結果顯示，在採用原始的預測流程時，調整剪輯窗口的步進長度對模型輸出的正確率無顯著影響。然而，在採用新的流程後，透過調整步進長度，使模型對每一段可能含有意圖的聲音剪輯都進行評分，並乘以意圖出現的機率。步長減少之後，參與評分投票的聲音片段數量增加，增強了含有意圖的聲音片段對最終分類結果的影響力。我們推斷是多次預測和投票的機制提升了雙模型整合運作時的效能，優於原始的流程。

對比於最直接的流程，我們對於流程和步進長度的調整之後，模型整體的正確率得到了提升，與一般的流程相比 ERR 為 17.6%。

在這 16 類意圖中「其他」這類意圖因為包括的語音資訊較為複雜，會對整體的結果造成影響，且在實際應用中並沒有作用，我們嘗試移除此類別的資料，再進行辨識，我們最終的輸出三意圖的正確率達到了 64.6%，可以在實際應用中發揮一定的作用。

4.2 Mobvoi Hotwords 中文關鍵詞發現實驗 (Experiments of Mobvoi Hotwords Chinese Keyword Spotting)

4.2.1 資料整理 (Data Arrangement)

在 Mobvoi Hotwords 資料集中，語句分為三類：你好問問、嗨小問、無關鍵詞。其測試集的數據分佈如表 5。我們將你好問問和嗨小問都劃分為同一類（有關鍵詞）無關鍵詞的語料劃分為一類，用這樣的資料訓練偵測模型，用來判斷語句中是否含有關鍵詞。然後我們單獨將有關鍵詞的語料提取出來，重新劃分為你好問問和嗨小問兩類，用來訓練二分類模型區分不同的關鍵詞。

表 5. Mobvoi Hotwords 測試集統計
[Table 5. Statistics of Mobvoi Hotwords Test Set]

關鍵詞	語句數量	語句時長（小時）
無關鍵詞	52,111	62.95
嗨小問	10,641	5.51
你好問問	10,641	5.64

4.2.2 模型訓練 (Model Training)

我們設定了最長輸入的語音為 5s，長語音會被切成多段進行輸入，短語音則補靜音到指定的長度。我們的意圖偵測模型和意圖分類模型都採用線性模型的下游模型。我們設定批次大小(batch size)為 32，使用 ADAM 優化器，學習率(leanrning rate)設定為 1e-4，同時引入有 warm up 的 scheduler 幫助模型快速收斂，訓練的 epoch 數為 30，模型已完全收斂。

4.2.3 實驗設計 (Experimental Design)

我們將我們的雙模型方法應用於關鍵詞發現任務上面，測試其對特定詞語的識別能力，並將我們的結果和此資料集上的最佳方法進行對比，評估方法為資料集原始論文提到的每小時錯誤響應次數 (false alarm per hour, FAH) 和 (false rejection rate, FRR)。

我們也進行了使用單一模型進行偵測和分類的對比實驗。同時將我們的結果和資料集原始論文 (Hou *et al*., 2019) 中效果最好的 RPN 方法進行對比。

4.2.4 實驗結果 (Experimental Results)

實驗結果如表 6。我們的方法 FAH 只比 RPN 方法高了 18.8% 的情況下，將 FRR 降低了 73.2%。同時相比於原先使用單下游模型的結果，使用我們的雙下游模型方法可以得到更低的 FRR，證明我們的方法在其他的複雜語音任務中也能有很好的表現，其具備一定的通用性。

表 6. 中文關鍵詞發現任務對比實驗結果
[Table 6. Chinese Keyword Spotting Task Comparative Experiment Results]

方法	FAH	FRR
RPN	5.00	2.20
HuBERT+ 單下游模型	**3.46**	2.44
HuBERT+ 雙下游模型	5.94	**0.59**

4.3 CATSLU 短語句分類實驗 (CATSLU Short Speech Sentence Classification Experiments)

4.3.1 資料整理 (Data Arrangement)

在 CATSLU 資料集中，語句分為五類，其中有一類雜訊，四類包含意圖。我們對其各個類別的意圖數量進行了統計，如表 7 所示。

表 7. CATSLU 資料集各個類別語句數量和時長統計
[Table 7. Statistics of the Number and Duration of Sentences in Each Category of the CATSLU Dataset]

類別	語句數量	語句時長（小時）
雜訊	3861	2.77
地圖	1672	1.47
音樂	1789	1.58
影片	1247	1.19
天氣	1676	1.38

採用雙下游模型進行辨識時，我們先將資料按照包含或不包含意圖分為兩類，訓練意圖偵測模型用來偵測是否為雜訊。再在有意圖的資料中根據意圖的類型分為四類，訓練意圖分類模型。

4.3.2 模型訓練 (Model Training)

在下游模型的選擇上，我們採用了更為複雜的模型。在進行 weighted sum 之後，我們將特徵向量用線性模型(linear model) 將其映射到 256 維，之後輸入隱藏層大小為 256 的一層 Bidirectional LSTM 層 (Graves & Schmidhuber, 2005)，之後對輸出經過均值池化(mean pooling) 後經線性模型 (linear model) 輸出。Whisper+BERT 的方法我們則採用完整音檔先通過 Whisper 轉譯成繁體中文，再輸入由 BERT 和一層線性層組成的文本分類模型進行分類。

4.3.3 實驗設計 (Experimental Design)

我們進行了三個方法的實驗，和新北市消防局無線電語料長語音意圖辨識實驗的方法相同。此資料集中的語音吐字較為清晰，且環境噪音少，且為短語句，使用 ASR 可以獲得較好的轉譯結果。

4.3.4 實驗結果 (Experimental Results)

實驗結果如表 8 所示。從表中可以看出，HuBERT+ 單下游模型的方法在處理乾淨的短語音的分類時，比使用 Whisper+BERT 的方法正確率要略高一些，我們在中文資料集上驗證了先前的研究(Borgholt *et al.*, 2021) 在英文資料集中的結論。

且我們使用的雙下游模型方法，比單下游模型的端到端意圖辨識方法具有更高的正確率。不只是在複雜的長語音意圖辨識任務中，在這種開源的短語音分類任務中，我們使用的雙下游模型方法也可以有很好的表現。

表 8. CATSLU 資料集之實驗結果
[Table 8. Experimental Results on the CATSLU Dataset]

方法	正確率 (%)
Whisper+BERT	85.7
HuBERT+ 單下游模型	89.1
HuBERT+ 雙下游模型	**89.4**

5. 結論與未來方向 (Conclusions and Future Directions)

我們提出的長語句端到端意圖辨識的方法，在三個語音任務上，相較於之前的方法都得到了提升。在意圖辨識任務中與 Whisper+BERT 方法相比，所需算力得到了明顯下降，更有希望在嵌入式設備和終端設備中運作。基於 CATSLU 的短語音分類任務可以作為一個評價中文端到端意圖辨識模型優劣的語音任務。在之後的工作中，我們將探索其他下游模型以更好的利用語音自監督模型提取的特徵，同時嘗試在嵌入式設備和終端設備上搭載我們的模型並進行評估。

參考文獻 (References)

Borgholt, L., Havtorn, J. D., Abdou, M., Edin, J., Maaløe, L., Søgaard, A., & Igel,

C. (2021). Do we still need automatic speech recognition for spoken language understanding? arXiv preprint arXiv:2111.14842.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (CVPR), 9268-9277. https://doi.org/10.1109/CVPR.2019.00949

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks, 18(5-6), 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

Hou, J., Shi, Y., Ostendorf, M., Hwang, M.-Y., & Xie, L. (2019). Region proposal network based small-footprint keyword spotting. IEEE Signal Processing Letters, 26(10), 1471-1475. https://doi.org/10.1109/LSP.2019.2936282

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3460. https://doi.org/10.1109/TASLP.2021.3122291

Jung, N., Kim, G., & Chung, J. S. (2022). Spell my name: keyword boosted speech recognition. In 2022 ieee international conference on acoustics, speech and signal processing (ICASSP 2022), 6642-6646. https://doi.org/10.1109/ICASSP43922.2022.9747714

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In Sixteenth annual conference of the international speech communication association, 3586-3589. https://doi.org/ 10.21437/Interspeech.2015-711

Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., & Bengio, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. arXiv preprint arXiv:1904.03670.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision (Tech. Rep.). Technical report, OpenAI, 2022. Retrieved from https://cdn. openai. com/papers/whisper. pdf.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classifi- cation? In Proceedings of China national conference on chinese computational linguistics, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T., Gugger, S., … Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., & Lee, H.-y. (2021). Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051.

Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., Xu, X., Bu, H., Chen, X., Zeng, C., Wu, D., & Peng, Z. (2022). Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In Proceedings of 2022 ieee international conference on acoustics, speech and signal processing (ICASSP 2022), 6182-6186. https://doi.org/10.1109/ICASSP43922.2022.9746682

Zhu, S., Zhao, Z., Zhao, T., Zong, C., & Yu, K. (2019). Catslu: The 1st chinese audio-textual spoken language understanding challenge. In Proceedings of 2019 international conference on multimodal interaction, 521-525. https://doi.org/10.1145/3340555.3356098