國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

邁向趨近人類的表徵學習：語音的非監督式句法剖析與音訊影像表徵的泛用性探討

Towards Human-like Representation Learning: Unsupervised Syntax Parsing of Speech and General-Purpose Audio-Visual Representations

曾元

Yuan Tseng

指導教授: 李琳山 教授

Advisor: Lin-shan Lee, Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202500549

國立臺灣大學碩士學位論文

口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE

NATIONAL TAIWAN UNIVERSITY

（論文中文題目）(Chinese title of Master's Thesis)

邁向趨近人類的表徵學習：語音的非監督式句法剖析與音訊影像表徵的泛用性探討

（論文英文題目）(English title of Master's Thesis)

Towards Human-like Representation Learning: Unsupervised Syntax Parsing of Speech and General-Purpose Audio-Visual Representations

本論文係曾元（R11942082）在國立臺灣大學電信工程學研究所完成之碩士學位論文，於民國 113 年 7 月 9 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Communication Engineering on 09/07/2024, have examined a Master's Thesis entitled above presented by Yuan Tseng (R11942082) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李琳山　曹昱　陳尚澤

（指導教授 Advisor）

賴穎暉　王新民　李宏毅

所長 Director:

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

摘要

「預訓練再微調」（pretrain-then-finetune）這一套訓練方法套用在語音辨識、語者驗證等不同語音處理任務中，都被證實有不錯的效果。在與自監督式學習（self-supervised learning）結合後，這套方法除了顯著地提升效能以外，也為語音科技帶來其他重要的效益，包括減少模型對標註資料的需求，以及簡化不同任務間模型的架構差異。這也顯示不久的未來有機會實現能夠從大量無標註資料與一些標注資料學習，並有能力同時處理多任務、多模態的一個通用模型，讓語音科技向實現人類等級模型的目標更進一步。

本論文中探究分析了延伸模型能力之深度與廣度的兩個方向的嘗試：首先，本論文提出一非監督式語音句法剖析任務，以探討在沒有成對資料的情況下，能否直接從語音得到一段語句之句法結構。實驗顯示在缺少成對資料的情況下，從口述語句得到正確的句法剖析樹極為困難。即便如此，模型仍展現出具備初步的判斷訓練資料的語言之分支結構的能力之跡象。

其次，本論文在多模態、多任務的更大框架下比較現有的自監督式訓練架構，檢驗現有的訓練架構是否能夠泛用在各種語音及音訊處理任務上。對一影音輸入，一個模型共可以取得音訊、影像、及混合三種內部表徵。接著以單一表徵作為輸入，對每一語音及音訊處理任務去訓練一個小模型，探討模型表徵的泛用性。在評估五個近期提出的模型後，結果顯示並沒有任一單一模型可以適用在所有任務

doi:10.6342/NTU202500549

上。

透過研究範圍更大、難度更高的任務，本論文希望探索現有自監督式表徵學習的一些可能性與局限性，並希望朝向更像人類能力之通用模型的目標前進。

關鍵字：自監督式學習，成分句法剖析，音頻-影像學習

doi:10.6342/NTU202500549

Abstract

The pretrain-then-finetune approach has been shown to be an effective direction for speech processing, with successful results in speech recognition, speaker verification, and a wide variety of other speech-related tasks. Combined with self-supervised learning, the paradigm brings major attractive advantages to speech technologies in addition to improved task performance, including reducing the dependency on large quantities of labeled data, and simplifying the task-specific components. This implies we are one step closer to constructing human-like models, able to perform different multi-modal tasks by learning from vast amounts of unlabeled data plus some limited labeled data.

This thesis focuses on two different directions towards the above goal: First, the unsupervised spoken constituency parsing task is proposed to examine the possibility of learning high-level linguistic structural information, such as syntax, directly from speech without any paired data. Experiments show that while it is still difficult at this moment for machines to learn to produce correct syntax trees from speech without any supervision, the

doi:10.6342/NTU202500549

model does indicate some initial evidence of being able to learn the branching direction of the language used for training.

Second, existing self-supervised audio-visual learning frameworks are broadly examined under a wider multi-modal, multi-task framework to determine how capable the existing approaches are on five speech and audio understanding tasks. For each model, three types of internal representations are obtained from auditory, visual, and both inputs, respectively. Next, model performance is measured by finetuning a small prediction head for each task, using each type of representation as input. The results of such an unified evaluation show that no single model can sufficiently generalize to all tasks.

By analyzing the applicability of self-supervised learning approaches to more difficult and broader tasks, this thesis aims to demonstrate the potential and shortcomings of existing technologies, in order to facilitate more research towards human-like audio-visual learning.

Keywords: self-supervised learning, constituency parsing, audio-visual learning

doi:10.6342/NTU202500549

目次

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

圖次

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

表次

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

第一章　導論

1.1　研究動機

現有計算系統對人類認知（human cognition）過程的模擬，仍存在極大的不足之處。譬如在常見的監督式學習（supervised learning）架構中，通常需要針對特定任務收集大量的輸入－輸出成對資料，並利用梯度下降（gradient descent）等演算法訓練一個類神經網路（neural network）模型將輸入映射到輸出。然而，這種學習架構與人類認知系統有著數個本質上的區別：

首先，以聲學為例，先進的語音辨識模型需要先接收數千小時的高品質語音－文字成對資料，才能精準辨認所聽到的語音之中的內容。相比之下，牙牙學語的嬰兒在學會識字以前便能純靠聽覺學習，可能僅用數百小時或更少的語音便足以讓嬰兒具備類似的識別能力。至於更高階的語言構成要素，如語意、語法等，現有的計算系統大多是從文字資料中習得。但世界上許多非書面語言的存在，顯示人類僅依賴語音輸入就能掌握語言的文法規則和語義理解能力。這也張顯人類語言習得（language acquisition）過程與現有計算系統的顯著區別。

另外，研究上通常將聽覺任務、視覺任務劃分為兩個不同的研究領域。甚至眾多的聽覺任務中，語音處理任務和音訊處理任務也一向被分開處理。過去在這樣的劃分下所建立的模型，往往僅能以單一模態的資料處理單一的任務，無法像

doi:10.6342/NTU202500549

人類一般，以一套包含感官與大腦的完整系統去處理不同模態資訊，廣泛的解決不同型態的任務。以語音辨識、音訊分類兩項聽覺任務為例，主流模型往往由完全不同架構組成，缺乏模型間的共通性，而人類則能以一套統一的聽覺系統接收所有聽到的聲音，再將其轉換成大腦能解讀的形式。

而自監督式表徵學習（self-supervised representation learning），恰是一個具潛力解決計算系統上述的兩個缺陷——模型需要大量成對資料做監督式學習，以及缺乏泛用的音訊－影像處理（audiovisual processing）模型——的架構：在優化自監督式的損失函數（loss function）過程中，自監督式模型能夠從沒有標記的資料，學到如何以資訊豐富的表徵來代表輸入。利用自監督式模型學習到的表徵，我們得以在更少的標註資料的情況下，建立比純監督式學習表現更好的模型。因此本論文希望探索如何利用自監督式模型去更好的模擬人類語言習得的過程，以及建立能更廣泛套用在不同聽視覺任務上的表徵模型。

1.2 研究方向

本論文描述了朝向開發更接近人類感知能力的計算模型所做的兩個嘗試：

第一、嘗試在只有沒有文字標註的語音資料與額外的非成對文字資料的情況下，試圖**建構一個有能力從語音學習到語法結構的模型**。我們以語言學中的句法剖析（constituency parsing）任務作為出發點，定義語音上的句法剖析，並且提出一個衡量正確性的指標。本論文將探討兩種不同的方式去處理這個問題：一者以非監督式的方法先從語音轉寫為文字，再從文字得到語句的句法樹（constituency tree）。另一者以自監督式表徵模型從語音抽取表徵，將語音分段，直接從每段的表徵取得句法樹。

第二、**建立一個多任務、多模態的評量基準** (benchmark)，**衡量音訊 - 影像表**

doi:10.6342/NTU202500549

徵模型的泛用性。近年有不少研究開始探討能否建立音訊－影像表徵模型，利用兩個模態的資料去更好的解決某些任務，但模型往往仍被侷限在特定的任務上。本論文提出評量基準的目標，是希望促成更多建構泛用的音訊－影像表徵模型的探討，以及更進一步研究語音處理和音訊處理之間是否存在著共通性等等。

1.3 研究貢獻

本論文的主要研究貢獻包括以下三點：

1. 本論文提出語音上的句法剖析任務，探討從非成對的語音、文字資料習得語言的文法結構之可行性。
2. 本論文提出一個非監督式的學習架構，不需要過渡階段將語音轉寫成文字，便能夠直接從自監督式語音表徵計算句法剖析樹。
3. 本論文系統性地比較五種不同自監督式音訊－影像表徵模型於五種多模態音訊處理及語音處理任務之表現，並發現現有音訊－影像表徵模型大多仍缺乏泛用性，無法有效套用在不同的音訊、影像處理任務上。

1.4 章節安排

本論文的章節安排如下：

- 第二章：閱讀本論文所需的背景知識。
- 第三章：描述對於口述語句之非監督式句法剖析的嘗試。[1]

[1]此章節內容基於作者於 ICASSP 2023 發表的論文

doi:10.6342/NTU202500549

- 第四章：介紹對現有音訊－影像表徵模型的多模態、多任務評測過程。[2]
- 第五章：本論文之結論與未來可能的研究方向。

[2]此章節內容基於作者於 ICASSP 2024 發表的論文

doi:10.6342/NTU202500549

第二章　背景知識

2.1　監督式學習

給定一個想解決的任務，若以 x 表示任務所提供之一筆輸入資訊，y 表示任務所需之一筆輸出資訊，則在監督式學習的框架下，會預設存在 N 筆成對資料 $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$，供模型作為學習的對象。以音訊事件分類（audio event classification）為例，x_i 即是數位音訊檔案，而 y_i 則是其對應的音訊類別。而監督式學習的目標便是利用 N 筆成對資料優化一個模型，使模型有能力在以 x_i 作為輸入時，模型預測 $\hat{y}_i$ 能夠盡量接近對應的正確輸出 y_i。

模型在優化過程（又稱為訓練過程、學習過程）中，必須定義一目標函數 L（objective function）來衡量模型預測與真正輸出之間的差距，此目標函數又被稱為損失函數（loss function）。每當模型接收到一筆輸入 x_i，首先會得到模型的預

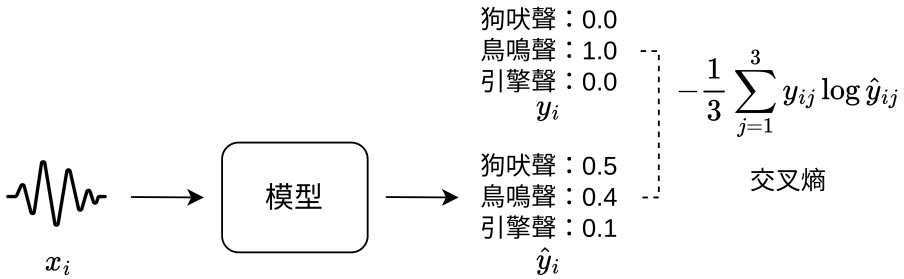


圖 2.1: 監督式學習示意圖，以音訊事件分類為例，假設輸入的音訊資料可能來自三個類別：狗吠、鳥鳴、及引擎聲。其中模型使用交叉熵（cross-entropy）作為損失函數 L。

doi:10.6342/NTU202500549

測 $\hat{y}_i$，接著算出模型預測 $\hat{y}_i$ 與真正輸出 y_i 之間的損失 $L(y_i, \hat{y}_i)$。得到 $L(y_i, \hat{y}_i)$ 後，便能夠調整模型以縮小損失值，降低預測值與真正輸出的距離。不過實務上為了提升訓練穩定性，通常會以數筆資料作為一個批次（batch），根據批次裡的平均損失值調整模型。在迴歸（regression）任務中，y_i 為連續數值，因此可以使用 L_2 或 L_1 距離作為損失函數。而分類任務中，常使用獨熱編碼（one-hot encoding）來表示離散的類別 y_i，以便於使用交叉熵（cross-entropy）函數作為損失函數。以圖2.1為例，圖中的音訊事件分類模型錯誤地將鳥鳴聲預測為狗吠聲。若以 $\hat{y}_{ij}$ 表示模型預測第 i 筆輸入資料屬於第 j 個類別之機率，此時的損失值可以表示為：

$$-\frac{1}{3}\sum_{j=1}^{3} y_{ij} \log \hat{y}_{ij} = -\frac{1}{3}\{0 \times \log 0.5 + 1 \times \log 0.4 + 0 \times \log 0.1\}$$

$$\approx 0.305$$

假設訓練一段時間後，模型預測的狗吠聲機率變成 0.4，鳥鳴聲機率變成 0.5，引擎聲機率不變，此時會預測出正確的分類，損失值為：

$$-\frac{1}{3}\sum_{j=1}^{3} y_{ij} \log \hat{y}_{ij} = -\frac{1}{3}\{0 \times \log 0.4 + 1 \times \log 0.5 + 0 \times \log 0.1\}$$

$$\approx 0.231$$

這顯示目標函數的設計需要讓越正確的預測之損失值，比越錯誤的預測之損失值小，才能讓模型在訓練過程中，越來越進步。

doi:10.6342/NTU202500549

2.2 深層類神經網路（Deep Neural Networks）

2.2.1 簡介

在電腦視覺領域中，克式（Alex Krizhevsky）於 2012 年所提出的 AlexNet 模型於影像分類上取得了重大的突破。AlexNet 被提出之前，當時普遍認為即使通用近似定理（universal approximation theorem）保證任何一個函數都存在一個近似的深層類神經網路，但實際上卻很難透過優化得到這個近似類神經網路，因此不容易將深層類神經網路有效地應用在生活場景中。而 AlexNet 借助圖形處理器（graphic processing unit）的平行化能力顯著地加速深層類神經網路的優化過程，並大幅超越傳統方法在影像分類上的表現後，便打破了深層類神經網路不實用的迷思，也帶動了往後十餘年深層學習（deep learning）的蓬勃發展。

深度類神經網路由許多層類神經網路堆疊組成。以下將會簡述幾種常用的類神經網路架構：全連接類層（fully-connected layers）、卷積層（convolutional layers）、遞迴層（recurrent layers）、以及轉換器層（transformer layers）。

2.2.2 全連接類層

全連接層為最基礎的類神經網路架構，由許多個全連接層堆疊而成的深層類神經網路又被稱為多層感知器（multi-layer perceptrons）。一個輸入長度 m 的向量 $\boldsymbol{x}$，輸出長度 n 的向量 $\hat{\boldsymbol{y}}$ 的全連結層裡，輸入輸出之間的關係可以表示為：

$$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

doi:10.6342/NTU202500549

其中 $\boldsymbol{W}$ 為一個 $n \times m$ 大小的矩陣，$\boldsymbol{b}$ 為一個長度 n 的向量，σ 為一個非線性函數，又被稱為激發函數（activation function）。

2.2.3 卷積層

卷積層的概念源自於信號處理中的濾波器，會將輸入與一個核（kernel）進行計算。以一維資料為例，一個一維卷積層會有兩組參數 $\{\boldsymbol{W} \in \mathbb{R}^k, \boldsymbol{b} \in \mathbb{R}\}$，其中 k 為核的大小。經過卷積層運算後，若輸入向量 $\boldsymbol{x}$ 長度為 l，則輸出向量 $\hat{\boldsymbol{y}}$ 長度為 l'，其中 $l' = \lfloor \frac{l-k}{s} + 1 \rfloor$，其中 s 為卷積的步幅（stride）。計算的詳細過程可以表示為：

$$\hat{\boldsymbol{y}}_i = \sum_{j=1}^{k} \boldsymbol{W}_j * \boldsymbol{x}_{s(i-1)+j} + \boldsymbol{b}$$

其中當第 $s(i-1)+j$ 個元素超過輸入 $\boldsymbol{x}$ 長度時，可以用零或是最近的元素填補（padding）。當然，以上的計算可以自然延伸到多維資料的情境。

使用卷積層的類神經網路模型又被稱為卷積類神經網路 (convolutional neural networks, CNN)。由於卷積運算僅考慮範圍內的鄰近資料，故卷積層輸出裡的任一元素只和一部份輸入有關聯，因此卷積層可以比全連結層更容易學到輸入的局部特徵。

2.2.4 遞迴層

上文所述的全連接層與卷積層類神經網路的一個共同特性是：它們的輸出皆只和輸入有關，因此輸出的每個元素之間彼此獨立。這類類神經網路又被稱為前饋式類神經網路（feedforward neural network）。但是，當任務中預測目標的每個元素之間具有因果關係，譬如一句話裡面的前後文，前饋式類神經網路便無法有效模擬，但遞迴類神經網路（recurrent neural network）便可以。

doi:10.6342/NTU202500549

這是因為遞迴類神經網路所指的正是輸出同時與輸入和其他輸出元素有關聯的類神經網路。最簡單的遞迴層又被稱為艾式網路（Elman network）。在艾式網路裡，過去的輸出資訊也會作為未來輸出元素的輸入，關係式如下：

$$\boldsymbol{h}_t = \sigma_h(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}_h)$$

艾式網路又可以與全連接層結合，成為喬式網路（Jordan network）。喬式網路的表示式則為：

$$\boldsymbol{h}_t = \sigma_h(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h \hat{\boldsymbol{y}}_{t-1} + \boldsymbol{b}_h)$$

$$\hat{\boldsymbol{y}}_t = \sigma_{\hat{y}}(\boldsymbol{W}_{\hat{y}} \boldsymbol{h}_t + \boldsymbol{b}_{\hat{y}})$$

上方兩遞迴類神經網路表示式中，下標 t 代表時間 t，x_t 為時間 t 時的輸入向量，而 y_t、h_t 分別為時間 t 時的全連接層輸出及遞迴層輸出向量，$\boldsymbol{W}_h$、$\boldsymbol{W}_y$、$\boldsymbol{U}_h$ 為可訓練的矩陣，$\boldsymbol{b}_h$、$\boldsymbol{b}_{\hat{y}}$ 為可訓練的向量。

然而，以梯度下降法訓練艾式網路與喬式網路時，模型會需要將梯度從序列最尾端反向傳播（backpropagate）到序列最初端，因此處理較長的序列時，往往會遇到梯度消失（gradient vanishing）或是梯度爆炸（gradient explosion）的問題。因此，後續研究提出了閘門循環單元（gated recurrent unit, GRU）與長短期記憶體（long short-term memory, LSTM）遞迴類神經網路，在模型中添加了「閘門」的設計，允許模型在學習過程中自動調節過去資訊的比重，或甚至完全捨棄不重要的過去資訊。

doi:10.6342/NTU202500549

2.2.5 轉換器層

轉換器（Transformer）在 2017 年被瓦氏（Ashish Vaswani）等人提出。在原始論文中，瓦氏等人發現轉換器在機器翻譯（machine translation）任務上表現優於遞迴類神經網路與卷積類神經網路，後續更是有許多研究發現轉換器架構同樣適用於電腦視覺、語音處理、自然語言處理的其他任務，譬如近年絕大多數的語言模型（language model）皆採用轉換器的變體作為模型架構。

轉換器與前文所述的幾種類神經網路最重要的核心區別在於多頭自專注機制（multi-head self-attention）。多頭自專注機制的輸入為一個向量的序列 $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T$，每個向量維度為 d，並共有 h 個頭。對於第 t 個輸入 $\boldsymbol{x}_t$，多頭自專注機制中會計算它與每一個輸入的專注權重（attention score），並按照權重將輸入加權平均得到第 t 個輸出 $\hat{\boldsymbol{y}}_t$。更具體的來說，多頭自專注機制中的每一個頭都會各自以三個大小為 $d_h \times d$ 的不同矩陣 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 將 $\boldsymbol{x}_t$ 線性投影成三個不同的向量 $\boldsymbol{x}_{Qt}$、$\boldsymbol{x}_{Kt}$、$\boldsymbol{x}_{Vt}$，分別稱為查詢向量（query vector）、鑰向量（key vector）、和值向量（value vector），其中 d_h 為每個頭的輸出向量長度，$d_h = d/h$。

以第一個頭所進行的計算而言，線性投影完後，第一個頭會根據所有的查尋向量與鑰向量計算專注權重。其中第 t 個查尋向量與第 s 個鑰向量的專注權重所代表的是兩者之間的關聯性大小，會決定第 t 個輸出向量中，來自第 s 個輸入的資訊所佔的比例。若將所有鑰向量視為一個大小 $T \times d_h$ 的矩陣，第 t 個輸入的專注權重之表示式為 $\text{softmax}(\frac{\boldsymbol{x}_{Qt} \cdot \boldsymbol{x}_K^T}{\sqrt{d_h}})$，則第一個頭的第 t 個輸出則可以被表示為：

$$\text{softmax}(\frac{\boldsymbol{x}_K \boldsymbol{x}_{Qt}}{\sqrt{d_h}})\boldsymbol{x}_V$$

最後整個多頭自專注機制的第 t 個輸出（維度為 d）由每個頭的第 t 個輸出（維度為 d_h）串接而成。多頭自專注機制之於單頭的自專注機制的優勢在於：當輸入向

doi:10.6342/NTU202500549

量之間有好幾種不同類型的關聯性時，多頭自專注機制中的每一個頭可以分別處理其中一種關係。

由於多頭自專注機制不考慮向量在輸入序列中的位置，因此無法有效處理位置資訊。故在轉換器論文中，除了多頭自專注機制以外，還會透過位置嵌入（positional embedding）的方式將向量在輸入序列中的位置對應到一個連續的向量，並加入到第一層轉換器層的輸入。

在原始論文中，轉換器為一編碼器－解碼器（encoder-decoder）架構，因此轉換器層又分為轉換編碼器層（transformer encoder layer）與轉換解碼器層（transformer decoder layer）兩類。由於轉換解碼器被設計根據其過去的輸出遞迴地產生下一個輸出，故轉換解碼器層的多頭自專注機制會被加上額外限制，將 $\boldsymbol{x}_t$ 對於所有 $\{\boldsymbol{x}_{t'}|t' > t\}$ 的專注權重設定為零。

2.3 自監督式學習（Self-Supervised Learning, SSL）

自監督式學習是一種非監督式的學習架構。在沒有標註資料的情況下，自監督式學習以預測部份輸入資料的變形作為訓練時的目標函數。過去研究發現，在進行監督式訓練前，先透過自監督式的訓練目標預先訓練模型（pretraining），可以用更少的標註資料在目標任務上取得更好的表現 [5, 23]。

以下將簡述近年提出的一些處理語音和多模態資料的自監督式學習方法，以及其中所使用的模型架構與目標函數。

2.3.1 自監督式語音模型

本論文中使用到的自監督式語音模型包括 wav2vec 2.0[5] 與 HuBERT [23]。

doi:10.6342/NTU202500549

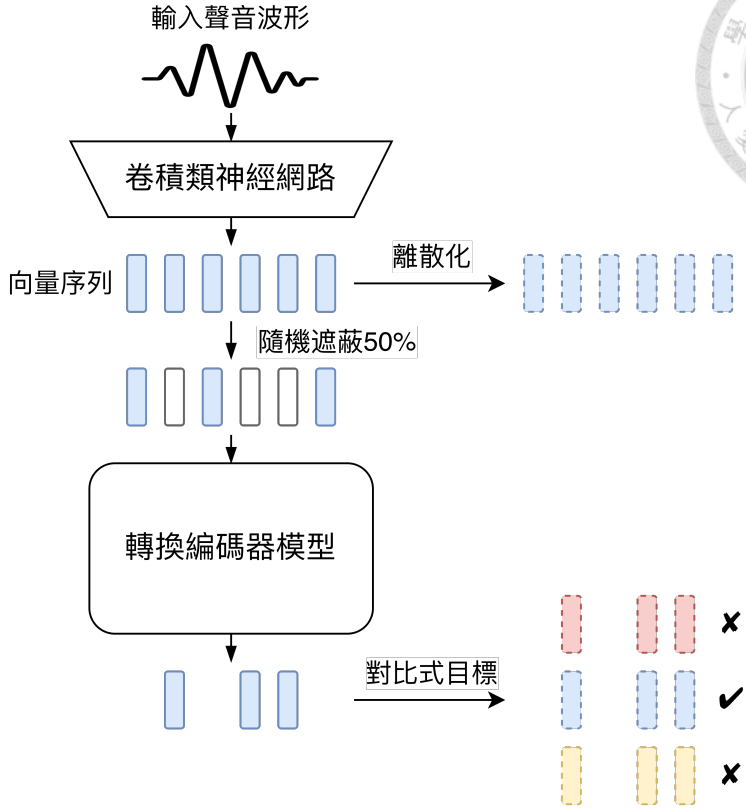


圖 2.2: wav2vec 2.0 架構及預訓練過程示意圖。其中以實線框長方形表示連續向量，虛線框長方形表示離散化的向量，「對比式目標」箭頭所指處之他色虛線框長方形表示對比式（contrastive）目標函數之負樣本（negative sample）。

如圖 2.2，wav2vec 2.0 的架構由一卷積類神經網路與數層轉換編碼器層組成。首先，卷積類神經網路以聲音波形作為輸入，將語音訊號下取樣至一個向量序列，其中每根向量會對應到 20 毫秒的聲波。接著，這些向量將會被離散化，變成離散的語音表徵，作為轉換編碼器層的預測目標。卷積類神經網路的輸出向量序列裡，會隨機選定一半的向量進行遮蔽，以一可訓練的遮蔽向量取代。被遮蔽過的向量序列會被作為轉換編碼器模型的輸入。wav2vec 2.0 模型的目標函數則正是以對比式學習（contrastive learning）的方式，從隨機選定的向量中，預測出遮蔽位置所對應的正確離散語音表徵。透過預測模型自身建立的離散語音表徵，模型在預訓練過程逐漸習得語音中的離散資訊。

在 wav2vec 2.0 論文中，使用 12 層轉換編碼器的模型在首先使用 Librispeech 資料集 [42] 中 960 小時的無標註語音資料做自監督式的預訓練後，再以僅僅 10 分

doi:10.6342/NTU202500549

鐘的標註資料進行監督式訓練，即可以在 test-clean 測試集得到 9.1 的語音辨識詞錯誤率（word error rate, WER）。相較之下，更早的研究在同樣的設定底下僅能達到 16.3 的詞錯誤率 [3]。

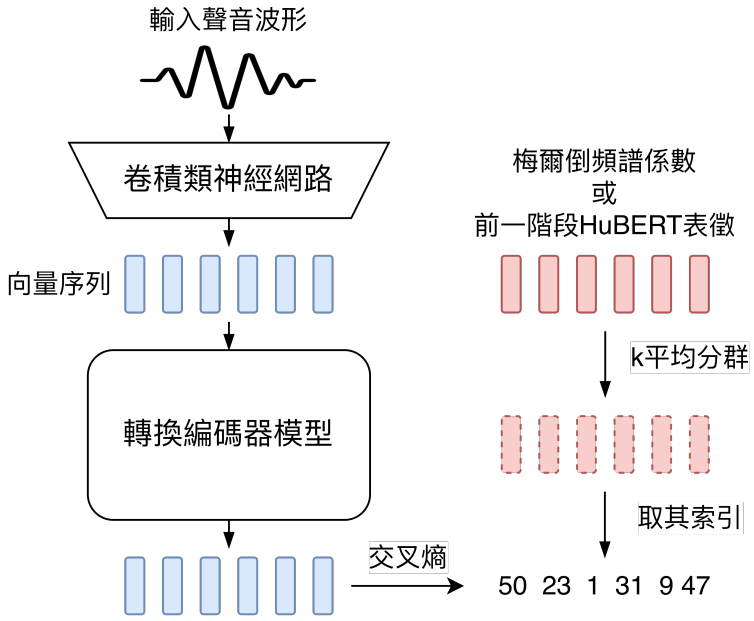


圖 2.3: HuBERT 架構及預訓練過程示意圖，其中實線虛線意義與圖2.2相同。HuBERT 採用階段性預訓練，圖中的紅色實線框向量在第一階段使用梅爾倒頻譜係數（mel-frequency cepstral coefficients, MFCC），後續階段則使用前一階段所預訓練的 HuBERT 模型轉換編碼器中間層的輸出。

如圖 2.3，HuBERT 的模型架構與 wav2vec 2.0 類似，預訓練過程使用的目標函數也是預測離散的語音表徵，但離散語音表徵的來源以及預測的方式不同。HuBERT 所預測的離散語音表徵會隨著預訓練的階段而有所不同。在初始的第一階段時，HuBERT 架構選擇將梅爾倒頻譜係數（mel-frequency cepstral coefficients, MFCC）以 k 平均分群演算法（k-means clustering）離散化，並以交叉熵損失函數使模型預測分群後的離散梅爾倒頻譜係數在碼書（codebook）之中的索引（indices）。第一階段預訓練完成後，模型的輸出表徵已蘊含一定量的資訊，但 HuBERT 論文中發現，可以將前一階段所預訓練的模型視為特徵抽取器（feature extractor），以模型轉換編碼器中間層所輸出表徵取代梅爾倒頻譜係數，並進行下一階段的預訓練。

doi:10.6342/NTU202500549

在 HuBERT 論文中，作者們發現在 960 小時無標註語音的自監督式預訓練，以及僅有 10 分鐘標註資料的監督式訓練後，使用 12 層轉換編碼器的第二階段 HuBERT 模型與同樣設定的 wav2vec 2.0 模型相比，語音辨識的表現近似。但當預訓練資料量級提升到六萬小時，並把模型放大至 24 層轉換編碼器，進行第三階段預訓練後，表現便會勝過同樣設定的 24 層 wav2vec 2.0 模型。

2.3.2 音訊 - 影像自監督式學習

認知心理學中著名的 McGurk 效應 [38] 顯示，我們人類從語音中聽到的內容，會被我們見到的資訊所左右。這不僅彰顯了人類認知中聽覺與視覺的密不可分，也啟發對於能夠同時從音訊和影像兩種模態的資料中學習的模型之研究。

以為唇語辨識任務設計的一模型為例，AV-HuBERT 模型是根據 HuBERT 架構所開發的音訊－影像自監督式模型。在預訓練階段時，AV-HuBERT 模型會同時接收頻譜圖和唇齒周圍的灰階影片，進行類似 HuBERT 模型的離散表徵預測。為了同時接收兩種模態的資料，AV-HuBERT 模型使用一卷積類神經網路將唇齒周圍影像處理成一影像向量序列，並以一線性轉換將頻譜圖轉換至一音訊向量序列。兩向量序列中的每一向量皆被設計成能夠對應到 40 毫秒的資訊，因此 AV-HuBERT 會先將對應同樣時段的音訊向量與影像向量串接後，輸入轉換編碼層做處理。

在預訓練過程中，AV-HuBERT 模型的預測目標同樣是預測 k 平均分群後的表徵之離散碼書索引。但是為了避免模型過於仰賴單一一種模態的資訊，預訓練過程中會採用一模態丟棄演算法（modality dropout），以 0.1 的機率將音訊向量序列或是影像向量序列設定為零。

以 LRS3-TED 資料集 [1] 中 433 小時的英文影音進行五階段的自監督式預訓

doi:10.6342/NTU202500549

練後，再以 30 小時的影片文字成對資料上監督式地訓練 AV-HuBERT 模型，可以在測試集上得到 51.8 的唇語辨識詞錯誤率。相較之下，更早的研究在同樣的設定底下僅能達到 71.9 的詞錯誤率 [36]。

2.4 成分句法剖析 (Constituency Parsing)

2.4.1 問題簡介

句法剖析這個任務的目的是分析一個句子的語法結構，並判定其是否符合語言學家所制定出語法規則。常見的語法規則包括成分句法（constituency grammar）與依存句法（dependency grammar）等，而本論文著重於成分句法剖析。

成分句法剖析假設任何一個語法通順的句子，都能以一個剖析樹（parse tree）表示其語法結構，而剖析樹裡的每一個節點則被稱為成分（又稱為詞組，constituent）。同類型的成分彼此替換後，仍可以得到語法上通順的句子。譬如在「今天的天氣真差」這一句子的剖析樹裡，如圖2.4所示，「今天」、「真差」、「今天的」都是成分，但「的天氣」則不是。透過句法剖析器（parser）得到的剖析樹，能夠以不同方式應用在語音合成（speech synthesis）、詞向量（word embedding）等任務上，包括提升模型的表現，以及增強模型的可解釋性（interpretability）。

2.4.2 非監督式成分句法剖析（Unsupervised Constituency Parsing）

由於語法規則具有高度複雜性，取得大量句子的剖析樹標註非常困難。因此，在缺乏標註資料的情況下，如何設計非監督式的訓練方式，使模型具備成分句法剖析的能力，便成為一個重要的研究課題。非監督式的句法剖析，被認為與人類

doi:10.6342/NTU202500549

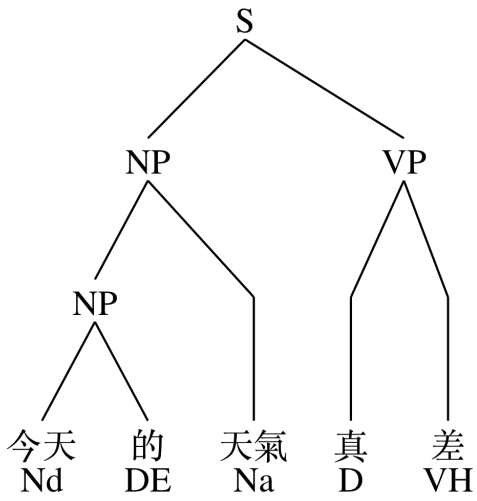


圖 2.4:「今天的天氣真差」這一個句子的剖析樹。根據其中「今天」、「天氣」為名詞,「真」為副詞,「差」為狀態不及物動詞。

學習語言的過程有關，因此又被稱為語法歸納（grammar induction）。為了降低任務的複雜度，過往研究大多限制剖析樹為二元樹（binary tree），假設不存在非投影樹（non-projective tree），同時也不判斷成分的詞性。

過去提出了許多不同的訓練方法來實現非監督式成分句法剖析。最早的研究以機率式無語境語法剖析器（Probabilistic Context Free Grammar, PCFG）以及其變形為主。在機率式無語境語法剖析器架構下，剖析樹裡的每種節點可能產生的所有分歧皆有一個機率值，以最大期望（expectation-maximization, EM）演算法調整。

而後續研究則又提出了五花八門的其他訓練方法，譬如：一些研究採用與語言學中成分定義相關的訓練方法，像是曹式 [9] 所提出的架構，先以語言學中的成分測試（constituency test）轉換輸入句子中的部份成分，再訓練一文法通順性模型（grammaticality model）檢查轉換後的句子文法是否通順，最後以文法通順性模型計算最通順的剖析樹。而另一些研究專注在設計具有樹狀的潛在變數（latent variable）之潛在結構模型（latent structure model）[31, 52]，以句子重組（sentence reconstruction）[16] 或是語言建模（language modeling）[47] 等非監督式任務作為

doi:10.6342/NTU202500549

訓練目標進行優化，最後再從訓練完畢的模型中產生出剖析樹。近年以類神經網路實現機率式無語境語法剖析器的研究 [30, 54, 57] 也與上述一類研究相關。

由於機率式無語境語法剖析器的框架中假設剖析樹節點類別有限多，無法直接套用於連續的語音輸入上，因此本論文第3章基於 DIORA 架構進行改良，使其適用於語音上的成分句法剖析。

doi:10.6342/NTU202500549

doi:10.6342/NTU202500549

第三章　探討自監督式模型能力之深度——語音非監督式成分句法剖析 (Unsupervised Spoken Constituency Parsing)

3.1　實驗動機

得益於近年自監督式表徵學習的突破，非監督式語音處理有了很大的進展。非監督式語音處理所研究的是：如何在沒有語音－文字成對資料的情況下，想方設法讓模型能夠從語音中學習有用的資訊。這和幼兒語言習得（child language acquisition）有許多相似之處。嬰幼兒在成長過程中，會漸漸從聽到的語音學習他們的母語，而非書寫語言（unwritten language）的存在，顯示這一個漸進式的語言習得過程是完全不需要語音－文字成對資料的。如果能讓計算系統具備類似的語言習得能力，便能將語音科技推廣到更多的低資源語言（low-resource language）。

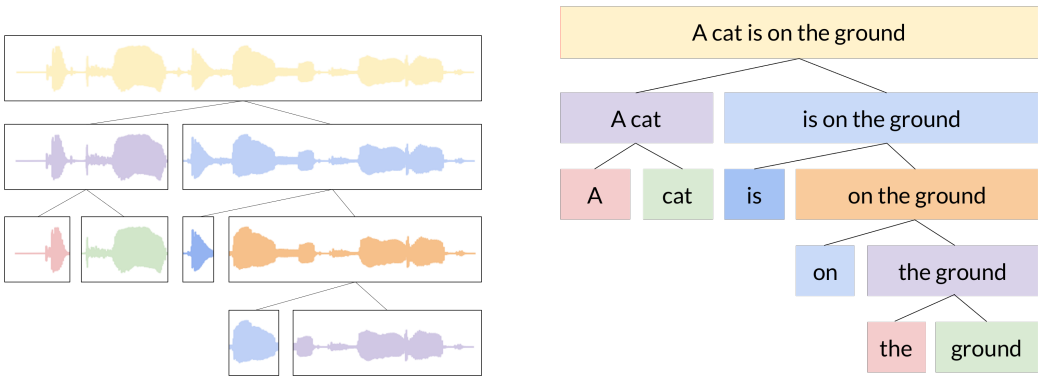
與語音波形或是傳統信號處理表徵相比，自監督式語音表徵與音素（phonemes）、詞彙（words）這些語言單位相關性更高 [17, 23, 40]。因此，以自監督式語音表徵為起點去進行音素或詞彙等級的非監督式語音任務，往往能得到更好的表現 [4, 35]。儘管如此，在語法（syntax）、語意（semantics）這類比較高階

doi:10.6342/NTU202500549

的語言構成要素上，自監督式語音表徵無法與文字語言模型內部表徵相比。過去研究顯示，自監督式語音表徵中語法、語意資訊含量較少，並同時和詞彙資訊交織，無法簡單的提取出 [46]。因此，在沒有成對資料的條件下，如何以非監督式的方式從語音中學習語法結構或是語意資訊，仍然是一個待解答的研究問題。

本章聚焦在如何從語音中以非監督式的方式學習語法結構的問題。在沒有成對文字資料的情況下，傳統的成分句法剖析評量指標便不再適用，因此本章重新定義語音上的成分句法剖析任務，並提出一評量指標來判斷語音剖析樹之正確率。另外，本章也提出兩種剖析器架構處理語音輸入：一者將語音辨識與文字句法剖析器進行串接，稱為串接式系統，另一者直接整合自監督式語音表徵與非監督式句法剖析架構，稱為直接式系統。

3.2 問題定義與正確性衡量



(a) 一口述句子之剖析樹　　(b) 圖 (a) 中對應文字之剖析樹

圖 3.1: 本章所定義的語音成分句法剖析樹。與文字上剖析樹類似，樹中每個不是葉端點（leaf）的節點皆會有兩個子節點（child node）。但在語音剖析樹中，每一個節點從一段文字，變成一段語音，相鄰的兩個子節點所對應的語音訊號串接起來即為它們的母節點所對應的語音訊號。

在本節中，我們將成分句法剖析樹的定義從文字延伸到語音輸入。在放寬的定義底下，剖析樹中的每個節點會對應到語音輸入的一段時間區間。因此，一語音成分句法剖析器必須首先將一段語音輸入進行分段（segmentation），再判斷哪

doi:10.6342/NTU202500549

幾段可以組合成更長的成分。

但是這時便會衍伸出一個問題：傳統上在衡量文字剖析樹的正確率時，最常使用的指標是模型預測的剖析樹與真正的剖析樹之間節點的精確率（precision）、召回率（recall）、與 F_1 分數（又稱為 PARSEVAL F_1[7]）。但這項指標假設剖析器的斷詞完全正確，無法反應剖析器將語音分段時發生的錯誤，因此不適合作為語音的成分句法剖析之衡量指標。

為了同時衡量剖析樹架構與分段的正確性，本章對 F_1 分數的計算方式進行調整。在比較模型預測結果與真正的剖析樹的結構之前，將會利用二分圖最大權重匹配（maximum weight bipartite matching）[19] 先找到兩個樹的端點之間最佳的一對一對應（one-to-one mapping），使得兩個樹間對應的端點之時間區段的重疊度最大化。計算端點之間的一對一對應關係，可以將節點之間的時間重疊度納入 F_1 分數的考量，使得模型預測結果與真正的剖析樹有更好的可比較性。而且，當模型預測結果與真正的剖析樹的端點完全相同時，這一指標便會與原始的 F_1 分數完全相同。

3.3 本章採用的剖析器架構

如第2.4.2節所述，雖然過去提出了許多不同的非監督式成分句法剖析器，但礙於語音具有連續性，因此無法使用預設剖析樹節點類別有限的框架來處理語音輸入。因此，本章採用改良後的 DIORA 剖析器架構 [16] 進行語句的剖析。

DIORA 是在 2019 年被德氏（Andrew Drozdov）等人所提出的一非監督式剖析器，由經過特殊設計的一編碼器與一解碼器組成，其計算過程與傳統的表格式剖析器（chart parser）類似。DIORA 的核心想法為：若不斷地將較短的成分合併成更長的成分，並以此方式壓縮句子，則依照剖析樹結構進行壓縮會保存最多資

doi:10.6342/NTU202500549

訊。

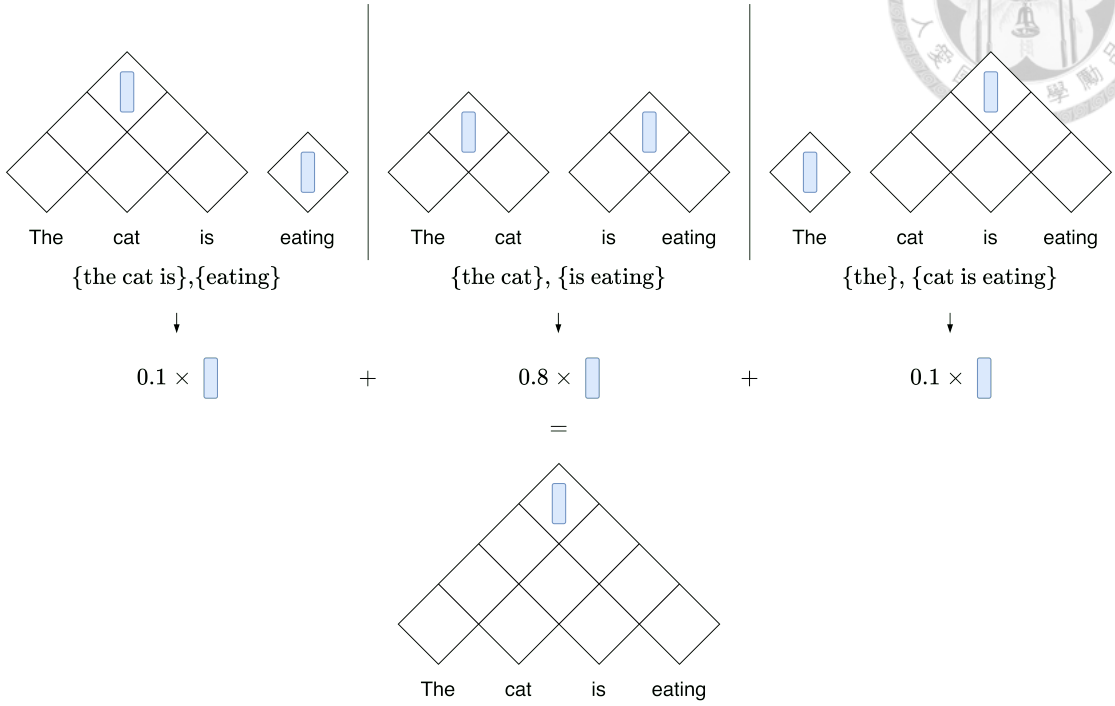


圖 3.2: DIORA 剖析器中編碼器之計算過程圖。表格裡的每一個格子皆對應到一個成分，而格子中的藍色長方形代表那一格所對應的成分之組合向量。

首先，編碼器會以詞向量（word embedding）序列作為輸入，由下而上地為每一可能的成分計算出一組合向量（composition vector），直到計算出代表整個句子的一個向量。一成分的組合向量是每個可能的剖析中，兩子樹之組合向量的加權平均。如圖3.2所示，「The cat is eating」這一成分共有三種可能的剖析。編碼器會對這三個剖析各自計算一向量與一相容性分數（compatibility score），並以相容性分數作為權重計算向量間的加權平均，得到「The cat is eating」這一成分的組合向量。

在編碼器計算出每個成分的組合向量以後，解碼器會反過來從剖析樹的根（root）開始，由上而下地為每一可能的成分計算出另一組合向量，直到計算出所有葉端點（leaf）的組合向量，再試著預測編碼器輸入的詞向量序列。解碼器中的組合向量是所有可能的相鄰節點的組合向量之加權平均。

doi:10.6342/NTU202500549

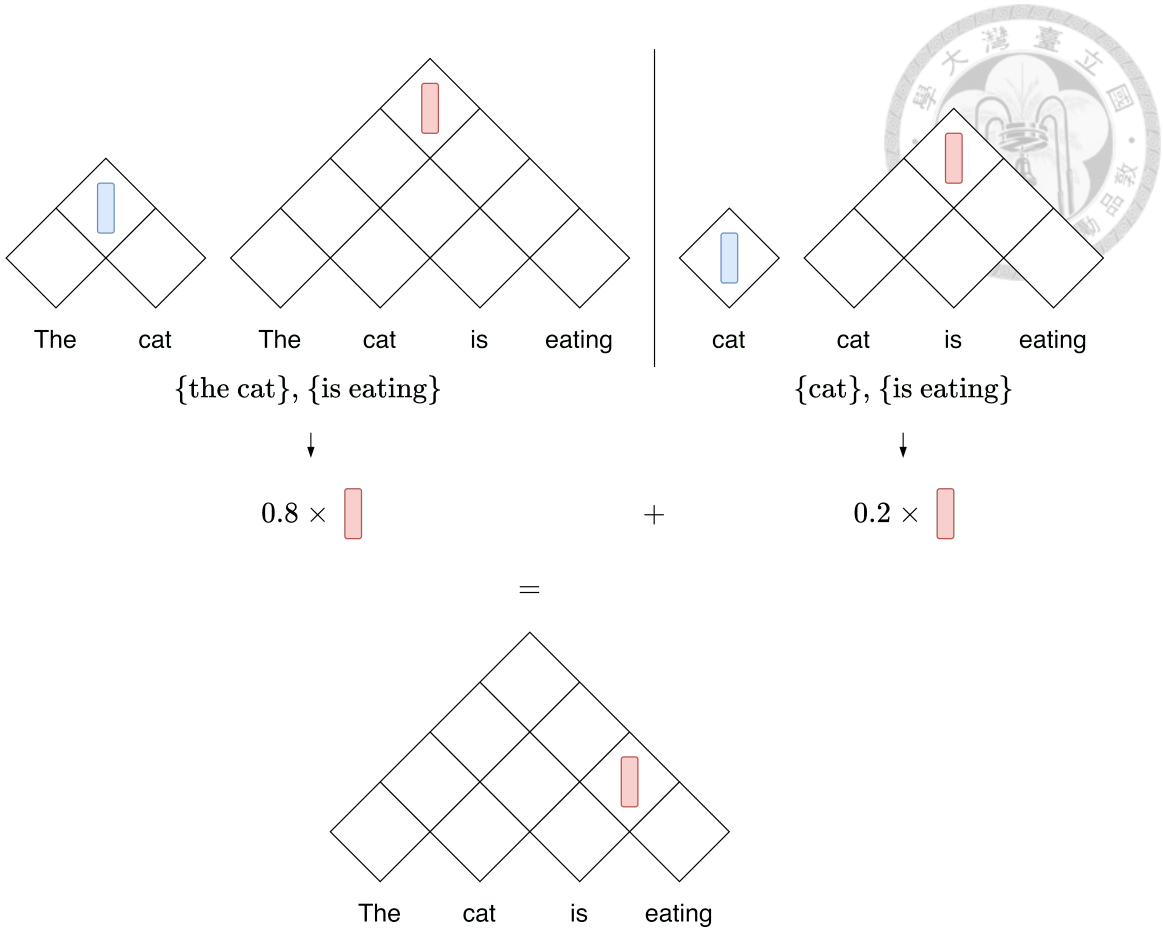


圖 3.3: DIORA 剖析器中解碼器之計算過程圖。表格裡的每一個格子皆對應到一個成分，而格子中的長方形代表那一格所對應的成分之組合向量。藍色長方形代表編碼器的組合向量，而紅色長方形代表解碼器的組合向量。

如圖3.3所示，「is eating」這一成分的相鄰節點有兩種可能，一者為「the cat」，另一者為「cat」。解碼器會對這兩個成分各自計算一向量與一相容性分數（compatibility score），並以相容性分數作為權重計算向量間的加權平均，得到「is eating」這一成分的組合向量。

最後，解碼器計算出一個葉端點的組合向量以後，便會計算一對比式損失函數，試著從隨機選定的向量中，正確預測出編碼器輸入序列中同個葉端點的詞向量。德氏等人發現在優化對比式損失函數的過程中，模型會漸漸學到如何根據剖析樹結構進行壓縮。訓練完模型後，便可以使用 CYK 演算法 [28, 55] 進行動態規劃（dynamic programming），從編碼器的權重中計算出機率最高的剖析樹。

doi:10.6342/NTU202500549

3.4 實驗方法

3.4.1 串接式系統：以語音辨識轉寫結果作為句法剖析器輸入

若要對語音輸入進行成分句法剖析，一個最直觀方式便是先用語音辨識將輸入轉寫成文字，再將轉寫結果作為一般的成分句法剖析器之輸入即可。但是，在沒有語音－文字或是語音－剖析樹這類成對資料的情況下，語音辨識和成分句法剖析都必須用非監督式的框架實現。因此我們語音辨識模型採用畢式（Alexei Baevski）等人延續本實驗室學長劉氏（劉達融）與陳氏（陳冠宇）等人所發展的架構方向 [11, 35]，進一步提出的 wav2vec-U 非監督式架構 [4]。

wav2vec-U 的框架中假設除了沒有標註的語音資料以外，還有額外的文字資料可以利用。由於自監督式語音表徵本身和音素有高度相關性，畢式等人發現可以用一系列的前處理步驟，將沒有標註的語音資料轉換成接近音素層級的表徵，再使用對抗式損失函數（adversarial loss）訓練一生成網路（generator），將這些表徵的機率分佈對應到文字音素序列的機率分佈，也就是從語音資料轉寫成音素序列。得到音素序列以後，便可以用 N 連詞串（N-gram）或是深層語言模型將音素序列轉換成詞彙，完成詞彙等級的非監督式語音辨識，本章將此設定稱為 UASR。另外，畢式等人還發現可以利用 wav2vec-U 的轉寫所得文字作為虛擬標註（pseudolabel），拿來作為另一模型的預測對象，用於進行監督式訓練，進而提升語音辨識的表現，本章將此設定稱為 UASR-ST。

為了觀察語音辨識錯誤率對成分句法剖析正確率之影響，本章將根據 wav2vec-U 的 UASR 與 UASR-ST 兩種設定，訓練兩個非監督式的語音辨識模型。從語音辨識模型得到轉寫結果後，轉寫結果裡的每個詞彙將被置換成對應的詞向量，作為剖析器之輸入。

doi:10.6342/NTU202500549

3.4.2 直接式系統：以語音表徵作為成分句法剖析器輸入

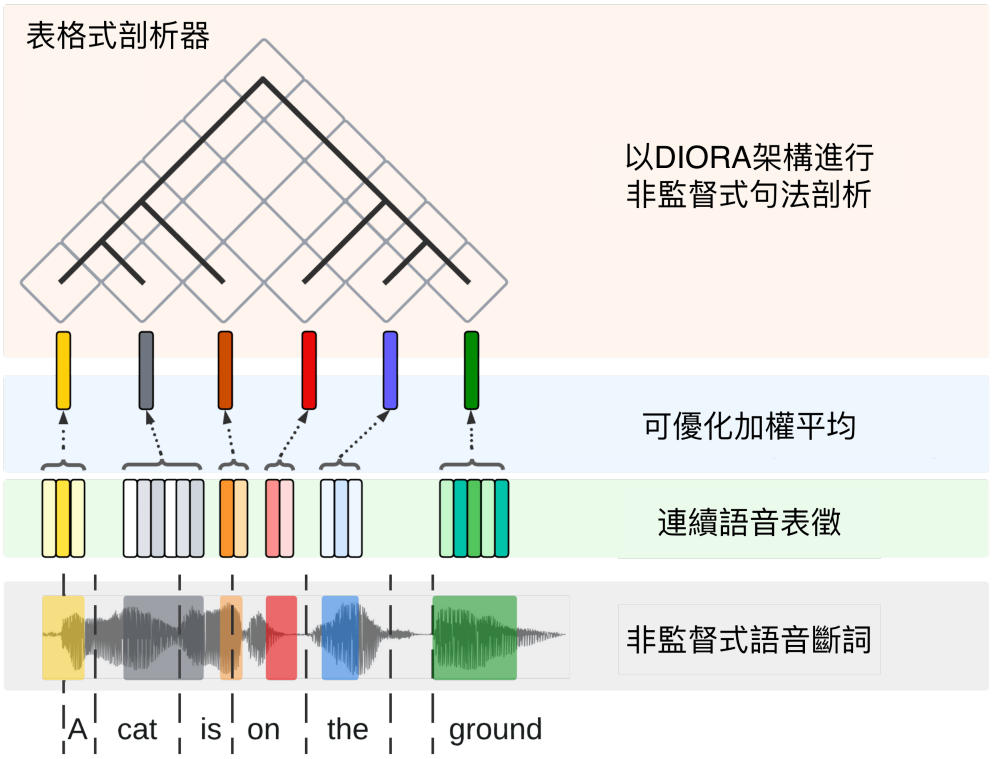


圖 3.4: 直接式系統之計算過程圖。將語音輸入以非監督式的方式進行斷詞後，系統會以可優化的加權平均計算出每一段語音之表徵，作為剖析器之輸入。輸入之轉寫文字僅作為圖解用。

由於剖析器之輸入為一向量序列，亦可以跳過文字轉寫步驟，直接從語音輸入得到詞等級（word-level）的連續表徵進行成分句法剖析。使用連續的語音表徵來表示輸入語句的目標是去更好地保留語音中的韻律資訊，讓剖析器可以做更好的預測。如圖3.4，直接式系統會先利用非監督式語音斷詞模型 [6, 27]，將語音進行分段。接著系統會用自監督式語音模型 [14] 計算出輸入訊號之連續語音表徵，並將屬於同一區段的表徵進行加權平均，得到詞等級的語音表徵。加權平均之權重由雙層多層感知器決定，而多層感知器之參數會與剖析器參數同時進行優化。

doi:10.6342/NTU202500549

3.5 實驗設定

為了有效衡量提出的兩種實驗方法，本章使用兩種高資源（high-resource）語言的資料來進行實驗，分別為 SpokenCOCO[24] 與 Zeroth-Korean[1] 這兩個資料集。SpokenCOCO 資料集包含 742 小時的英文語音資料，來自兩千餘個語者對 MSCOCO 資料集 [34] 裡的影像標題（image caption）之朗讀，其中每張影像對應到五個標題，而本章實驗僅使用錄製的音檔，不使用影像資料。本章跟隨前作採用 8.3 萬／5 千／5 千的影像切分 [56] 來分出訓練、驗證、與測試集，並用剩下的 3.1 萬張影像之標題文字作為額外文字資料訓練非監督式的語音辨識模型。而韓文實驗中使用的 Zeroth-Korean 資料集之訓練與測試集有 51.6 ／ 1.6 小時的韓文語音，來自 105 ／ 10 個語者。本章劃分出訓練集的 10 個語者作為驗證資料。

由於上述的兩個資料集僅有語音－文字的成對資料，沒有斷詞或是剖析樹的標註，因此本章使用一剖析器套件 [32] 對正規化（normalized）的文字標註進行成分句法剖析。在去除剖析樹中的標點符號後，便得到語音資料之剖析樹標註。而關於斷詞部分，本章使用一強迫對齊（forced-alignment）套件 [37] 來得到語音中每個詞所對應到的時間區段。

另外，在 DIORA 剖析器之原始論文中，德氏等人使用了另外訓練的 ELMo 詞向量 [44] 來提升模型正確率。由於 ELMo 詞向量需要大量文字作為訓練資料，為了更貼近低資源語言習得的情境，本章參考萬氏（Bo Wan）等人之實驗 [50]，使用隨機初始化的詞向量。因此本章實驗之超參數設置也與萬氏等人相同，批次大小為 32，學習率為 5×10^{-3}。

值得一提的是，在訓練剖析器時，本章採用過去研究的作法 [50, 56]，將訓練集中最常出現的一萬個詞彙建立詞表（vocabulary set），詞表以外的罕用詞（rare

[1] https://github.com/goodatlas/zeroth

doi:10.6342/NTU202500549

words）皆會被對應到同樣的一「未知」詞。初期實驗顯示，這一設定對結果有重要的影響，若不捨棄罕用詞，則剖析器表現會大幅降低，推測是因為罕用詞出現次數過少，對剖析器的學習造成了負面干擾。

模型選用（model selection）則是根據在驗證集上最小的損失函數來決定。串接式系統總共訓練 10 個訓練階段（epoch），而直接式系統收斂較快，因此僅訓練 2000 個批次。

3.6 實驗結果

3.6.1 串接式系統結果

如第3.4.1節所述，本章使用 100 小時的無標註語音和 15 萬句額外文字作為訓練資料，根據 wav2vec-U 的 UASR 與 UASR-ST 兩種設定去訓練非監督式語音辨識模型。訓練完語音辨識模型後，便可以對訓練集進行轉寫。最後兩模型在 SpokenCOCO 訓練集的轉寫結果之詞錯誤率（word error rate, WER）分別為 28.25% 和 13.15%。表3.1中第 (A)、(C)、(E) 列的結果顯示，隨著詞錯誤率越大，剖析器的正確性同樣會隨之下降。

在假設有額外文字資料的情況下，剖析器的訓練資料便有兩種選擇：以整個訓練集的轉寫文字訓練剖析器，或是用較少的額外文字資料訓練剖析器。表3.1中第 (B) ／ (C) 列與第 (D) ／ (E) 列間的分數差距顯示，即便整個訓練集的轉寫文字之資料量是額外文字資料的三倍以上，用較少的額外文字資料訓練剖析器仍然可以得到更好的結果。統計 UASR-ST 設定下的轉寫結果裡面詞彙之使用率可以發現，轉寫文字僅使用了原先訓練集的 1.6 萬個詞彙中的 8 千多個詞彙。因此使用轉寫文字作為訓練資料的剖析器無法處理罕用詞，便會導致較差的分數。

doi:10.6342/NTU202500549

	訓練資料	訓練文字屬性	測試文字屬性	F_1
(A)	整個訓練集	標準答案	標準答案	57.15 ± 2.09
(B)	額外文字資料	標準答案	UASR-ST 轉寫結果	44.08 ± 1.64
(C)	整個訓練集	UASR-ST 轉寫結果	UASR-ST 轉寫結果	40.53 ± 1.65
(D)	額外文字資料	標準答案	UASR 轉寫結果	34.97 ± 1.32
(E)	整個訓練集	UASR 轉寫結果	UASR 轉寫結果	31.01 ± 1.17

表 3.1: 串接式系統五次訓練之 F_1 分數平均與標準差，圖中以橫線分隔剖析器文字訓練資料之不同錯誤率。「訓練資料」一行所註記的是剖析器是使用額外文字資料還是訓練集的轉寫結果作為訓練資料。第 (A) 列列出使用完全正確的標準答案（ground truth）文字進行句法剖析的結果，是串接式系統的上限。

3.6.2 直接式系統結果

如第3.4.2節所述，直接式系統由三個部分組成：語音表徵，可優化的加權平均，以及非監督式的斷詞模型。關於表徵部分，本章以一多語言（multilingual）的 XLSR-53 模型之第 14 層轉換器層輸出作為語音表徵。XLSR-53 使用 wav2vec 2.0 架構模型，並以 53 種語言資料進行預訓練，因此具備處理不同語言的語音輸入的能力。

	斷詞方式		F_1
	訓練階段	測試階段	
(A)	標準答案	標準答案	57.11 ± 0.00
(B)	每 0.5 秒	標準答案	57.10 ± 0.01
(C)	每 0.5 秒	每 0.5 秒	3.88 ± 0.00
(D)	UAST-ST	UAST-ST	40.44 ± 1.72
(E)	UASR	UASR	28.49 ± 0.57

表 3.2: 使用不同斷詞方式的直接式系統之 F_1 分數。圖中第 (B) 列與第 (C) 列訓練時，將每 0.5 秒的語音輸入視為一段，而第 (D) 列與第 (E) 列使用語音辨識結果與輸入訊號的強制對齊作為斷詞方式。第 (A) 列列出完全正確的斷詞下的結果，為直接式系統的上限。

而關於斷詞模型而言，本章考慮了數種不同的斷詞方式。在初期實驗中，本

doi:10.6342/NTU202500549

章先嘗試以極其簡單的方式將每 0.5 秒的語音輸入進行分段，不意外地得到極差的結果（表3.2中第 (C) 列）。但是，當測試階段給模型提供正確的斷詞時（表3.2中第 (B) 列），即便在訓練階段只使用簡單方法作斷詞，模型表現仍然可以得到很大的提升。為了提升斷詞方式的精準度，初期實驗曾經以其他非監督式斷詞模型 [18] 進行分段，但結果不盡理想。因此，後續實驗以前一節中的非監督式語音辨識模型來提升斷詞精準度。在表3.2中第 (D)、(E) 列中，會將語音辨識模型之轉寫結果與輸入語音訊號進行強制對齊，得到轉寫結果中每個詞的時間區段，並以此進行後續計算。比較表3.2中第 (D) 列與表3.1中第 (C) 列，或是表3.2中第 (E) 列與表3.1中第 (E) 列可以發現，在僅使用分段資訊的條件下，直接式系統與串接式系統的表現是接近的。因此推斷當斷詞方法的正確性夠高，可能就足以讓直接式系統預測出精準的剖析樹。

3.6.3 直接式系統之分支方向

	英文	韓文
規則式		
左分支樹	24.68	27.15
右分支樹	57.11	7.60
直接式		
每 0.5 秒分段	57.10 ± 0.01	18.53 ± 8.99

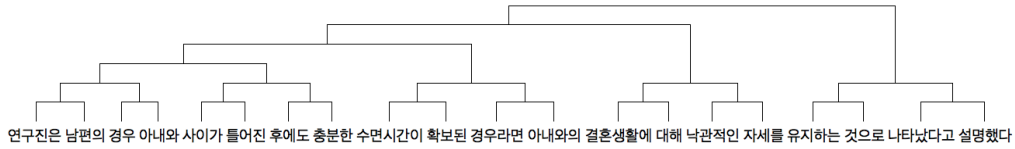
表 3.3: 將表3.2中第 (C) 列的直接式系統，與規則式（rule-based）系統在英文（右分支語言，right-branching）語音輸入和韓文（左分支語言，left-branching）語音輸入上的比較結果。

由於英語是中心語前置型（head-initial）的語言，英文的剖析樹在去除標點符號後，經常是向右分支的。相對來說，像中文、韓文、日文這種中心語後置型（head-final）的語言的剖析樹經常是向左分支的。

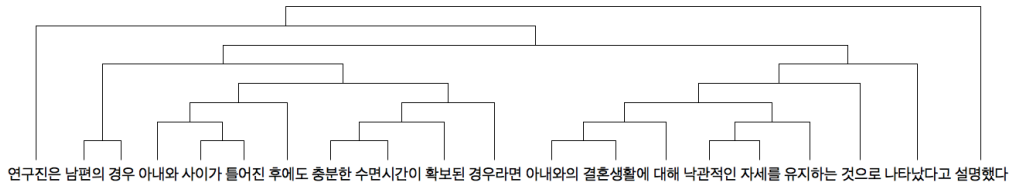
在 SpokenCOCO 資料集上進行初期實驗時，發現直接式系統在訓練後經常只

doi:10.6342/NTU202500549

會輸出向右分支的剖析樹。為了確認這一現象不是某種程度的模型崩潰（model collapse），我們在 Zeroth-Korean 這一韓文資料集上重複表3.2中第 (C) 列的實驗，並將結果列於表3.3。在五次訓練中，直接式系統有三次會傾向輸出類似左分支樹的結構（如圖3.5）。因此推測直接式系統雖然表現略低於串接式系統，在訓練的過程中仍然可以學到語言分支方向這一類的語法資訊。



（一）模型預測之一剖析樹舉例



（二）標準答案之一剖析樹舉例

圖 3.5: 在 Zeroth-Korean 資料集上，一模型預測之剖析樹與其對應的標準答案剖析樹。為了方便視覺化，圖中僅列出語音訊號所對應的文字節點。可以觀察到模型所預測的剖析樹有一定的左分支結構，符合韓文的語法。

3.7 本章結論

本章首次提出一語音上的成分句法剖析任務，並探討如何在非監督式的條件下完成它。本章提出了兩種方法實現這一目標：由語音辨識模型與文字剖析器組成的一串接式系統，與直接對連續語音表徵進行剖析的直接式系統。對串接式系統而言，本章發現在少量文字資料上訓練剖析器，效果會比在大量有錯誤的轉寫文字資料上訓練剖析器還要好。而對直接式系統來說，在同樣的斷詞下，實驗結果顯示直接式系統所預測的剖析樹正確率與串接式系統相近。同時，直接式系統的預測結果有學到語言分支方向的跡象。

doi:10.6342/NTU202500549

第四章　探討自監督式模型能力之廣度——模型於音訊 - 影像任務之效用評比

4.1　實驗動機

自監督式學習能夠讓模型從沒有標註的資料中，學到資訊豐富的表徵。過去已有多篇研究顯示，這些模型自動學到的表徵，往往比人工設計的傳統表徵更能泛用在各種任務上 [5, 15, 22]。儘管如此，這些研究中所探討的設定大多侷限在處理個別模態的資料。因此如何以自監督式的方式，讓模型能夠像人一樣從多模態資料中學習（multimodal learning），便成為一個很自然的延伸研究方向。

本章節將聚焦在多模態學習中的音訊－影像表徵學習（audio-visual representation learning）。在音訊處理及語音處理領域中，許多重要的任務能夠從聲音與影像之間的成對關係受益。以音訊事件分類為例，不同的聲音類別經常是由可以看見的不同發聲源所產生的，因此同時根據兩種模態的資訊進行音訊事件分類往往可以得到表現更好、更強健的模型 [25, 26]。再如語音辨識（speech recognition）中，語者的嘴形可以在吵雜環境中輔助模型進行更精準的辨識 [41, 45]。

儘管自監督式音訊－影像表徵學習已被證明在多個音訊和語音處理任務中

doi:10.6342/NTU202500549

有效，但目前仍不清楚現有框架是否像單模態自監督式表徵學習一樣具有普遍適用性。為了回答這個問題，本章透過五個多模態的音訊－影像下游任務（downstream tasks）來比較不同的音訊－影像表徵模型，建立一評量基準。

4.2 實驗設定

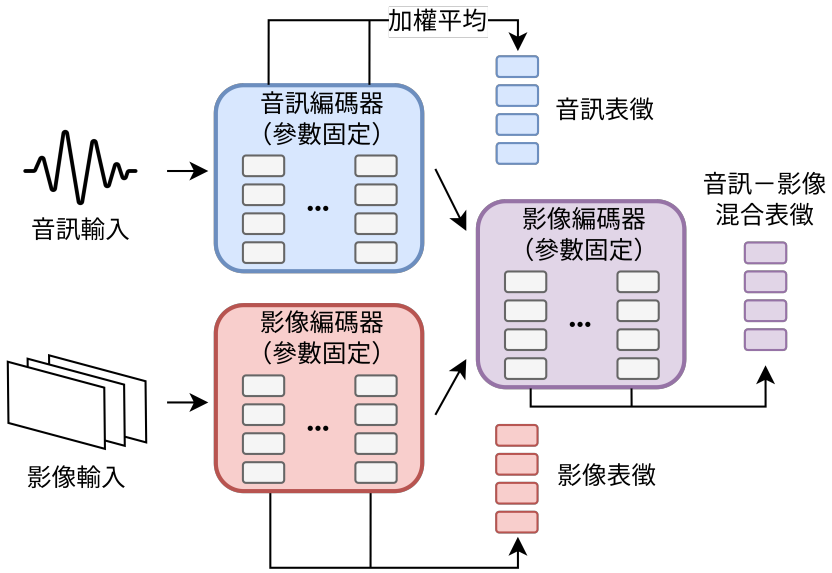


圖 4.1: 本論文所提出的評量基準考慮以下三種評量情景：僅使用音訊模態、僅使用影像模態、或同時使用兩種模態所提取的表徵。本章沿用 SUPERB 基準 [53] 的做法，將從轉換編碼器中間層提取的表徵計算加權平均，並作為對每個個別任務的下游小模型作為輸入進行訓練。第 4.2.2節詳細說明了評量基準涵蓋的所有下游任務。

如圖4.1所示，一個音訊－影像表徵模型經常由三個元件組成：兩個處理單一模態資料的編碼器，與一內化音訊影像資訊的混合編碼器。在下文中將統稱單一模態編碼器的輸出及中間層表徵為音訊表徵或影像表徵，並以混合表徵統稱混合編碼器輸出及中間層表徵。本節實驗依據這三種表徵在下游任務的表現，來衡量一個模型的泛用程度，同時與單模態自監督式表徵模型做比較。

如4.1所述，本章希望在多個下游任務上衡量音訊－影像表徵之泛用性，而不是在每個任務上取得最佳的表現，因此本章選擇將音訊－影像表徵模型的參數固

doi:10.6342/NTU202500549

定，視其為特徵抽取器，並針對每個下游任務訓練一僅有少量參數之模型，稱為下游模型。為了盡量降低下游任務的訓練過程對任務表現之影響，除了使用少量參數以外，也將大部分超參數（hyperparameter）固定，只針對學習率（learning rate）在 10^{-1} 到 10^{-5} 的範圍內進行搜索。

4.2.1 所評量之自監督式表徵模型

本章選定四個音訊－影像表徵模型進行評量，分別為 AV-HuBERT [48]、RepLAI [39]、李氏（Sangho Lee）等人提出的模型 [33]（以下稱為 AVBERT），以及 MAViL [25]。這四個模型皆為了不同任務所設計，彼此間的模型架構、預訓練目標、及前處理方式皆不盡相同。由於缺乏預訓練自監督式表徵模型所需的大量資源與時間，本章節僅針對每一篇論文所公開之模型暫存點進行評量，因此無法將每個模型所使用的預訓練資料、模型架構等因素統一。為了可以更了解預訓練階段加入影像資訊的優勢，本章也將會評量 HuBERT 這一自監督式語音表徵模型的表現，並與訓練方式相似的 AV-HuBERT 模型相比。為了驗證自監督式學習是否帶來表現的提升，我們也與信號處理中的兩種傳統表徵作比較，分別為對數梅爾濾波器表徵（log-mel filterbank, FBANK）以及方向梯度直方圖（histogram of oriented gradients, HoG）。

4.2.2 任務及資料集簡介

為了衡量音訊－影像表徵模型的泛用性，本章實驗選定兩個音訊處理任務以及三個語音處理任務來進行全方位地評測：音訊事件分類、動作辨識（action recognition）、語音辨識、語者驗證（speaker verification）、以及情緒辨識（emotion recognition）。以上每一個任務在過去皆有研究指出以多模態的方式處理可以

doi:10.6342/NTU202500549

得到更好的表現。模型架構而言，除了語音辨識屬於序列至序列（sequence-to-sequence, seq2seq）任務以外，其餘四個任務皆為語句層級（utterance-wise）的分類任務，因此語音辨識任務使用兩層的雙向長短期記憶體模型架構，而其餘四個任務皆使用兩層的感知器架構。

以下將條列出每個下游任務所用到的資料集與訓練細節。

1. **音訊事件分類**：本章使用兩個資料集來評量表徵的音訊事件分類表現，分別是 AudioSet [20] 和 VGGSound [10]。兩資料集中的影片長度皆大約為 10 秒，但兩資料集的特性略有不同，VGGSound 中的發聲源必定會出現在影片畫面中，但 AudioSet 中的發聲源則不一定會出現在畫面中。本章使用 AudioSet 類別平衡的子集進行下游模型的訓練，大約有 2 萬筆影片，而每支影片裡可能同時有多個音訊事件，一共 527 種音訊事件。本章以二分類交叉熵（binary cross entropy）作為損失函數，來優化下游模型進行多標籤分類（multi-label classification），並以在測試集上的平均精確率（mean average precision, mAP）作為評量指標。VGGSound 包含約 20 萬筆影片，按照資料集的劃分，取 18.3 萬筆作為訓練集和 1.5 萬筆作為測試集。每支影片裡標有一個音訊事件，一共 309 種音訊事件。本章以測試集的首選正確率（top-1 accuracy, Acc.）作為評量指標。

2. **動作辨識**：本章使用兩個資料集來評量表徵的動作辨識表現，分別是 Kinetics-Sounds [2] 和 UCF101 [49]。兩資料集的特性同樣略有不同，Kinetics-Sounds 影片中的聲音和畫面之間關聯性較高，但 UCF101 影片中的配音可能與動作無關。Kinetics-Sounds 是從 Kinetics 400 [29] 資料集中特別篩選出的子集，分成訓練、驗證及測試集，分別有 2.3 萬、1.6 千、和 3.1 千筆影片。該資料集中的 32 個動作類別是以比較可能同時出現在畫面與音訊

doi:10.6342/NTU202500549

中作為原則篩選出的。UCF101 則包含 1.3 萬筆影片，分為 101 個動作類別。本章使用官方提供的第一個訓練測試劃分進行評估。兩個動作辨識資料集的下游模型皆以訓練交叉熵作為損失函數，並以各自測試集的準確率作為評量指標。

3. **語音辨識**：本章使用 LRS3-TED 資料集 [1] 來評量表徵的語音辨識能力。該資料集包含從線上講座中提取的 433 小時演講影片。本章使用鏈結式時序分類（connectionist temporal classification, CTC [21]）作為損失函數，優化一個隱藏層具有 1024 維度的雙層雙向長短期記憶體模型，進行字符等級（character-level）的語音辨識。本章以測試集上的字符錯誤率（character error rate, CER）作為評量指標。在解碼過程中，不使用額外的語言模型重新排序，而是直接使用貪婪解碼（greedy decoding）。

4. **語者驗證**：本章使用 VoxCeleb2 資料集 [13] 來評量表徵的語者辨識能力。該資料集包含超過 100 萬個視頻片段。為了解省評量模型的計算成本，每位語者僅使用五部影片作為訓練資料，一部影片用於驗證資料。官方測試集用於生成目標和非目標試驗以進行測試。本章使用可加性間距軟性最大化函數（additive margin softmax [51]）作為損失函數來優化下游模型，並以測試集中的相同錯誤率（equal error rate, EER）作為評量指標。

5. **情緒辨識**：本章使用 IEMOCAP 資料集 [8] 來評量表徵的情緒辨識能力。根據常用的實驗設定，本章將快樂與興奮兩類別合併，難過與煩躁兩類別合併，以進行四類別的分類（快樂、悲傷、憤怒、剩餘類別為中性）。本章以會議一作為測試集，並以準確率作為評量指標。

doi:10.6342/NTU202500549

4.3 實驗結果與討論

4.3.1 五表徵模型之評量結果

五種模型的三種表徵之音訊處理與語音處理表現分別列於表4.1及4.2。過去文獻指出，模型最終層的輸出表徵往往不是資訊最豐富的 [43, 53]，因此我們將所有轉換編碼器的輸出與中間層表徵計算一可優化的加權平均，作為每個任務的下游模型之輸入。其中加權平均之權重是在訓練下游模型的過程中，與下游模型參數一同被優化。在計算單模態的音訊表徵與影像表徵時，只會使用對應的單一模態之輸入與編碼器計算加權平均表徵。計算音訊－影像混合表徵時，便同時以兩種模態作為輸入，計算混合編碼器之加權平均表徵。不過為了能夠更公平地比較 HuBERT 及 AV-HuBERT，在計算 AV-HuBERT 的音訊表徵和影像表徵時（圖4.1、4.2中 * 符號位置），會將另一模態編碼器的輸出設定為零，並以混合編碼器之表徵作為單一模態表徵，使 HuBERT 及 AV-HuBERT 的設定更為接近。

從表4.1可以觀察到，由於音訊事件分類中音訊資訊的重要性較高，大部分模型的音訊表徵在音訊事件分類比影像表徵更佳。而動作辨識中影像資訊的重要性較高，所以大部分模型的影像表徵表現也比較好。如第4.2.2節所述，VGGSound 和 Kinetics-Sounds 這兩個資料集中音訊資訊和影像資訊的關聯性較高。比較兩資料集上混合表徵與單一模態表徵可以發現，當資料中音訊與影像的關聯性高時，設計同時考慮雙模態資訊的模型可以達到更好的表現。相對的，在 AudioSet 與 UCF101 這類影片可能出現無關的音訊或是影像的資料上，混合表徵便不一定比單一模態表徵更好。

另外，從表徵模型在音訊處理任務上的綜合表現來說，表現最好的是 MAViL 模型，次佳的則是 AVBERT 模型。這兩個模型的共通點在於，兩者所使用的預訓

doi:10.6342/NTU202500549

表徵學習方法	參數量	音訊事件分類		動作辨識	
		AudioSet-20K (mAP ↑)	VGGSound (Acc. ↑)	Kinetics-Sounds (Acc. ↑)	UCF101 (Acc. ↑)
音訊表徵					
FBANK	0	2.8	7.76	24.73	19.91
HuBERT	95M	14.3	30.21	51.46	36.06
AV-HuBERT*	90M	12.6	31.14	49.02	38.58
RepLAI	5M	12.3	27.01	45.90	33.85
AVBERT	10M	20.5	37.67	55.28	43.26
MAViL	86M	**21.6**	**39.91**	**57.28**	**45.68**
影像表徵					
HoG	0	1.5	3.81	18.70	25.67
AV-HuBERT*	103M	2.4	5.90	24.73	37.55
RepLAI	15M	5.5	13.5	46.68	56.69
AVBERT	37M	11.5	28.73	62.67	77.42
MAViL	87M	**18.0**	**32.08**	**74.01**	**79.37**
混合表徵					
AV-HuBERT	103M	13.3	32.69	52.23	41.46
AVBERT	43M	22.9	44.54	71.31	71.76
MAViL	187M	**26.7**	**47.22**	**79.51**	**77.98**

表 4.1: 五種自監督式表徵模型之音訊事件分類與動作辨識評量結果，其中每行上方的箭頭方向朝上表示分數越高越好，朝下表示分數越低越好。每個任務中三種不同表徵的最佳結果以粗體字顯示，次佳結果則以底線顯示。參數量行中，M 代表一百萬。可以觀察到 MAViL 在音訊處理任務上表現較佳。圖中 * 號位置代表 AV-HuBERT 音訊表徵與影像表徵抽取方式與其他模型之表徵略有不同。

練資料皆是 AudioSet。AudioSet 的資料量大，資料領域（domain）也比較貼合音訊處理任務，因此表現會比使用其他預訓練資料的表徵要來得高。

若是語音處理任務而言，表4.2可以觀察到 HuBERT 與 AV-HuBERT 兩模型的表現較佳。如第2.3.2節所述，HuBERT 與 AV-HuBERT 兩模型使用的模型架構類似，預訓練目標函數也相近。因此比較 HuBERT 與 AV-HuBERT 的音訊表徵可以觀察到：預訓練階段中成對影像的輔助是否會提升音訊表徵的效果。比較後可以發現，AV-HuBERT 的音訊表徵表現與 HuBERT 相差不遠，只有在 VoxCeleb2、VGGSound、及 UCF101 上略優於 HuBERT，因此推斷成對影像並沒有輔助音訊表徵的學習。

doi:10.6342/NTU202500549

表徵學習方法	參數量	語音辨識 LRS3-TED (CER ↓)	語者驗證 VoxCeleb2 (EER ↓)	情緒辨識 IEMOCAP (Acc. ↑)
音訊表徵				
FBANK	0	21.43	27.16	51.52
HuBERT	95M	**2.96**	15.58	**62.14**
AV-HuBERT*	90M	3.01	**14.45**	58.54
RepLAI	5M	66.09	32.58	57.53
AVBERT	10M	80.23	23.74	60.94
MAViL	86M	24.43	20.71	59.46
影像表徵				
HoG	0	71.46	36.32	35.83
AV-HuBERT*	103M	**50.91**	**11.90**	26.59
RepLAI	15M	71.33	36.95	40.72
AVBERT	37M	72.29	20.00	**45.8**
MAViL	87M	74.03	24.58	43.03
混合表徵				
AV-HuBERT	103M	**2.75**	**9.46**	46.45
AVBERT	43M	70.12	18.31	**61.87**
MAViL	187M	30.18	19.67	54.94

表 4.2: 五種自監督式表徵模型之語音處理任務評量結果，其中每行上方的箭頭方向朝上表示分數越高越好，朝下表示分數越低越好。每個任務中三種不同表徵的最佳結果以粗體字顯示，次佳結果則以底線顯示。參數量行中，M 代表一百萬。可以觀察到 HuBERT 和 AV-HuBERT 在語音處理任務上表現較好。圖中 * 號位置代表 AV-HuBERT 音訊表徵與影像表徵抽取方式與其他模型之表徵略有不同。

整體來說，五種自監督式表徵模型都無法同時在所有任務上都取得好的表現。為語音任務設計的 HuBERT 與 AV-HuBERT 模型無法執行音訊任務，而在音訊任務上表現較好的 MAViL 與 AVBERT 模型也不能有效處理語音任務。

4.3.2 逐層貢獻度分析

如第4.3.1節所述，對每個下游任務而言，下游模型的輸入都是待測表徵模型的所有轉換編碼中間層表徵之加權平均。由於加權平均的權重會與下游模型參數一同被優化，可以檢視加權平均中每一層中間層輸出所佔的權重比例，來比較對特定下游任務而言，哪一層中間層表徵貢獻度最大 [12]。不過因為每一中間層表

doi:10.6342/NTU202500549

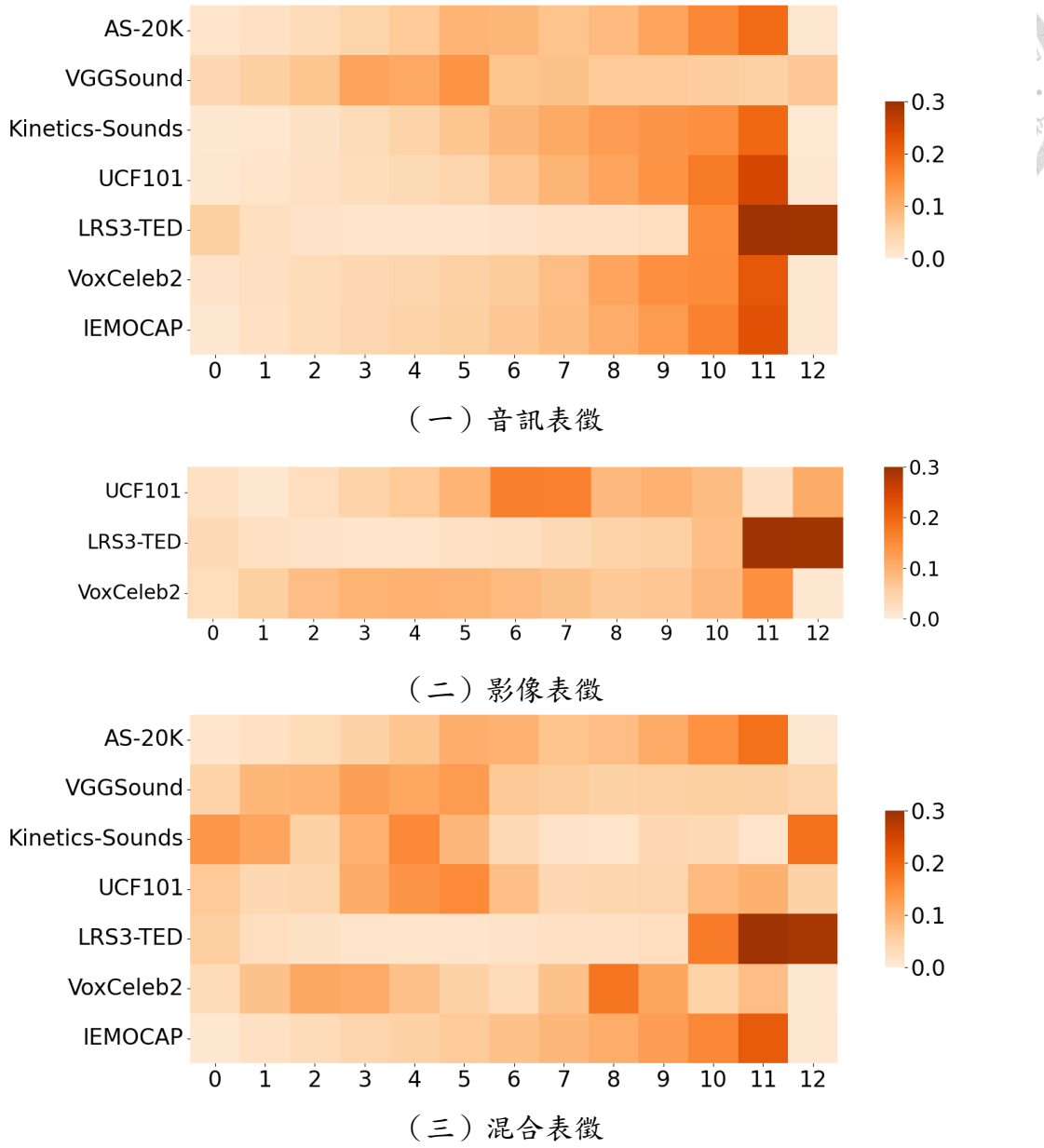


圖 4.2: AV-HuBERT 模型三種表徵的加權平均權重之熱度圖。第 0 層代表轉換編碼層的輸入。表現劣於平均之資料集結果（譬如 AV-HuBERT 影像表徵於 AudioSet-20K）已被移除。可以觀察到加權平均中貢獻最大的往往不是最終的輸出。

徵的平均範數（norm）本來就不同，分析前會先將權重與表徵的平均 L_2 範數相乘，以消除範數差異對貢獻度的影響。

對 AV-HuBERT 表徵而言，圖4.2可以看到除了在語音辨識任務外，大多數任務上最終的輸出的貢獻度通常較小。以音訊表徵而言，大多數任務中貢獻最大的是倒數第二層轉換編碼器之輸出。同時，當混合表徵優於音訊表徵之表現時（VGGSound、Kinetics-Sound、UCF101、VoxCeleb2），混合表徵中靠前的幾層之

doi:10.6342/NTU202500549

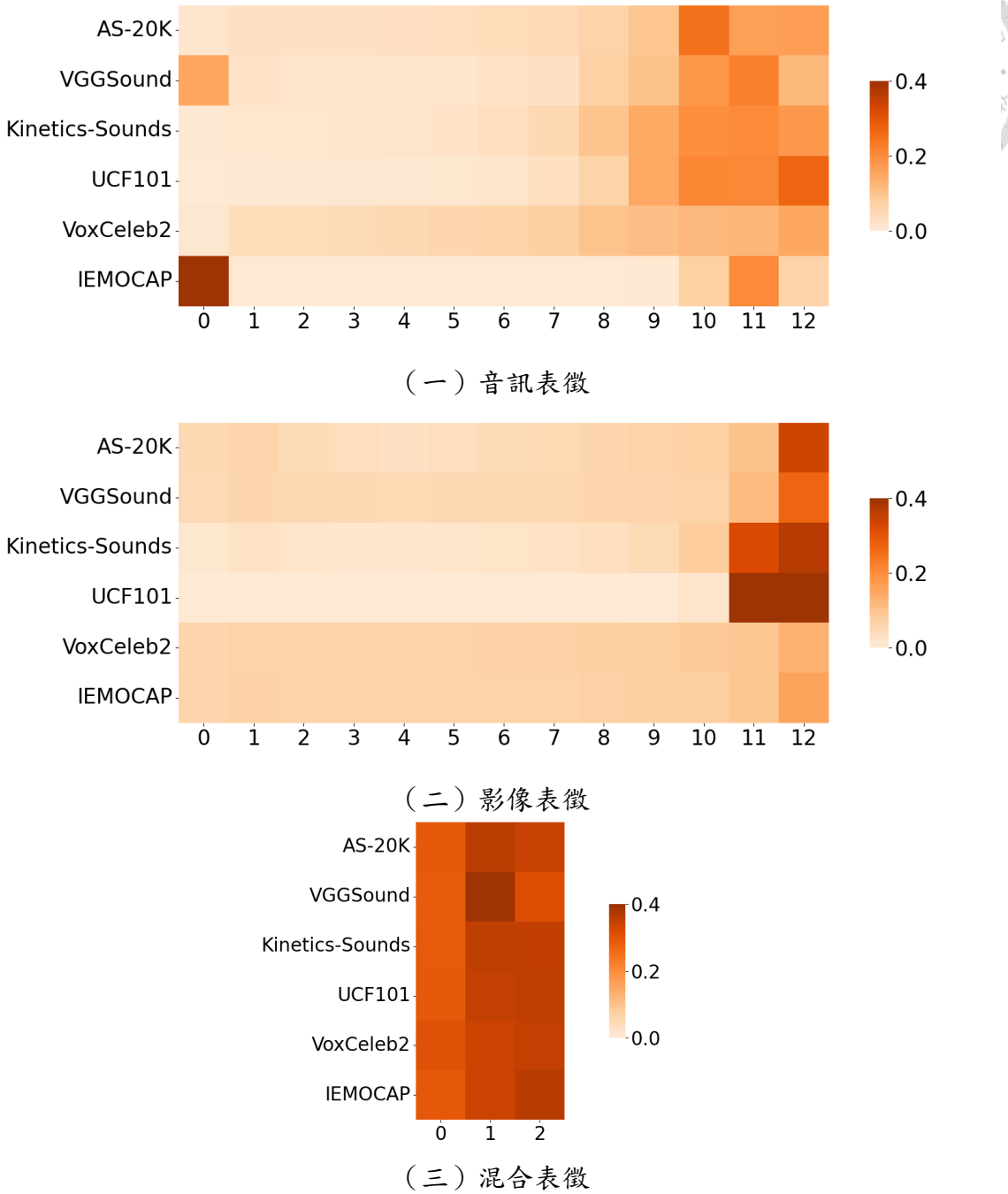


圖 4.3: MAViL 模型三種表徵的加權平均權重之熱度圖。第 0 層代表轉換編碼層的輸入。表現劣於平均之資料集結果（譬如 MAViL 影像表徵於 LRS3-TED）已被移除。可以觀察到加權平均中貢獻最大的往往不是最終的輸出。

貢獻度會比較大。這有可能代表 AV-HuBERT 中的靠前的幾層和影像資訊更為相關，而靠後的幾層則包含更多的音訊資訊。

而對 MAViL 表徵來說，由圖4.3可以觀察到貢獻度較大的往往是音訊編碼器的最後三層，以及視訊編碼器和混合編碼器的最後兩層。儘管如此，音訊表徵

doi:10.6342/NTU202500549

在 IEMOCAP 的情緒辨識任務上的表現是一個例外。對於 IEMOCAP 而言，貢獻度最大的音訊表徵是模型一開始的輸入。整體而言，AV-HuBERT 與 MAViL 表徵的逐層貢獻度之高度變化性顯示，只使用模型的最終輸出作為表徵是比較不明智的。

4.3.3 監督式訓練對表徵泛用性之影響

本章截至目前為止所討論的都是透過自監督式方式訓練出的表徵模型。然而，監督式的訓練在讓自監督式表徵模型更適應於特定任務的同時，也會影響模型表徵的泛用性。故本節額外評量經過監督式訓練後的 AV-HuBERT 與 MAViL 兩模型，並將結果呈現於表4.3。表中的 AV-HuBERT 模型被訓練在 433 小時的成對影像－文字資料上，進行唇語辨識，故過程中未使用音訊資訊。而表中的 MAViL 模型則是訓練在整個 AudioSet 資料集的 5800 小時音訊上，進行音訊事件分類。

以 AV-HuBERT 而言，從結果可以觀察到學習唇語辨識使模型影像表徵的泛用性有微幅提升，但其他表徵的泛用性大幅降低。這可能代表模型在優化過程中，逐漸喪失了處理音訊資訊的能力。而以 MAViL 模型而言，監督式訓練使模型在兩個音訊處理任務上表現有很大的提升。同時，模型在語者驗證的表現也有微幅提升，但語音辨識表現則持平，情緒辨識表現則大打折扣。這顯示了監督式學習的確可能提升模型表徵在其他任務的表現，但同樣也可能會別的任務上有很大的負面作用。因此，未來仍然需要更多研究探討如何以自監督式、監督式、或是其他的訓練方法來提升表徵的泛用性。

doi:10.6342/NTU202500549

	訓練任務	音訊事件分類		動作辨識	
		AudioSet-20K (mAP ↑)	VGGSound (Acc. ↑)	Kinetics-Sounds (Acc. ↑)	UCF101 (Acc. ↑)
AV-HuBERT					
Audio	唇語辨識	12.6(-0.6)	22.83(-8.31)	38.19(-10.83)	28.70(-9.88)
Video		2.5(+0.1)	6.12(+0.22)	25.35(+0.62)	42.03(+4.48)
Fusion		5.1(-8.2)	17.11(-15.58)	38.52(-13.71)	40.74(-0.72)
MAViL					
Audio	音訊事件分類	28.3(+6.7)	44.79(+4.89)	62.93(+5.65)	50.10(+4.42)
Video		20.9(+2.9)	36.68(+4.58)	77.39(+3.38)	86.93(+7.56)
Fusion		39.1(+12.4)	55.94(+8.72)	84.93(+5.42)	88.07(+10.09)

	訓練任務	語音辨識	語者驗證	情緒辨識
		LRS3-TED (CER ↓)	VoxCeleb2 (EER ↓)	IEMOCAP (Acc. ↑)
AV-HuBERT				
Audio	唇語辨識	13.89(-10.88)	22.38(-7.93)	53.92(-4.62)
Video		35.48(+15.43)	11.40(+0.50)	32.69(+6.10)
Fusion		22.66(-19.91)	11.35(-1.89)	43.58(-2.87)
MAViL				
Audio	音訊事件分類	23.99(+0.44)	21.77(-1.06)	58.17(-1.29)
Video		78.59(-4.56)	23.93(+0.65)	39.15(-3.88)
Fusion		30.65(-0.47)	18.61(+1.06)	46.35(-8.59)

表 4.3: 經監督式訓練後的 AV-HuBERT 與 MAViL 模型表徵結果，其中唇語辨識的訓練時未使用到音訊資訊。訓練後與原先表徵相比之絕對進步幅度以括號表示在每個結果後，紅色表示退步，藍色表示進步。

4.3.4 本章結論

本章提出一套評量基準，以表徵模型在五個下游任務的表現，來衡量自監督式音訊－影像表徵在不同類型的任務上之泛用性。實驗結果顯示，本章所評估的五種表徵模型中，未有任何一種模型可以在所有下游任務都取得好的表現。另外，音訊－影像表徵模型與單一模態的表徵模型類似，不同轉換器層的輸出表徵所包含之資訊也有所不同。譬如本章推斷 AV-HuBERT 靠前／靠後的轉換器層之輸出表徵分別與影像／音訊資訊較為相關。最後，本章發現在經過監督式訓練後，表徵雖然會更適應特定的任務，但同時在其他任務上的表現可能會大大降低。

doi:10.6342/NTU202500549

第五章　結論與展望

5.1　研究總結

本論文包含兩個不同面向的實驗，嘗試在更接近人類感知的低資源條件下建立有用的計算模型。

第3章以語音上的非監督式成分句法剖析為目標，試著做出有類似於幼兒語言習得能力的模型，可以從沒有標註的語音資料學習語法資訊。雖然本論文中提出的串接式、直接式兩架構表現上無法超越文字上的非監督式成分句法剖析，但實驗結果顯示模型仍有學到簡單語法資訊的跡象，展現出在不將語音轉寫成文字的條件下，也能夠直接對語音輸入進行非監督式成分句法剖析。另外，本論文所提出的評量指標也讓後續研究可以衡量非書面語言的剖析樹之正確率。

第4章則聚焦在自監督式音訊－影像表徵模型的評比上，試著從五個不同的音訊處理及語音處理任務的角度來衡量現有表徵模型的能力。實驗結果顯示現有的表徵模型往往僅適用於某一部分的任務，缺乏像人類一般以一套認知系統處理各種模態任務的能力。同時，本論文也發現監督式學習雖然可能讓表徵模型在特定任務上有所提升，但同樣也有可能大幅降低表徵模型於其他任務的適用性。

doi:10.6342/NTU202500549

5.2 未來展望

關於語音上的非監督式成分句法剖析，本論文所提出的兩套框架仍然有其局限性。譬如說，不論是串接式，還是直接式的系統，都是由無法同時優化的數個部分所組成。未來希望能對改進直接式系統，建立一端對端（end-to-end）的非監督式成分句法剖析語音模型，同時優化斷詞模型與剖析器。另外，也希望本論文所提出的語音非監督式成分句法剖析任務可以對低資源語音處理的其他任務有所助益，像是語音合成（speech synthesis），語音問題回答（spoken question answering），以及語音內容檢索（spoken content retrieval）。

而對音訊－影像表徵而言，本論文對現有表徵模型的系統性比較之公平性可以再提升，因為現有模型在預訓練階段所使用的預訓練資料、模型架構、損失函數等變因都不盡相同。若有足夠資源，希望可以研究在資料統一的情況下，比較什麼樣的模型架構和損失函數可以在不同任務上取得最好的表現。另外，本論文主要專注在評量現有表徵模型之多任務泛用性，對如何提升模型的泛用性則著墨較少，僅分析某兩個任務下的監督式訓練對表徵的影響。如何提升一音訊－影像表徵模型在不同任務上的泛用性尚有很大的研究空間，期許以後可以嘗試不同的自監督式、監督式、或甚至是強化學習的訓練方式，讓模型有更好的表現。

doi:10.6342/NTU202500549

參考文獻

[1] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496, 2018.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In ICCV, 2017.

[3] A. Baevski, M. Auli, and A. Mohamed. Effectiveness of self-supervised pre-training for speech recognition. arXiv preprint arXiv:1911.03912, 2019.

[4] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli. Unsupervised speech recognition. Advances in Neural Information Processing Systems, 2021.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In NeurIPS, 2020.

[6] S. Bhati, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak. Segmental Contrastive Predictive Coding for Unsupervised Word Segmentation. In Interspeech, 2021.

[7] E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. A procedure for quantitatively comparing the syntactic cov-

doi:10.6342/NTU202500549

erage of English grammars. In Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991, 1991.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008.

[9] S. Cao, N. Kitaev, and D. Klein. Unsupervised parsing via constituency tests. In EMNLP, 2020.

[10] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In ICASSP, 2020.

[11] K.-Y. Chen, C.-P. Tsai, D.-R. Liu, H.-Y. Lee, and L. shan Lee. Completely Unsupervised Phoneme Recognition by a Generative Adversarial Network Harmonized with Iteratively Refined Hidden Markov Models. In Interspeech, 2019.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 2022.

[13] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In Interspeech, 2018.

[14] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In Interspeech, 2021.

doi:10.6342/NTU202500549

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.

[16] A. Drozdov, S. Rongali, Y.-P. Chen, T. O' Gorman, M. Iyyer, and A. McCallum. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In NAACL-HLT, 2019.

[17] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. De Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux. The zero resource speech challenge 2021: Spoken language modelling. arXiv preprint arXiv:2104.14700, 2021.

[18] T. S. Fuchs, Y. Hoshen, and J. Keshet. Unsupervised Word Segmentation using K Nearest Neighbors. In Interspeech, 2022.

[19] Z. Galil. Efficient algorithms for finding maximum matching in graphs. ACM Computing Surveys (CSUR), 1986.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In ICASSP, 2017.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In ICML, 2006.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In CVPR, 2022.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed.

doi:10.6342/NTU202500549

Hubert: Self-supervised speech representation learning by masked prediction of hidden units. TASLP, 2021.

[24] W.-N. Hsu, D. Harwath, C. Song, and J. Glass. Text-free image-to-speech synthesis using learned segmental units. In ACL-ICJNLP, 2021.

[25] P.-Y. Huang, V. Sharma, H. Xu, C. Ryali, h. fan, Y. Li, S.-W. Li, G. Ghosh, J. Malik, and C. Feichtenhofer. Mavil: Masked audio-video learners. In NeurIPS, 2023.

[26] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. In NeurIPS, 2022.

[27] H. Kamper. Word segmentation on discovered phone units with dynamic programming and self-supervised scoring. arXiv preprint arXiv:2202.11929, 2022.

[28] T. Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. Coordinated Science Laboratory Report, 1966.

[29] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[30] Y. Kim, C. Dyer, and A. M. Rush. Compound probabilistic context-free grammars for grammar induction. In ACL, 2019.

[31] Y. Kim, A. M. Rush, L. Yu, A. Kuncoro, C. Dyer, and G. Melis. Unsupervised recurrent neural network grammars. In NAACL-HLT, 2019.

[32] N. Kitaev and D. Klein. Constituency parsing with a self-attentive encoder. In ACL, 2018.

doi:10.6342/NTU202500549

[33] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song. Parameter efficient multimodal transformers for video representation learning. In ICLR, 2021.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[35] D.-R. Liu, K.-Y. Chen, H.-Y. Lee, and L.-S. Lee. Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings. In Interspeech, 2018.

[36] P. Ma, R. Mira, S. Petridis, B. W. Schuller, and M. Pantic. LiRA: Learning Visual Speech Representations from Audio Through Self-Supervision. In Interspeech, 2021.

[37] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Interspeech, 2017.

[38] H. McGurk and J. MacDonald. Hearing lips and seeing voices. Nature, 1976.

[39] H. Mittal, P. Morgado, U. Jain, and A. Gupta. Learning state-aware visual representations from audible interactions. In NeurIPS, 2022.

[40] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In NeurIPS SAS Workshop, 2020.

[41] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Audio-visual speech recognition using deep learning. Applied Intelligence, 2015.

doi:10.6342/NTU202500549

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In ICASSP, 2015.

[43] A. Pasad, J.-C. Chou, and K. Livescu. Layer-wise analysis of a self-supervised speech representation model. In ASRU, 2021.

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In M. Walker, H. Ji, and A. Stent, editors, NAACL, 2018.

[45] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE, 2003.

[46] G. Shen, A. Alishahi, A. Bisazza, and G. Chrupała. Wave to Syntax: Probing spoken language models for syntax. In Interspeech, 2023.

[47] Y. Shen, S. Tan, A. Sordoni, and A. Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. In ICLR, 2019.

[48] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In ICLR, 2022.

[49] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[50] B. Wan, W. Han, Z. Zheng, and T. Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In ICLR, 2021.

[51] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. IEEE Signal Processing Letters, 2018.

doi:10.6342/NTU202500549

[52] Y.-S. Wang, H.-Y. Lee, and Y.-N. Chen. Tree transformer: Integrating tree structures into self-attention. In EMNLP-IJCNLP, 2019.

[53] S. wen 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, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In Interspeech, 2021.

[54] S. Yang, Y. Zhao, and K. Tu. Neural bi-lexicalized PCFG induction. In ACL-IJCNLP, 2021.

[55] D. H. Younger. Recognition and parsing of context-free languages in time n3. Information and Control, 1967.

[56] Y. Zhao and I. Titov. Visually grounded compound pcfgs. In EMNLP, 2020.

[57] H. Zhu, Y. Bisk, and G. Neubig. The return of lexical dependencies: Neural lexicalized PCFGs. TACL, 2020.

doi:10.6342/NTU202500549