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

用於自監督式語音模型之各項任務泛用序列壓縮法

Once-for-all Sequence Compression for Self-supervised Speech Models

陳宣叡

Hsuan-Jui Chen

指導教授: 李宏毅 博士

Advisor: Hung-yi Lee Ph.D.

中華民國 112 年 7 月

July, 2023

doi:10.6342/NTU202304502

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

用於自監督式語音模型之各項任務泛用序列壓縮法

Once-for-all Sequence Compression for Self-supervised Speech Models

本論文係陳宣叡（R10942105）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 112 年 7 月 18 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 18th July, 2023 have examined a Master's thesis entitled above presented by Hsuan-Jui Chen (R10942105) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李宏毅

（指導教授 Advisor）

王新民　　曹昱　　陳尚澤

系主任/所長 Director: 貝蘋

doi:10.6342/NTU202304502

摘要

自監督式語音模型（self-supervised speech models）在現今多項語音下游任務中達到了最先進的結果，同時展現了其在不同下游任務的泛用性，為了降低自監督式語音模型的運算量以在不同的裝置運算限制之下運行，多種不同的技術被應用來降低自監督式語音模型的運算成本，其中序列壓縮法利用語音模型的特性，以減少序列長度的方式降低運算量。本論文提出各項任務泛用序列壓縮法，讓單一的預訓練模型能根據下游任務需求動態的改變其序列壓縮率。

首先，本論文將所提出之各項任務序列壓縮法應用在兩種自監督式語音模型：知識蒸餾模型及對比式預訓練模型上，並將結果驗證在 SUPERB 基準中的多項語音下游任務當中。所提出之方法將前作預訓練模型所使用的單一序列壓縮率擴展到連續可用的壓縮率區間，同時將驗證的序列壓縮率進一步推進到了最大 48 倍的壓縮率。

接著，為了更進一步節省使用網格搜尋（grid search）尋找最佳結果所帶來的額外運算量，本論文實驗同時優化下游任務模型及上游預訓練模型壓縮率，比較此設定所得之結果和以網格搜尋所得之最佳結果間的差異，初步驗證了所提出之框架在不需要網格搜尋的前提下亦能找到最佳下游任務結果。

關鍵字：自監督式學習、序列壓縮法、各項任務泛用訓練

doi:10.6342/NTU202304502

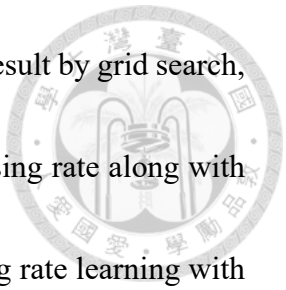
doi:10.6342/NTU202304502

Abstract

Self-supervised speech models achieve state-of-the-art results in many speech downstream tasks, showing their generalizability across different tasks. In order to operate under multiple device computational constraints, several methods have been applied to lower the computational cost of self-supervised speech models. The thesis proposed a once-for-all sequence compression method for self-supervised speech models enabling a single pre-trained model to change the sequence compressing rate on demand at inference time.

To begin with, the thesis applied the proposed once-for-all sequence compressing method on two self-supervised speech models: a knowledge distillation and a contrastive learned pre-train model, then evaluate the result on several downstream tasks from the SUPERB benchmark. The proposed method extends the original single sequence compressing rate into a continuous range of operating compressing rates, in addition, pushes the upper limit of sequence compressing to 48 times.

doi:10.6342/NTU202304502



To further reduce the computational cost of finding the optimal result by grid search, the thesis experiments with the ability to tune the upstream compressing rate along with the downstream model. Comparing the result of adaptive compressing rate learning with the overall best result obtained by grid search shows that the proposed framework has the ability to find the close to optimal result without grid search.

Keywords: self-supervised learning, sequence compression, once-for-all training

doi:10.6342/NTU202304502

目錄

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

圖目錄

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

表目錄

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

第一章　導論

1.1　研究動機

隨著機器學習在各個領域中蓬勃發展，其應用在語音處理中更是扮演了一個重要的角色，語音各項下游任務包含了：語音辨識、情緒辨識、語者辨識等任務，對於現代人類的便利生活有著功不可沒的貢獻，而其中自監督式語音模型 [1, 6, 9] 展現了在多個語音下游任務中突出的表現 [25]，其出色的任務間轉移能力讓大家對於各項任務泛用性的特質更加重視。

自監督式語音模型的優勢在於在預訓練階段中使用了大量的未標註資料，使其在應用到各項下游任務的時候僅需要微調各項下游任務專用的參數即可達到最先進的（state-of-the-art）結果。儘管自監督式語音模型的各項任務通用特質大幅的解決了在不同任務間轉換時所需要的時間以及運算量，但是自監督式模型本身隨之而來的運算量也是相當可觀的。因此多個用以壓縮自監督式語音預訓練模型的研究被提出，包含了：知識蒸餾（knowledge distillation）[4]、網路裁減（weight purning）[10]、序列壓縮（sequence compression）[13]，而其中序列壓縮的研究更在這個領域中有著突破性的發展，在特定任務如關鍵詞辨識（keyword spotting）中，可以在 8 倍序列壓縮率的情況下僅有 0.6 個百分點的表現落差相較於原始未壓縮的模型 [13]。

doi:10.6342/NTU202304502

然而在序列壓縮的研究中 [13] 發現了各項語音下游任務對於序列的壓縮率有不同的容忍程度，舉例來說，儘管關鍵詞辨識任務在 8 倍壓縮率下表現出色，使用字母單元（character unit）的自動語音辨識（automatic speech recognition, ASR）在大於 4 倍壓縮率的情況下完全無法達到可接受的結果。為了達到在各項語音下游任務中都有最適合的表現運算量權衡（performance-effiency trade-off），也就是在可接受的表現範圍內有最大的壓縮率，前作 [13] 所使用的方法需要以多次預訓練來得到多個不同序列壓縮率的預訓練模型來達成。為了解決以上限制，本論文提出了各項任務泛用序列壓縮法：一個可以根據不同的下游任務或是不同的運算量預算來即時調整序列壓縮率的預訓練方式，來更進一步推進序列壓縮自監督式語音模型的泛用性。

1.2 研究方法

本論文提出各項任務泛用序列壓縮法，利用了常見使用在自動語音辨識中的連續整合發放機制（continuous integrate-and-fire, CIF）[7] 以及可變間距序列壓縮法（variable-length sequence compression）[13]，搭配泛用性預訓練（once-for-all pre-training）[2] 的技巧，在預訓練階段隨機取樣各種不同序列壓縮率進行預訓練，來達成各項任務泛用序列壓縮的目標。

本論文之宗旨在於分析所提出的各項任務泛用序列壓縮法在各項語音下游任務中的表現，本論文將此各項任務泛用序列壓縮法應用在兩種常見的語音自監督式預訓練模型中：對比式（contrastive learning）預訓練模型以及知識蒸餾（knowledge distillation）模型，並探討此各項任務泛用序列壓縮法在不同下游任務中的表現，以及所帶來的運算量減低。

doi:10.6342/NTU202304502

1.3 主要貢獻

本論文的主要貢獻包含了以下三點：

- 提出各項任務泛用的序列壓縮法，將前作 [11, 13, 24] 少量離散的可運作序列壓縮率（discrete operational compressing rate）延伸到了連續區間的可運作序列壓縮率（continuous operational compressing rate），並將此壓縮方式驗證在兩種常見的自監督語音模型上：對比式預訓練模型以及知識蒸餾模型。
- 將此架構的模型驗證在各項不同的下游任務，並將前作 [13] 最大 8 倍壓縮率提高到了最大 48 倍壓縮率。
- 探討此一架構的額外優勢：使用自適應序列壓縮率（adaptive compressing rate learning）讓不同下游任務挑選最適當的序列壓縮率，進一步避免使用網格搜尋（grid search）所帶來的昂貴運算成本。

1.4 章節安排

本論文的章節安排如下：

- 第二章：介紹本論文相關背景知識，包含深層網路模型、語音自監督式學習與序列壓縮法。
- 第三章：介紹各項任務泛用序列壓縮法以及應用在對比式預訓練和知識蒸餾模型上的結果。
- 第四章：介紹自適應序列壓縮率之應用及結果。
- 第五章：論文總結與未來展望。

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

第二章　背景知識

2.1　深層神經網路

2.1.1　簡介

本節簡介深層類神經網路（deep neural network）。類神經網路（neural network）的基本運算單元為感知器（perceptron）[17]，其架構如圖2.1（左）所示，運算機制表達如下，

$$y = \sigma(b + \mathbf{w}^T \mathbf{x}) \tag{2.1}$$

，其中 $\mathbf{x} \in \mathbb{R}^N$ 為輸入訊號，$\mathbf{w} \in \mathbb{R}^N$ 為感知器對於輸入訊號各個維度的權重（weight），b 為純量參數偏差（bias），$\sigma(\cdot)$ 為激發函數（activation function），而 y 為對應感知器的輸出。

單層類神經網路是由多個感知器並聯所組成，而深層類神經網路則是堆疊了多個單層類神經網路，被稱為全連結層（fully connected layer），如圖2.1（右）所示，每一層（layer）由多個感知器並聯，而各層與層之間的感知器均互相連結，因此稱為全連結層。全連結層為最基本的神經網路，研究對於根據不同的應用延伸出了相對應不同的架構，例如針對視覺特化出的卷積層類神經網路

doi:10.6342/NTU202304502

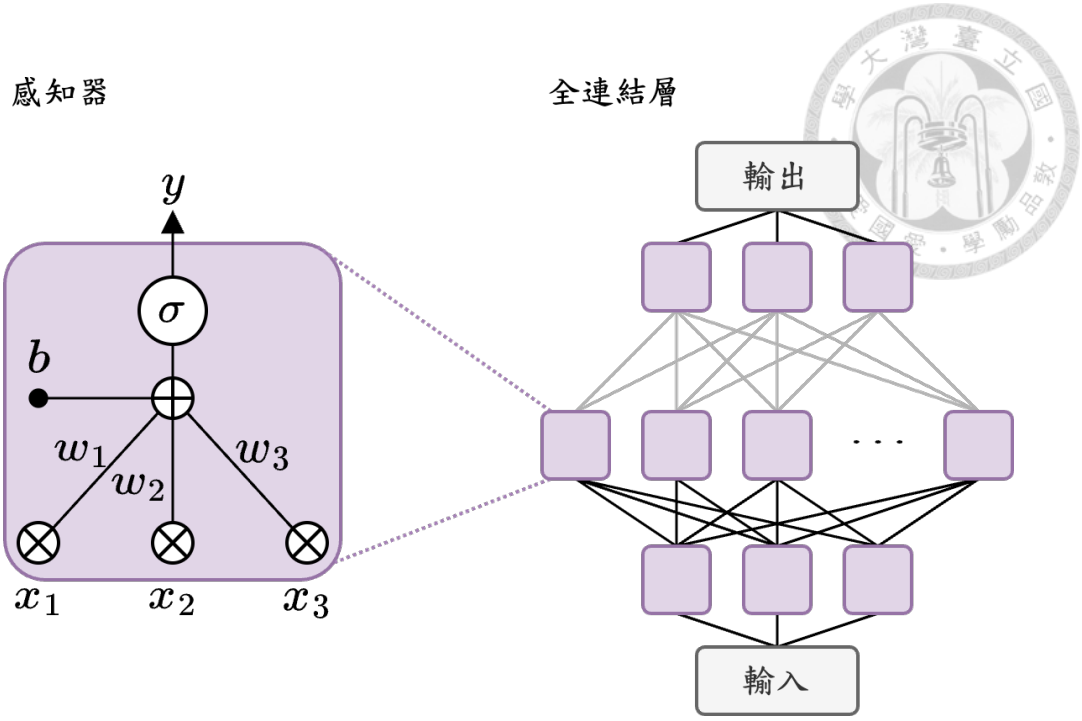


圖 2.1: 感知器與全連結層示意圖，圖左為單一感知器運作機制，圖右為全連結層結構。

（convolution neural network, CNN）及根據序列任務特化出的轉換器式類神經網路（Transformer），以下本節對於這兩種常見的特化網路進行簡介。

2.1.2 卷積式類神經網路

卷積式類神經網路（convolutional neural network, CNN）的特化是根據影像處理的濾波器（filter），卷積式類神經網路根據其濾波器的維度可以區分為一維、二維及三維卷積式神經網路，本小節對於語音模型中常用到的一維卷積式類神經網路加以介紹。

一維卷積式類神經網路中有兩個參數決定濾波器運作的方式：濾波器寬度 d（kernel width）及步幅 s（stride），其架構如圖2.2所示，其中濾波器寬度決定輸出訊號對應可觀測的輸入訊號範圍（receptive field），而步幅則決定了濾波器每次運

doi:10.6342/NTU202304502

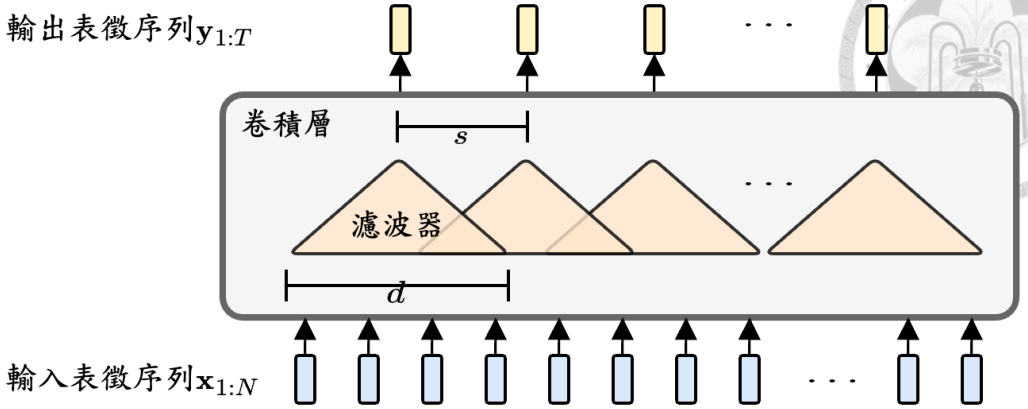


圖 2.2: 一維卷積式類神經網路示意圖，其中 d 為濾波器寬度；s 為濾波器步幅。

作之間橫移的距離，其數學表達式如下，

$$y_i = \sigma(b + \sum_{j=1}^{d} \mathbf{w}_j \mathbf{x}_{si+j}) \tag{2.2}$$

，其中 $\mathbf{x} \in \mathbb{R}^N$ 為輸入訊號，$\mathbf{w} \in \mathbb{R}^d$ 為濾波器權重，$\sigma(\cdot)$ 為激發函數。

在語音模型當中，一維卷積式類神經網路常被放在最前端，當作是抽取原始音訊（raw waveform）的模組，本論文將此類型的卷積式類神經網路稱為卷積層特徵截取模組（CNN feature extractor）。步幅大於 1 的一維卷積式網路可以視為一種次採樣層（subsample layer），而此類的卷積層特徵截取模組通常會堆疊多個步幅大於 1 的一維卷積層，並在層與層之間加入合計層（pooling layer），常見的合計層有最大合計（max pooling）及平均合計（mean pooling）。常見自監督式語音模型的卷積層特徵截取模組 [1, 4, 9]，會將取樣頻率為 16 千赫茲（KHz）的原始音訊；或等效於採樣間距 0.0625 毫秒（milliseconds, ms）的原始序列，次採樣至採樣間距為 20 毫秒的表徵序列。

doi:10.6342/NTU202304502

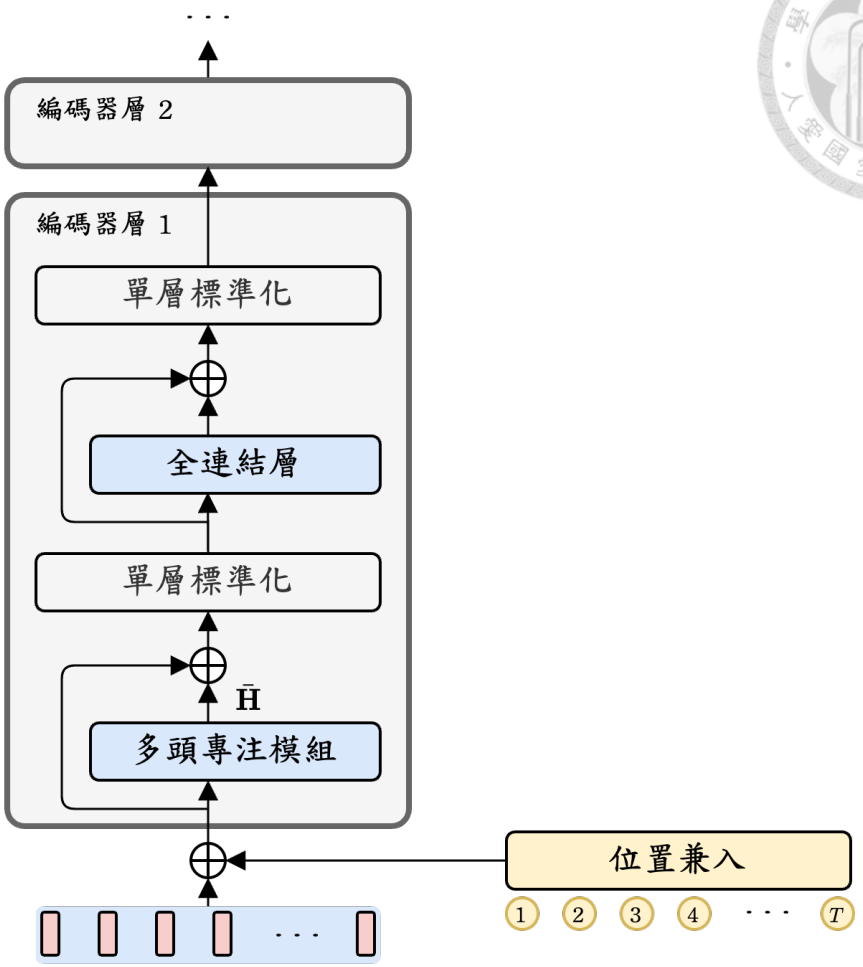


圖 2.3: 多層轉換器式類神經網路編碼器架構示意圖。

2.1.3 轉換器式類神經網路

轉換器式類神經網路（Transformer）[21] 由瓦氏（Ashish Vaswani）提出，其原始的設計包含了編碼器（encoder）以及解碼器（decoder）模組，在自監督式語音模型中多僅採納其中編碼器模組，本小節對於轉換器式類神經網路編碼器（Transformer encoder）進行簡介。

轉化器式類神經網路編碼器（或簡稱編碼器）架構圖如圖2.3所示，其中最重要的模組為多頭專注模組（multi-head attention），對於輸入特徵 $\mathbf{X} = \mathbf{x}_{1:T}$ 會依據三個不同的可學習線型轉換，轉換成為尋向量 **Q**、鑰向量 **K**、值向量 **V**，其中 d_h 為多頭專注模組的最終輸出維度，n_h 為多頭專注個數。專注機制的運作方式如

doi:10.6342/NTU202304502

下，首先循向量 **Q** 與鑰向量 **K** 進行矩陣乘法得到專注權重（attention weight），接著根據該權重對於值向量 **V** 進行加權，得到最終表徵 **H** 表示如下，

$$\mathbf{H} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}\right)\mathbf{V} \tag{2.3}$$

，其中 softmax 為軟性最大函數，此處代表的是音框層級（framewise）的軟性最大，也就是對於輸入表徵序列的每一個表徵向量分別做軟性最大所組成的序列。最終多頭專注模組將各個專注頭所得到的表徵沿著表徵維度做拼接及線型轉換後，得到最終的多頭專注模組的輸出 $\bar{\mathbf{H}}$。接著經過殘差連結、單層標準化、全連結層組成為單一編碼器層，同樣的架構重複堆疊即成為多層轉化器編碼器層（multi-layer Transformer encoders），在自監督式語音模型中多層編碼器層經常被接在卷積式特徵截取模組之後，作為學習語音自監督式損失的主要模組。

由於在多頭專注模組中並無音框在序列中的位置資訊，因此在第一層編碼器層之前，位置兼入模組會將序列位置資訊以相加的方式加入輸入表徵中，而兼入的位置表徵為參數化模型或是非參數化函數（如正餘弦）的輸出。

2.2 語音自監督式學習

2.2.1 簡介

自監督式學習是一種從未標註資料中學習的方式，勒氏（Yann LeCun）描述自監督式學習為一種「從其餘的部分預測輸入資料的任意部分」（“Predict any part of the input from any other part.”）的方式，其最大特點為不需要標註資料，加上未標註資料的取得相對於標註資料容易，現今最先進的模型經常將自監督式學習當作是監督式學習的預訓練（pre-train），接著在對應的下游任務中進行監督式學習。

doi:10.6342/NTU202304502

在語音的各項任務中，自監督式學習展現了其優異的表現及泛用能力 [1, 6, 9, 25]，而如何從未標註的資料中有效學習是自監督式模型設計中重要的一部分，也就是自監督式損失函數（self-supervised loss function）的設計，根據不同的損失函數類型可以將現今的自監督式語音模型加以分類，本節將概述常見的兩種損失函數設計：對比式預訓練模型及知識蒸餾模型。

2.2.2 對比式預訓練模型

對比式學習（contrastive learning），為一種自監督式學習方式，其中心思想為針對一個樣本找出其正樣本（positive sample）以及負樣本（negative sample），藉由損失函數拉近此樣本與正樣本的距離同時拉遠與負樣本的距離，對比式損失函數表示為，

$$L_{\text{contrastive}} = -\mathbb{E}_x \left[\log \frac{\sum_{z \in x_{\text{pos}}} \exp(\text{sim}(x, z))}{\sum_{z \in x_{\text{neg}}} \exp(\text{sim}(x, z))} \right] \tag{2.4}$$

，其中 x 為輸入表徵，x_{pos} 及 x_{neg} 分別為樣本 x 的正樣本及負樣本集合。

貝氏（Alexei Baevski）所提出的 Wav2Vec 2.0[1] 及是使用對比式學習的語音預訓練模型，其架構圖如圖2.4所示，首先對於通過卷積層特徵擷取模組（CNN feature extractor）的表徵向量進行向量量化（vector quantization）得到一組量化表徵序列，接著未經量化的表徵向量隨機遮罩特定音框之後經過編碼器層，自監督式學習的任務為針對時刻 t 輸出的表徵向量 $\mathbf{c}_t$，拉近與對應同時刻音框所對應的量化特徵 $\mathbf{q}_t$，而負樣本 $\mathbf{Q}_t$ 則從其餘被遮罩的音框對應的量化特徵中隨機取樣，損失函數表達為，

$$L_{\text{Wav2Vec 2.0}} = -\mathbb{E}_x \left[\log \frac{\exp(\text{sim}(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\mathbf{q} \in \mathbf{Q}_t} \exp(\text{sim}(\mathbf{c}_t, \mathbf{q})/\kappa)} \right] \tag{2.5}$$

doi:10.6342/NTU202304502

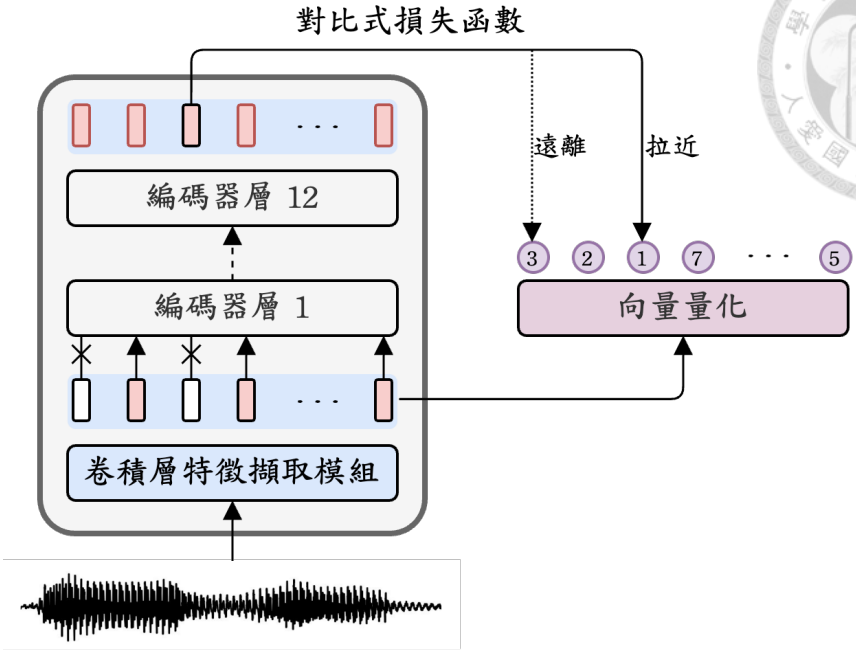


圖 2.4: Wav2Vec 2.0[1] 模型架構示意圖。

，其中相似函數 $\text{sim}(\cdot,\cdot)$ 為餘弦相似度（cosine similarity）；κ 為溫度參數（temperature）。實驗結果發現以此對比式損失函數預訓練的模型能夠學習到相當優異的語音表徵，在貝氏提出此模型時，其預訓練模型微調（fine-tune）在少量一小時的語音標註資料，贏過了以監督式訓練在一百小時語音標註資料的結果。

2.2.3 知識蒸餾模型

知識蒸餾（knowledge distillation）[8] 由辛氏（Geoffrey Hinton）所提出，做為壓縮模型（model compression）的一種方式，其中心思想為以小模型（學生模型）模擬大模型（教師模型）的輸出分佈，可以應用在監督式模型或自監督式模型上，如果教師模型使用自監督式學習，則知識蒸餾得到的學生模型亦為自監督式模型。用來拉近教師模型及學生模型的損失函數可以是任意的機率分佈距離函數，常見被使用在知識蒸餾學習的距離函數為凱式分歧度（Kullback–Leibler

doi:10.6342/NTU202304502

divergence, KL divergence），使用凱式分歧度的蒸餾損失函數表達如下，

$$L_{\text{distillation}} = D_{\text{KL}}(P(x)||Q(x)) \tag{2.6}$$

$$= \mathbb{E}_{x\in\mathcal{D}}\left[\log\frac{P(x)}{Q(x)}\right] \tag{2.7}$$

，其中 $P(\cdot)$ 及 $Q(\cdot)$ 分別為教師及學生模型的輸出分佈函數，x 為從知識蒸餾訓練集 $\mathcal{D}$ 中取樣的樣本。

在自監督式語音模型中，不同的輸入音訊長度會對應到不同的輸出表徵序列長度，常見的作法是做音框層級的知識蒸餾（framewise distillation），在模型架構設計上須確保教師模型以及學生模型對同一段音訊輸出的表徵序列長度相同。另外研究發現自監督式語音模型各個不同層對應的表徵向量包含了不同的資訊[5, 16]，因此張氏（Heng-Jui Chang）提出的模型 DistilHuBERT[4] 使用了多個預測模組（projection head）來學習教師模型（選用的教師模型為 HuBERT[9]）之選定層數的輸出表徵向量序列，其架構圖如圖2.5所示。

張氏所選用於知識蒸餾的距離函數為最小化曼哈頓距離（或稱 $\ell 1$ 距離, $L_{\ell 1}$）及最大化餘弦相似度（cosine similarity, $L_{\cos}$），要拉近的兩個向量 $\mathbf{h}_t^{(l)}, \hat{\mathbf{h}}_t^{(l)} \in \mathbb{R}^D$ 分別代表在第 t 個音框由教師模型第 l 層及學生模型對應層數的預測模組的表徵向量，損失函數為，

$$L_{\text{DistilHuBERT}}^{(l)} = L_{\ell 1}^{(l)} + \gamma L_{\cos}^{(l)} \tag{2.8}$$

$$= \sum_{t=1}^{T}\left[\frac{1}{D}\left\|\mathbf{h}_t^{(l)} - \hat{\mathbf{h}}_t^{(l)}\right\|_1 - \gamma\log\sigma\left(\cos\left(\mathbf{h}_t^{(l)}, \hat{\mathbf{h}}_t^{(l)}\right)\right)\right] \tag{2.9}$$

，其中 T 為總音框個數、σ 及 $\cos(\cdot,\cdot)$ 分別代表 S 函數（Sigmoid function）及餘弦相似度、γ 為權重。張氏的研究發現，以此知識蒸餾損失學習得到的學生模型，相對於教師模型減少了 75 個百分點的參數量及 73 個百分點的運算時間，同時在

doi:10.6342/NTU202304502

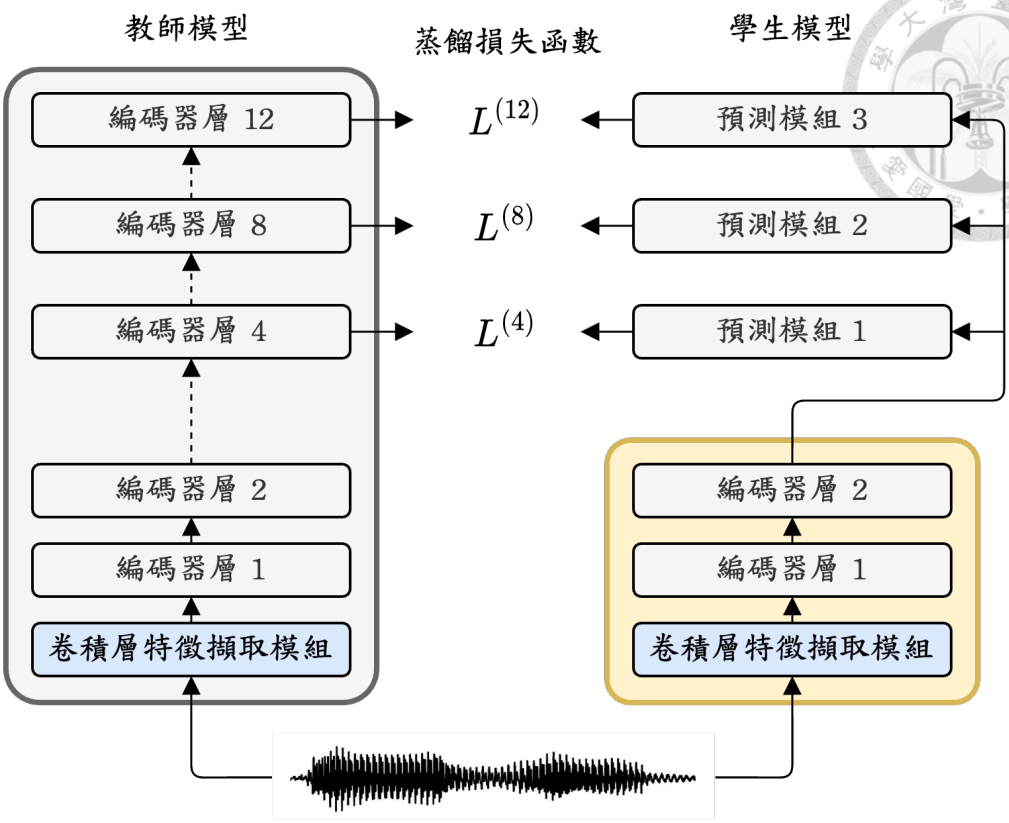


圖 2.5: DistilHuBERT[4] 模型架構示意圖。

多項下游任務中保有接近於教師模型的表現。

2.2.4 語音下游任務

為了評斷自監督式語音模型的表現，由楊氏（Shu-wen Yang）提出了一個統一的基準（benchmark）：SUPERB 基準（Speech processing Universal PERformance Benchmark）[25]，作為一個公平的平台，比較不同的自監督式語音模型，此基準包含了多項語音下游任務，並規範了各個下游任務中所使用的資料集、訓練參數與驗證方式等，以確保不同模型之間的公平性。蔡氏（Hsiang-Sheng Tsai）對 SUPERB 基準提出了延伸任務，專注於語義和生成能力，稱之為 SUPERB-SG[19]，包含兩者基準的最新自監督式語音模型的比較結果可以在 SUPERB 基準的公開網站[1]中取得。

[1] https://superbbenchmark.org

doi:10.6342/NTU202304502

表 2.1: 本論文所使用的語音下游任務與所屬任務類別配對。

任務類別	任務內容
內容相關 （Content）	音素辨識（Phoneme Recognition） 語音辨識（ASR） 關鍵詞辨識（Keyword Spotting） 按例查詢（Query by Example）
語者相關 （Speaker）	語者辨識（Speaker Identification）
語義相關 （Semantics）	意圖辨識（Intent Classification） 填空任務（Slot Filling）
副語言相關 （Paralinguistics）	情感辨識（Emotion Recognition）
語義和生成相關 （Semantic & Generative）	語音翻譯（Speech Translation）

本論文驗證在 SUPERB 基準多個下游任務的子集合中，包含了音素辨識（Phoneme Recognition）、語音辨識（Automatic Speech Recognition, ASR）、關鍵詞辨識（Keyword Spotting）、按例查詢（Query by Example）、語者辨識（Speaker Identification）、意圖辨識（Intent Classification）、填空任務（Slot Filling）、情感辨識（Emotion Recognition）、語音翻譯（Speech Translation），根據任務的不同面向，這九個任務可以分類為：內容相關、語者相關、語義相關、副語言相關、語義和生成相關等任務，任務內容及其所屬的任務分類如表2.1所示。

根據下游模型的模型架構，本論文所使用的下游任務可以進一步分類為以下四種類別：鏈結式時序分類（connectionist temporal classification, CTC）、序列至序列（sequence-to-sequence）、序列層級合計（sequence-level pooling）、序列層級比對（sequence-level comparison），其中各個任務對應到的任務類型如表2.2所示，以下本小節對於這四種不同的任務類型進行簡介。

doi:10.6342/NTU202304502

表 2.2: 本論文所使用的語音下游任務與所屬任務類型配對。

任務類型	任務內容
鏈結式時序分類（CTC）	音素辨識（Phoneme Recognition） 語音辨識（ASR） 填空任務（Slot Filling）
序列至序列（sequence-to-sequence）	語音翻譯（Speech Translation）
序列層級合計（sequence-level pooling）	關鍵詞辨識（Keyword Spotting） 語者辨識（Speaker Identification） 意圖辨識（Intent Classification） 情感辨識（Emotion Recognition）
序列層級比對（sequence-level comparison）	按例查詢（Query by Example）

2.2.4.1 序列至序列

序列至序列（sequence-to-sequence）模型可以用於解決輸入及輸出序列長度不同之任務，其下游任務模型由編碼器（encoder）及解碼器（decoder）兩個部分組成，首先編碼器會將自監督式語音模型的輸出表徵作為輸入，接著解碼器以自回歸（auto-regressive）的方式，根據編碼器的輸出及已經解碼出的字符序列（token sequence）解碼此刻字符輸出，直到任務完成。序列至序列下游模型架構如圖2.6所示，本論文使用序列至序列的下游任務有語音翻譯，其使用的評斷指標為 BLUE 分數（BLUE score）。

2.2.4.2 鏈接式時序分類

鏈接式時序分類（connectionist temporal classification, CTC）可以用以解決序列至序列問題中輸出序列長度嚴格小於輸入序列長度，且輸入及輸出序列為單調對齊（monotonic align）的任務，在鏈接式時序分類器的輸出序列中可以選擇輸出字符（token）或是空字符（null token），接著塌縮函數會進行兩個步驟的塌縮：合併重複及去除空字符，由於塌縮函數的使用，其輸出的序列長度勢必會小於或等

doi:10.6342/NTU202304502

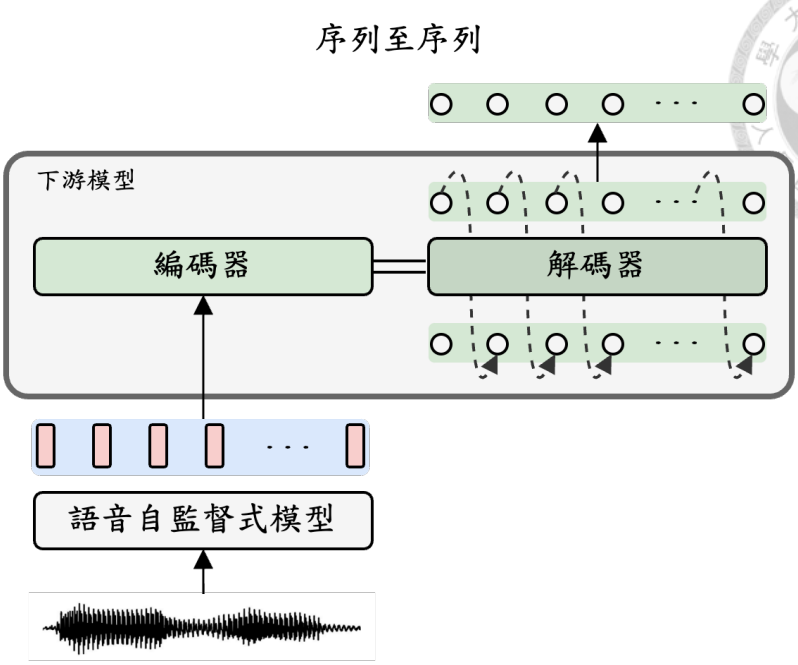


圖 2.6: 序列至序列下游模型架構示意圖。

於輸入序列長度，因此使用鏈接式時序分類器的下游任務有一個嚴格的物理限制：由自監督式語音模型輸出的表徵序列長度要大於或等於對應下游任務的目標序列長度。鏈接式時序分類下游模型架構如圖2.7所示，本論文使用鏈接式時序分類的下游任務有音素辨識、語音辨識、填空任務，其中音素辨識使用的字符為音素單元（phone unit）而語音辨識及填空任務使用的字符為字母單元（character unit），這一類任務的評斷指標通常使用相對應字符錯誤率（token error rate），較特別的是填空任務在 SUPERB 基準中使用另外的 F1 分數（F1 score）作為評斷任務類別正確率的指標。

2.2.4.3 序列層級合計

序列層級合計（sequence-level pooling）可以用於單一序列辨識單一類別的任務，其做法會將自監督式語音模型的輸出表徵在序列層級做合計（pooling）如最大合計（max pooling）或是平均合計（mean pooling），接著針對這一個合計過後的表徵向量進行分類（classification）任務。序列層級合計下游模型架構如圖2.8所

doi:10.6342/NTU202304502

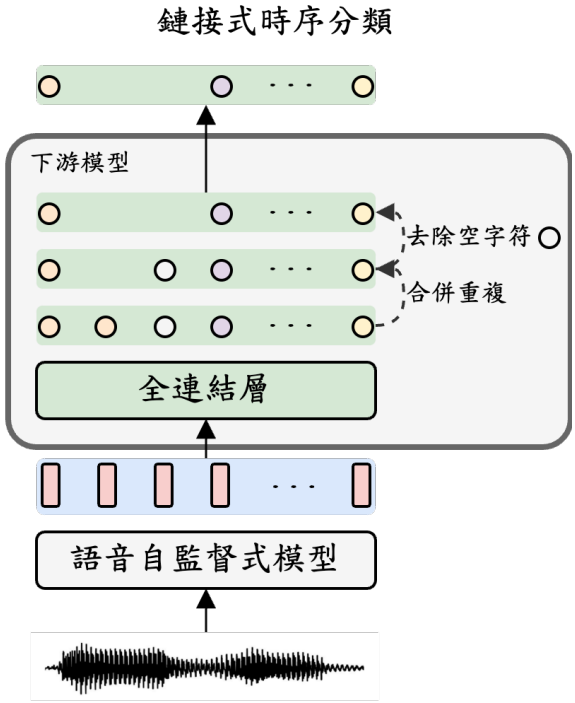


圖 2.7: 鏈接式時序分類下游模型架構示意圖。

示，本論文使用序列層級合計的下游任務有關鍵詞辨識、語者辨識、意圖辨識、情感辨識，其中關鍵詞辨識是實作在獨立詞彙上（isolated words），此類任務使用的評斷指標為分類正確率（accuracy）。

2.2.4.4 序列層級比對

序列層級比對（sequence-level comparison）用於給定一段目標語音決定受測語音和目標語音的相似度，其作法為將受測語音及目標語音經過自監督式語音模型得到對應的表徵序列，接著對這兩段表徵序列以動態時間校正算法（dynamic time warping, DTW）計算出兩序列的相似度。序列層級比對下游模型架構如圖2.9所示，本論文使用序列層級比對的下游任務有案例查詢，案例查詢所使用的受測語音為一個詞彙（spoken term），目標語音為資料庫中的語音文檔（audio document），其使用的評斷指標為 MTWV 分數（maximum term weighted value, MTWV）。

doi:10.6342/NTU202304502

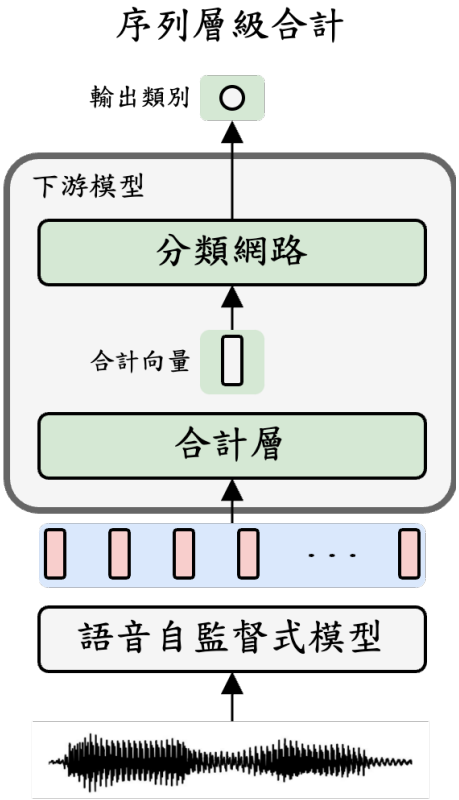


圖 2.8: 序列層級合計下游模型架構示意圖。

2.3 序列壓縮法

2.3.1 簡介

序列壓縮法可以用以統稱在時間維度上減少序列長度的作法，最簡單的丟棄（drop）和拼接（concatenating）及是常見在語音辨識系統中用以減少序列長度的方法 [3, 14, 20]，能夠有效的降低語音辨識系統的運算量。前作對於自監督式語音模型也提出對應的序列壓縮法，如李氏（Yeonghyeon Lee）[11] 及吳氏（Felix Wu）[24] 提出在自監督式語音模型中將次採樣（subsample）配對於上採樣（upsample），也就是在模型初期使用次採樣將序列縮短，在輸出階段使用上採樣將序列還原成原來的長度，以較短的序列長度通過編碼器層換取運算量降低，兩者的實驗均發現序列壓縮可以在表現差異不大的前提下大幅降低自監督式語音模型的運算量，

doi:10.6342/NTU202304502

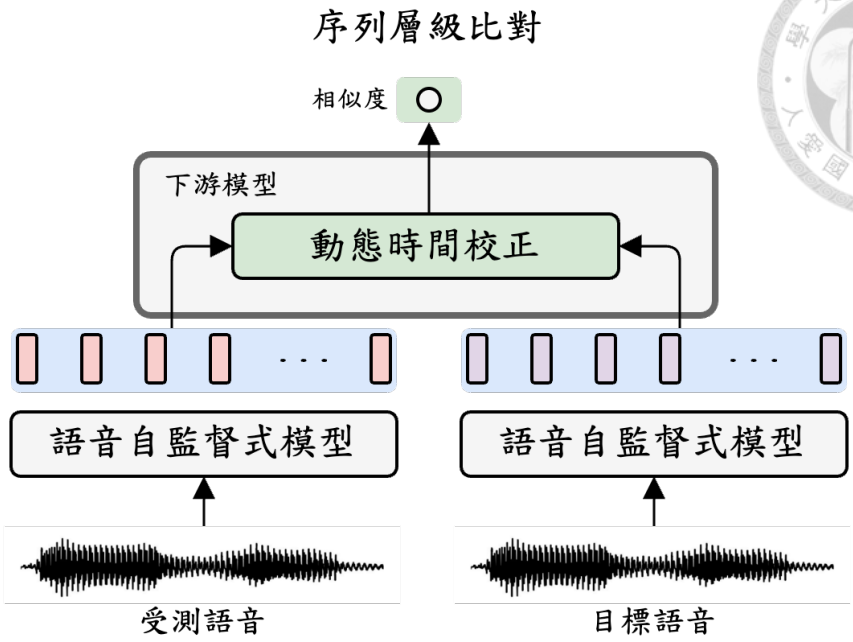


圖 2.9: 序列層級比對下游模型架構示意圖。

然而李氏及吳氏的研究僅實驗在有限的 2 倍序列壓縮率（40 毫秒採樣間距）。

由孟氏（Yen Meng）[13] 發表的研究中，進一步的將序列壓縮率提高到了最高 8 倍壓縮率（160 毫秒平均採樣間距），不同於常見的固定間距序列壓縮法（fixed-length sequence compression）：固定間距（interval）的輸入音訊對應到一個輸出表徵向量，孟氏結合了連續整合發放機制（continuous integrate-and-fire, CIF）[7] 提出了可變間距序列壓縮法（variable-length sequence compression），預訓練模型可以在預訓練階段和原本的自監督式目標（self-supervised objective）一同學習分割邊界（segmentation boundary）。其研究中發現在大壓縮率（平均採樣間距接近平均音素長度）的情況下，可變間距序列壓縮法的表現勝過於固定間距序列壓縮法，本節將簡述所使用的連續整合發放機制及可變間距序列壓縮法。

doi:10.6342/NTU202304502

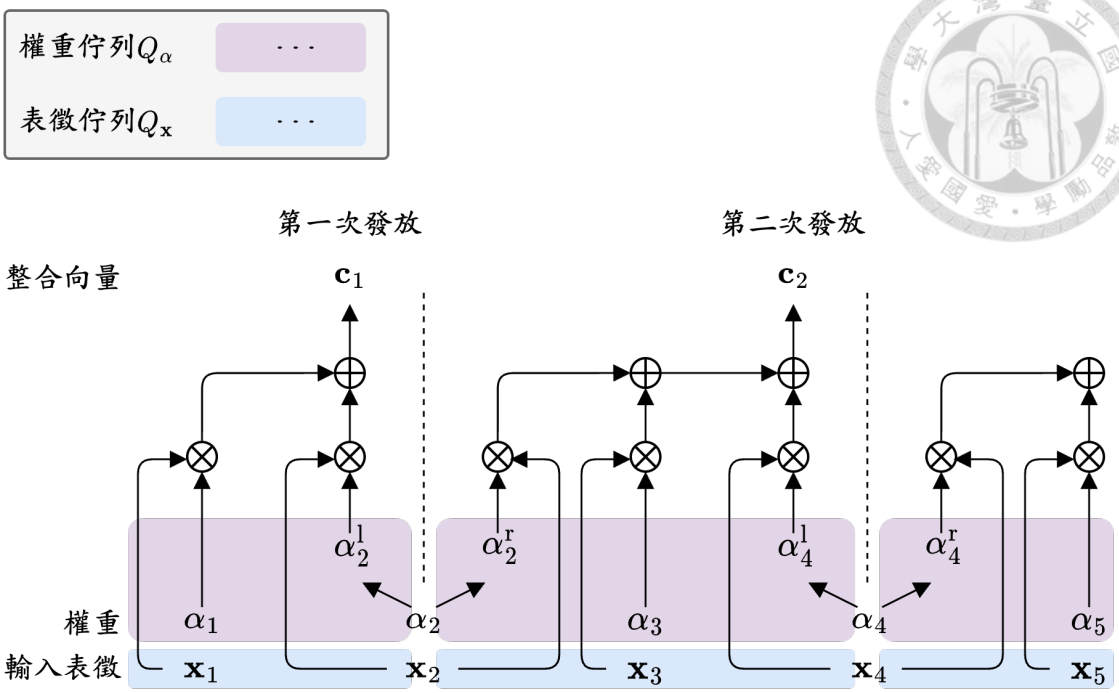


圖 2.10: 連續整合發放函數運作示意圖。

2.3.2 連續整合發放機制

連續整合發放機制（continuous integrate-and-fire, CIF）[7] 由董氏（Linhao Dong）所提出，為一種軟性和單調對齊（soft and monotonic alignment）的序列壓縮方式，其研究驗證在語音辨識任務上，和傳統使用在語音辨識的對齊機制：鏈接式時序分類（connectionist temporal classification, CTC）達到接近的結果，本小節將簡介其連續整合發放機制。

連續整合發放機制可以分成兩個主要的模組：權重預測模組（α prediction module）及連續整合發放函數（CIF function），其運作機制圖2.10所示：權重預測模組由一維卷積層類神經網路所組成，輸入長度為 T 的表徵序列 $\mathbf{x}_{1:T}$，並輸出對應權重 $\alpha_{1:T}$，其中 $0 \leq \alpha_i \leq 1;\ \forall i$。接著，連續整合發放函數同時輸入表徵序列 $\mathbf{x}_{1:T}$ 以及權重預測模組輸出權重 $\alpha_{1:T}$，在每一個時刻 t，連續整合發放函數會維持兩個佇列（Queue），分別儲存輸入表徵及其對應的權重值，根據在 t 時刻佇列的權重合及事先定好的閥值 β 分為以下兩種操作方式：

doi:10.6342/NTU202304502

1. 情境一：當佇列中累積權重合未達閥值 β 時，

$$\sum_{\alpha_k \in Q_\alpha} \alpha_k + \alpha_t < \beta \tag{2.10}$$

，將此刻對應的輸入表徵 $\mathbf{x}_t$ 及對應權重 α_t 分別存入佇列中，接著無需任何動作直接進入下一個時刻 $t+1$。

2. 情境二：當佇列中累積權重合超過閥值 β 時，

$$\sum_{\alpha_k \in Q_\alpha} \alpha_k + \alpha_t \geq \beta \tag{2.11}$$

，則需要進行發放（fire）機制，在發放時，首先將此時刻的權重 α_t 取出並分為左右兩個部分 $\alpha_t^{\mathrm{l}}, \alpha_t^{\mathrm{r}}$，其中 $\alpha_t^{\mathrm{l}} + \alpha_t^{\mathrm{r}} = \alpha_t$ 並使得佇列中的權重和 α_t^{l} 的合等於閥值，

$$\sum_{\alpha_k \in Q_\alpha} \alpha_k + \alpha_t^{\mathrm{l}} = \beta \tag{2.12}$$

，接著將總和為閥值的這些權重與對應佇列中的表徵向量取出進行加權合，加權合的結果為整合向量。接著連續整合發放函數清空佇列，並將剩餘的權重 α_t^{r} 及對應的表徵向量 $\mathbf{x}_t$ 存入對應佇列中，接著進入下一個時刻 $t+1$。

連續整合發放函數重複以上步驟直到整個序列發放完畢（時刻 T），最終整合向量所組成的序列及為輸出表徵序列。

2.3.3 可變間距序列壓縮法

可變間距序列壓縮法（variable-length sequence compression）由孟氏 [13] 所提出，其方法在自監督式語音模型中的卷積層特徵截取模組及編碼器層中間

doi:10.6342/NTU202304502

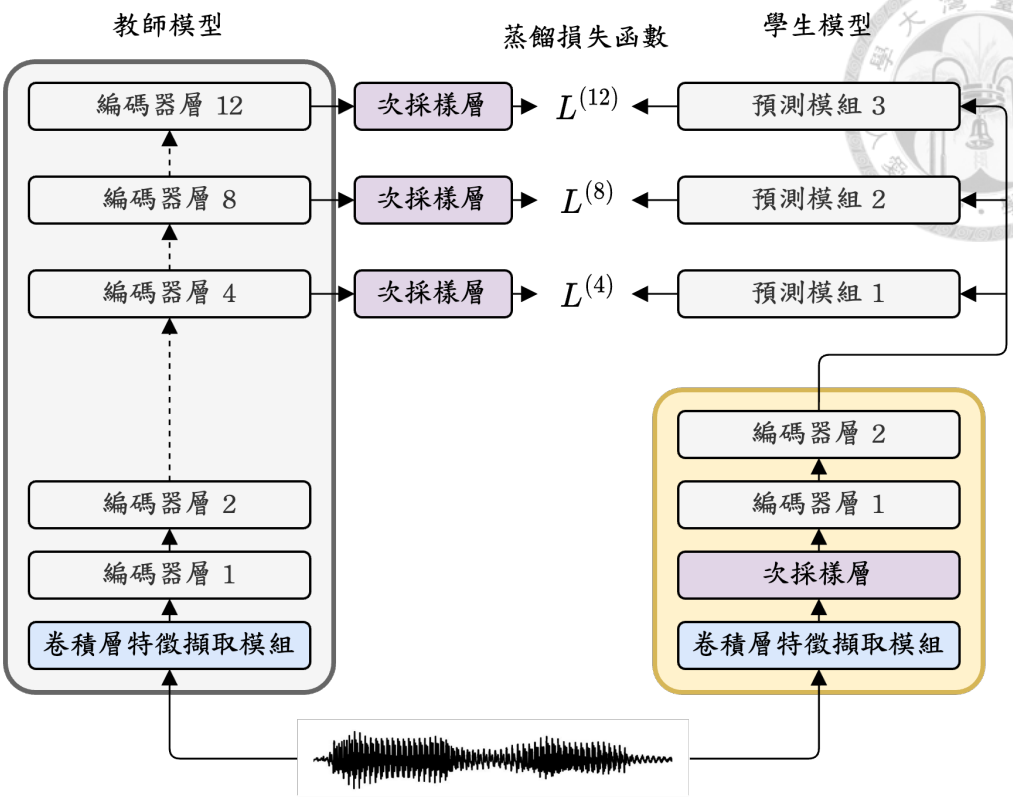


圖 2.11: 可變間距序列壓縮法 [13] 用於 DistilHuBERT 架構示意圖。

加入連續整合發放次採樣層（CIF subsample layer），其應用於知識蒸餾模型 DistilHuBERT 上的模型架構如圖2.11所示。

孟氏的研究發現上述架構中，預訓練模型在無額外引導的情況下，並無法學習出有意義的聲學邊界，在表現上相對於傳統的固定間距序列壓縮法並無可觀測的進步，因此提出額外的損失函數稱為分割學習引導（boundary guidance），用以引導權重預測模組預測出對應到有意義分割的權重。分割學習引導運作方式如下，首先針對預訓練的資料集使用非監督式方法得到一組聲學邊界，舉例來說：使用非監督式語音辨識（unsupervised ASR）[12] 將目標語音訊號做音素辨識，標註出每個音框對應到的音素標示，並將所有音素變換時刻的音框當作是聲學邊界。舉例來說，給定一段語音訊號，將所得到的非監督式聲學邊界表示為 $t_{1:K} = \{t_1, t_2, \ldots, t_K\}$，第一個分割層級損失函數（segment-based loss, L_{seg}）用以拉近連續整合發放下採樣層中權重預測模組所預測的邊界和非監督式聲學邊界的

doi:10.6342/NTU202304502

距離，

$$L_{\text{seg}}(\alpha) = \sum_{k=1}^{K} \left| \sum_{j=1}^{t_k} \alpha_j - k \right|. \tag{2.13}$$

，另外一個音框層級損失函數（frame-based loss, L_{frame}）則用以拉近音框層級（framewise）的距離，首先根據聲學邊界建構出一組權重 $\alpha_{1:T}^{\text{sup}}$，

$$\alpha_i^{\text{sup}} = \frac{1}{t_{k+1} - t_k} \text{ for } t_k \leq i < t_{k+1} \tag{2.14}$$

，接著此損失函數拉近連續整合發放次採樣層中權重預測模組所預測的權重 $\alpha_{1:T}$ 和 $\alpha_{1:T}^{\text{sup}}$ 之間的距離，

$$L_{\text{frame}}(\alpha, \alpha^{\text{sup}}) = \|\alpha - \alpha^{\text{sup}}\|_1 \tag{2.15}$$

，最終的損失函數則表示為，

$$L_{\text{total}} = L_{\text{SSL}} + \gamma_{\text{seg}} L_{\text{seg}} + \gamma_{\text{frame}} L_{\text{frame}} \tag{2.16}$$

，其中 L_{SSL} 為原自監督式模型的損失函數；γ_{seg} 及 γ_{frame} 為對應分割層級及音框層級損失函數之權重。實驗結果顯示，當壓縮比率接近平均音素發音長度時（接近 100 毫秒的採樣間距），其提出的可變間距序列壓縮預訓練模型，能夠在多數的語音下游任務中有優於同樣壓縮率的固定間距序列壓縮預訓練模型的結果。

2.3.4 時間及空間複雜度

由於編碼器中的自專注模組需要計算序列表徵兩兩之間的專注權重，對於輸入長度為 N 的序列，時間及空間複雜度為 $O(N^2)$，也就是在未經優化的情況下，運算複雜度和輸入表徵的序列長度有著二次（quadratic）相關性。有一個方向的

doi:10.6342/NTU202304502

研究著重於降低編碼器和輸入序列長度的複雜度相關性，舉例來說，Big Bird [26] 提出可以將時間複雜度降低至 $O(N)$ 的專注機制，然而在實際應用的情況下，這一類型的專注機制為了達成次二次相關（sub-quadratic）的專注機制，需要在低次項上花費額外的運算或是在空間複雜度上有所犧牲 [18]，使得這一類的專注機制雖然在複雜度上較低，但是低次方項所造成的多餘運算量會導致在序列大於某一長度的情況下才會有明顯的運算量降幅，因此多數的自監督式語音模型為了簡單起見依然是使用第2.1.3節描述的專注機制。而序列壓縮法是一個可以更直接降低運算量的方法，對於輸入表徵序列長度進行壓縮後再通過編碼器層，以直接降低輸入序列長度方式，對於不論是使用傳統二次相關或是次二次相關的專注機制，均可以有效降低自監督式語音模型所需要的運算量。

doi:10.6342/NTU202304502

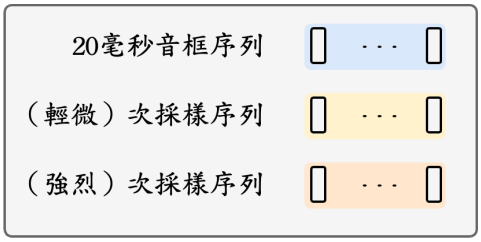
第三章　各項任務泛用序列壓縮法

3.1　各項任務泛用序列壓縮法用於自監督式語音模型

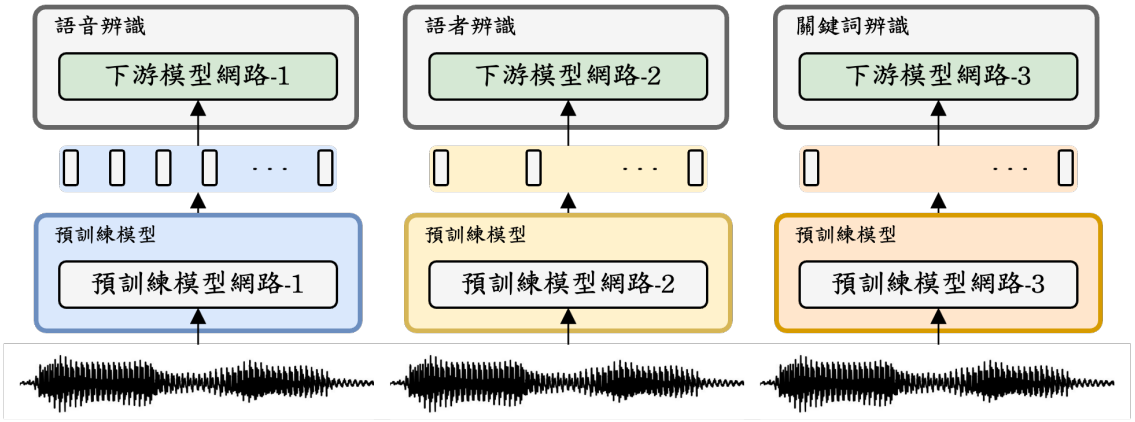
3.1.1　簡介

各項任務泛用（once-for-all）[2] 技術由蔡氏（Han Cai）所提出，其中心思想為「訓練單一網路，根據不同情境特化並有效率的部署」（“Train one network and specialize it for efficient deployment.”），在此技術之前，如果要在裝置端根據不同的運算或儲存空間資源達到最佳的表現運算量權衡，每一個不同的情況都需要進行一次訓練。由蔡氏所提出的技術，利用網路裁減配合各項任務泛用技術達到「訓練一次得到多個模型」（”Train once get many.”）的目的，在每一次部署的時候，根據不同的運算限制從單一預訓練模型中選取不同大小的子網路，作為裝置端部署的網路。王氏（Rui Wang）所提出的 LightHuBERT [23] 將各項任務泛用技術應用在自監督式語音模型當中，同樣是利用子網路的選取的方式，在不同的運算限制下選取不同大小的子網路。前作 [23] 將各項任務泛用技術應用在自監督式語音模型時僅著重在模型參數量上，在不同部署情境下，預訓練模型的輸出序列長度均保持固定，然而在語音模型當中，序列的長度是另外一個影響運算量的重要因素，本論文要探討的即是將各項任務泛用技術應用在自監督式語音模型的序列壓縮上。

doi:10.6342/NTU202304502



單一序列壓縮率模型



各項任務泛用序列壓縮率模型

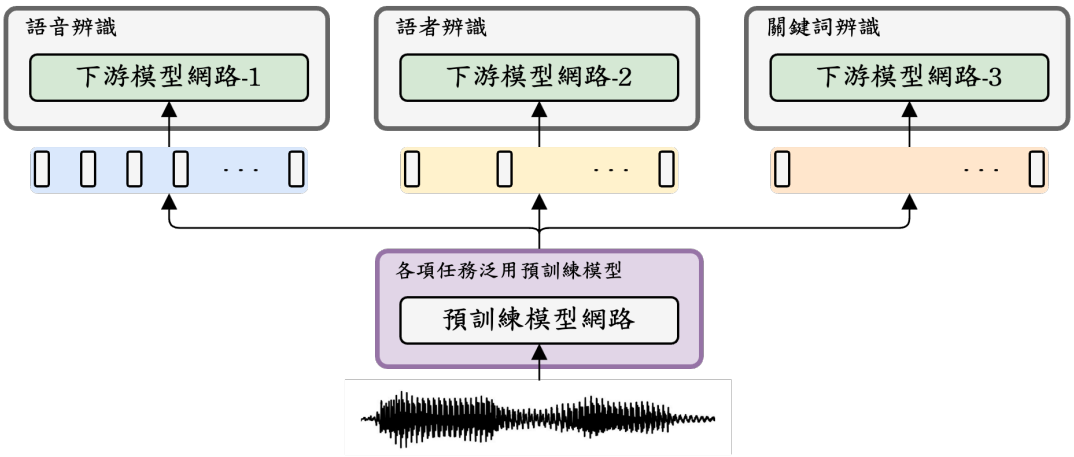


圖 3.1: 單一序列壓縮率及各項任務泛用序列壓縮率預訓練模型對比圖。上半圖為單一序列壓縮率模型，若欲對不同的下游任務選取最恰當的序列壓縮率則需多個預訓練模型；下半圖為各項任務泛用序列壓縮率模型，在同樣多個下游任務的情況下僅只需要單一個預訓練模型即可根據不同下游任務的需求改變序列壓縮率。

doi:10.6342/NTU202304502

在自監督式語音模型的應用當中，序列壓縮能有效的降低所需要的運算量，然而前作在實現序列壓縮時多採用單一序列壓縮率，也就是當模型預訓練完成之後僅有一個可以運作的序列壓縮率，亞氏（Apoorv Vyas）提出的模型 [22] 雖然有支援多個壓縮率的使用情況，然而僅限於兩種不同的序列壓縮率：未壓縮（20 毫秒採樣間距）及兩倍壓縮（40 毫秒採樣間距）的序列壓縮率，且僅實驗在語音辨識任務上。孟氏的研究 [13] 中發現不同的語音下游任務，對於序列壓縮的容忍程度差異甚大，舉例來說，語音辨識在接近 2 倍壓縮（40 毫秒採樣間距）之後表現急遽下降，而關鍵詞辨識任務在達到 8 倍壓縮（160 毫秒採樣間距）的情況下，表現僅僅下降了 0.6 個百分點，因此如果要對於各種不同應用情境的語音下游任務都有恰當的序列壓縮率以最大化表現運算量權衡，則前作 [11, 13] 使用的單一序列壓縮率模型需要使用多個預訓練模型來達到這個目標，如圖3.1上半所示，如此一來，不僅所需要儲存的預訓練模型參數量增加、預訓練所需之運算量也因重複多次預訓練而大幅提升，同時也減低了單一預訓練模型的泛用性。為了更進一步推進序列壓縮預訓練模型的泛用性，本論文將各項任務泛用（once-for-all）技術應用在序列壓縮自監督式語音模型的預訓練當中，在僅需要存取單一預訓練模型參數及單一次預訓練過程的情況下，讓單一預訓練模型在下游任務階段能夠根據需求選擇恰當的序列壓縮率，如圖3.1下半所示。

3.1.2 模型通用架構

為了方便討論，本章將自監督式語音模型架構廣泛的拆分為三個部分：卷積層特徵截取模組（CNN feature extractor）、轉換器編碼器層（Transformer encoder layers）、自監督式學習對象（self-supervised learning targets），如圖3.2中所示，給定一段的語音訊號，卷積層特徵截取模組首先將輸入語音轉換成為一個序列表徵向量（sequence of feature vectors），接著序列表徵向量經過多層編碼器層後得到最

doi:10.6342/NTU202304502

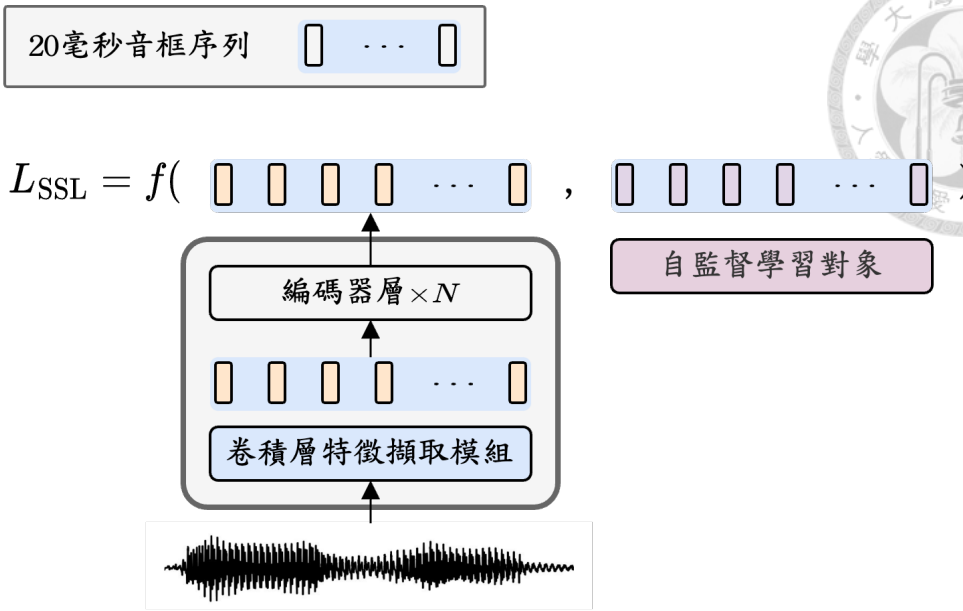


圖 3.2: 自監督式語音模型通用架構示意圖；L_{SSL} 為自監督式損失函數。

終輸出表徵（output representation），最終自監督式損失函數（self-supervised loss function, L_{SSL}）在輸出表徵和自監督式學習對象之間計算，舉例來說，Wav2Vec 2.0 是計算兩者間的對比式損失（contrastive loss）而 DistilHuBERT 則是計算兩者之間的蒸餾損失（distillation loss）。

本論文所提出的架構在自監督式模型中加入可調節式次採樣層（dynamic subsample layer）如圖3.3所示，可調節式次採樣層被放置於卷積層特徵截取模組及多層編碼器層之間，將序列表徵向量做次採樣之後再輸入至多層編碼器層，由於編碼器層輸入和輸出序列長度一致的特性，最終的輸出表徵序列也會是經過次採樣的序列。最終在計算自監督式損失函數之前，自監督式學習對象會進行同樣的次採樣（使用圖3.3的次採樣層-副），讓輸出序列表徵和自監督式學習對象的採樣頻率相同，因此原本的自監督式損失函數可以直接套用於新採樣頻率的序列。

doi:10.6342/NTU202304502

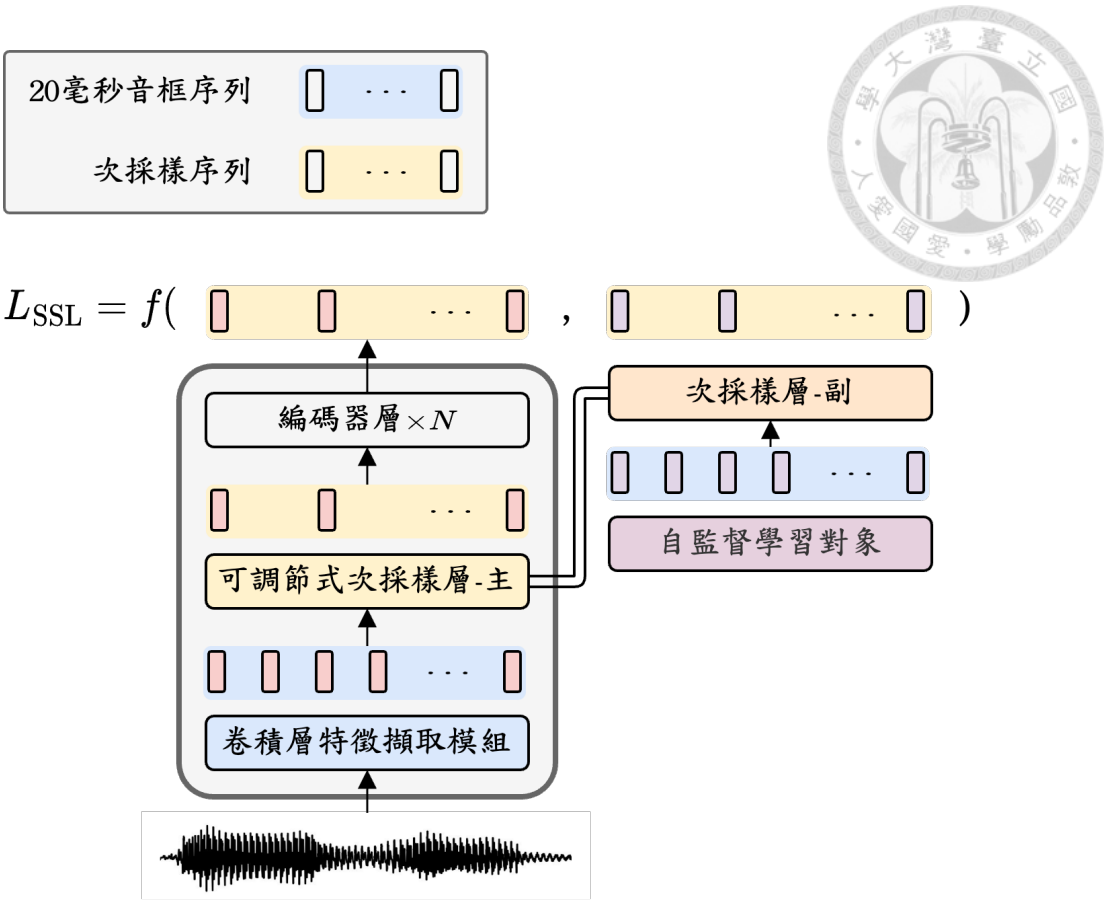


圖 3.3: 本論文提出的各項任務泛用序列壓縮法通用架構示意圖，其中次採樣層（-副）會直接複製（clone）可調節式次採樣層（-主）的次採樣方式。

3.1.3 可調節式次採樣層

可調節式次採樣層設計如圖3.4所示，首先，輸入的 20 毫秒音框序列會經過權重預測模組預測出一組原始權重 $\alpha_{1:T}$，接著此權重會經過權重改動模組，我們引入一個用以調節壓縮率的純量參數 $\lambda \in [0, 2)$ 作為權重預測模組的輸入，根據不同的 λ 值，改動之後的權重以 $\alpha_{1:T}^{\text{mod}}$ 表示，分成以下三種情境討論：

- 情境一：當 $\lambda \in [0, 1)$ 時，改動 α^{mod} 公式如下，

$$\alpha_i^{\text{mod}} = \lambda \alpha_i + (1 - \lambda) \tag{3.1}$$

doi:10.6342/NTU202304502

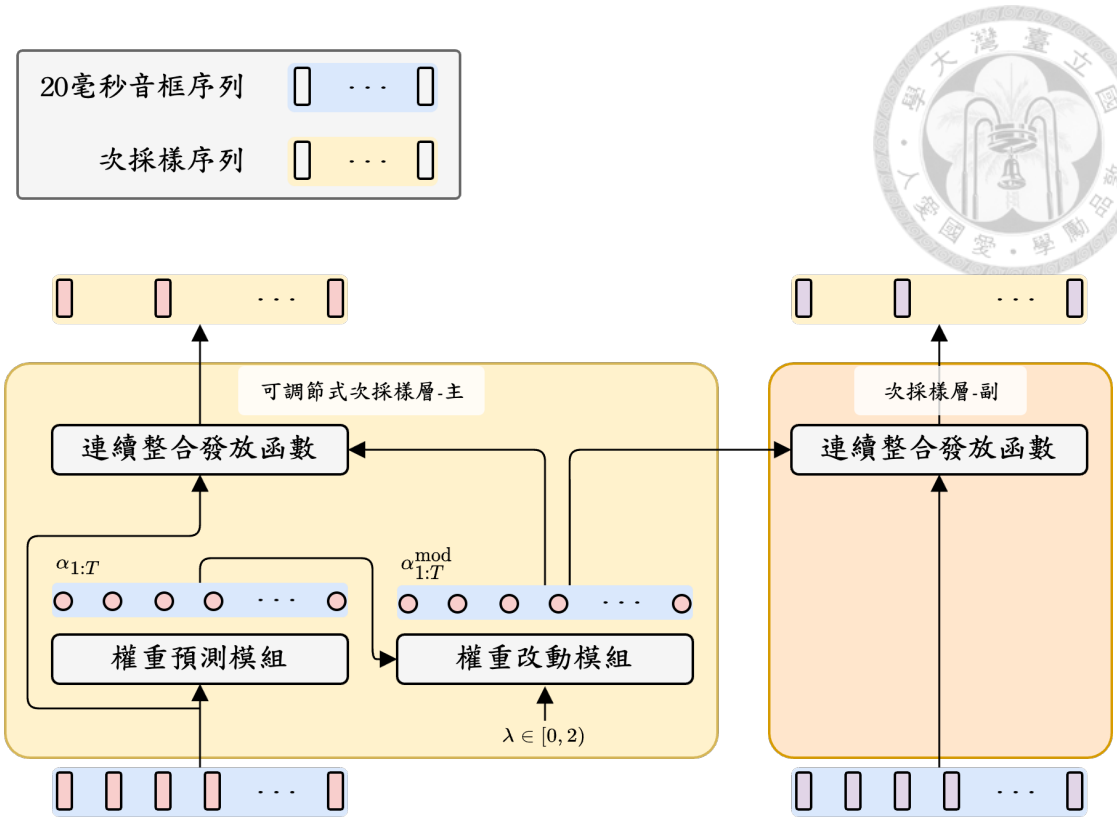


圖 3.4: 可調節式次採樣層架構示意圖。

- 情境二：當 $\lambda \in [1, 2)$ 且 $(2-\lambda)\sum_{i=1}^{T} \alpha_i \geq 1$ 時，改動 α^{mod} 公式如下，

$$\alpha_i^{\text{mod}} = (2-\lambda)\alpha_i \tag{3.2}$$

- 情境三：當 $\lambda \in [1, 2)$ 且 $(2-\lambda)\sum_{i=1}^{T} \alpha_i < 1$ 時，改動 α^{mod} 公式如下，

$$\alpha_i^{\text{mod}} = \frac{\alpha_i}{\sum_{i=1}^{T} \alpha_i} \tag{3.3}$$

在式3.1及式3.2中，當 $\lambda = 1$ 時兩式均化簡為 $\alpha_i^{\text{mod}} = \alpha_i$；在式3.2及式3.3中，當 $(2-\lambda)\sum_{i=1}^{T} \alpha_i = 1$ 時兩式均簡化為 $\alpha_i^{\text{mod}} = (2-\lambda)\alpha_i$，因此以上三式可以用單

doi:10.6342/NTU202304502

一個連續且分段線性的函數 F 結合，

$$F(\alpha_i, \lambda) = \begin{cases} \lambda\alpha_i + (1-\lambda), & \text{if } 0 \leq \lambda < 1 \\ (2-\lambda)\alpha_i, & \text{if } 1 \leq \lambda < 2 \text{ and } (2-\lambda)\sum_{i=1}^{T}\alpha_i \geq 1 \\ \frac{\alpha_i}{\sum_{i=1}^{T}\alpha_i}, & \text{if } 1 \leq \lambda < 2 \text{ and } (2-\lambda)\sum_{i=1}^{T}\alpha_i < 1 \end{cases} \tag{3.4}$$

，同時確保其分段可微之特性。此處的權重改動可以視為是將原本的權重 $\alpha_{1:T}$ 做上取樣（up-scaling）及下取樣（down-scaling），在原本連續整合發放機制中，權重預測模組會經過 S 函數（Sigmoid function）將權重限制在 0 到 1 中間，來確保在接下來通過連續整合發放函數的時候單一個音框（time frame）不會對應到多個觸發事件（fire event），基於同樣的理由，我們希望經過權重改動模組改動後的權重 $\alpha_{1:T}^{\text{mod}}$ 同樣落在 0 到 1 中間。

在上取樣的情況中，如果直接使用倍率上取樣（naive scaling）則無法避免在某些時間點出現大於 1 的權重值，因此在上取樣的情況中，我們將原始的權重 $\alpha_{1:T}$ 值與 1 進行內差，如式3.1所示。在下取樣的情況中，如果使用同樣的式3.1，則更改過後的 α^{mod} 值會出現負值，這個情況也是我們希望避免的，因此在下取樣的情況中我們使用的是倍率下取樣，如式3.2所示。另外，為了避免在下取樣的情況中自監督式模型沒有輸出任何表徵導致損失函數無法計算，在下取樣的情況中加入另外一個判斷條件：下取樣之後預計的總發放次數為 $(2-\lambda)\sum_{i=1}^{T}\alpha_i$ 次，如果預計發放次數為大於等於一次的話，則使用3.2式的下取樣公式，如預計發放次數小於一次的話，使用3.3式的權重改動以確保最少一次的發放事件會發生。

最終，連續整合發放函數輸入原始 20 毫秒音框序列及 $\alpha_{1:T}^{\text{mod}}$ 做為整合發放的權重進行發放，連續整合發放函數的輸出序列即為壓縮過後的序列。最終的輸出表徵序列長度可以表示為 $\sum F(\alpha_i, \lambda)$，其與 λ 值的關係如圖3.5所示，總結來說，

doi:10.6342/NTU202304502

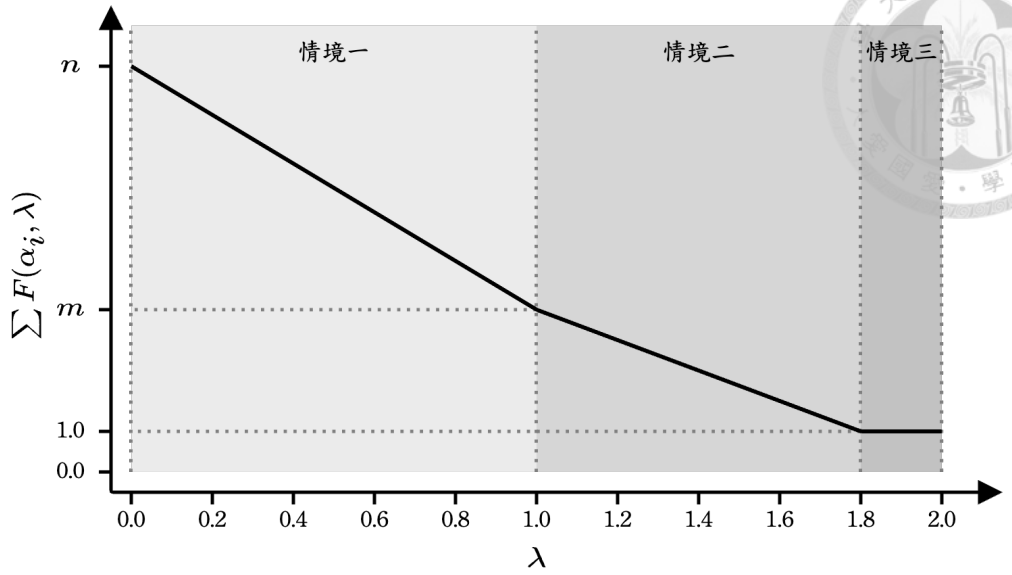


圖 3.5: $\sum F(\alpha_i, \lambda)$ 對 λ 示意圖，其中 $\sum F(\alpha_i, \lambda)$ 為對應 λ 經過權重改動模組後，對應的輸出序列長度，其中 n 為原始未經壓縮的自監督式語音模型輸出序列長度；m 為使用孟氏 [13] 單一壓縮率模型對應的輸出序列長度。

在 $\lambda = 0$ 的情況下，此預訓練模型等價於原始未使用序列壓縮的自監督式語音模型，也就是每一個固定間距為 20 毫秒的音框（time frame）對應到一個輸出的表徵向量。在 $\lambda = 1$ 的情況下，此模型等同於單一個由孟式 [13] 所提出的可變間距序列壓縮法模型。當 λ 從 2 的下界逼近 2 的時候，也就是 $\lambda \to 2^-$ 時，此模型對任一長度的輸入語音均輸出單一個表徵向量。

3.1.4 模型預訓練方式

在模型預訓練階段，針對每一筆的批次資料（batch data），λ 值以任意的機率分佈在 $\lambda \in [0, 2)$ 中取樣，根據取樣到的 λ 值對這一筆的批次資料進行權重改動，權重改動模組與連續整合發放次數關係示意圖如圖3.6所示。λ 取樣的函數可以是任意的機率密度函數，在本論文的實驗中，λ 取樣的機率分佈使用連續型均勻分布（uniform distribution），只改動連續型均勻分布的上下界範圍如圖3.7所示，共測試了三個不同的機率分佈：型 1、型 2、型 3 分別為 $\lambda \in [0, 1]$、$\lambda \in [1, 1.5]$、

doi:10.6342/NTU202304502

權重改動模組

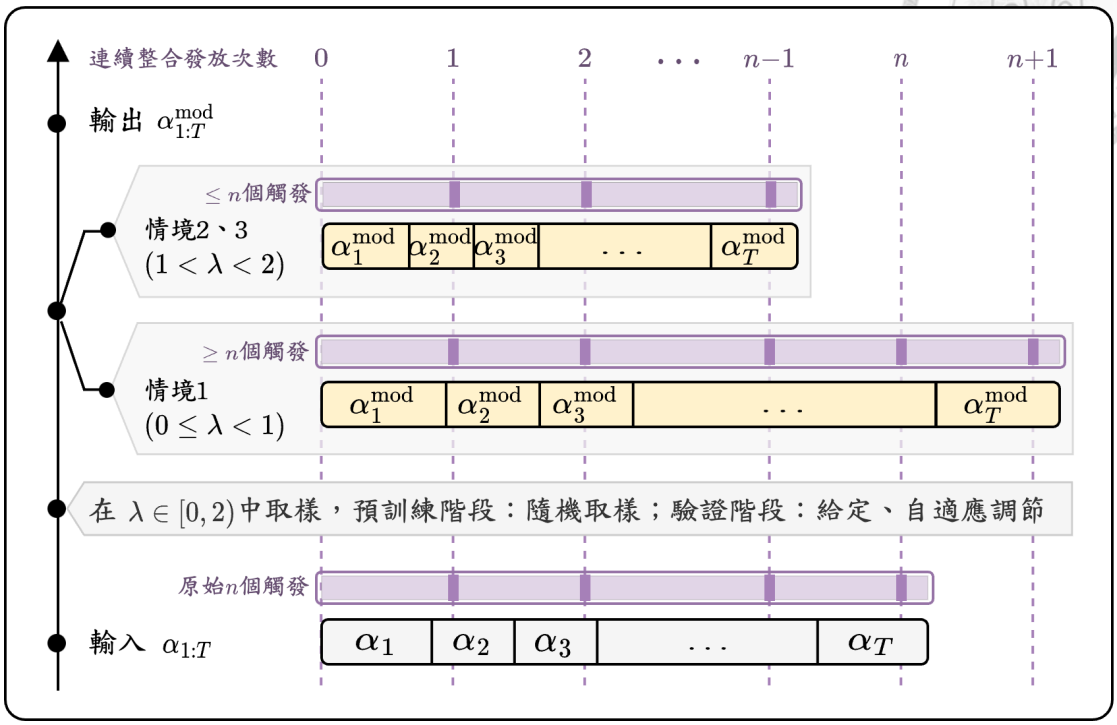


圖 3.6: 權重改動模組示意圖，λ 取樣的方式在預訓練階段使用隨機取樣；在驗證階段可以使用兩種取樣方式：給定或自適應調節，第三章中探討給定 λ 的驗證結果；第四章中探討自適應調節 λ 的驗證結果。

$\lambda \in [0, 2)$ 的連續型均勻分佈機率函數。對於每一筆批次資料，根據所取樣到的 λ 值進行權重改動之後，將改動後權重 $\alpha_{1:T}^{\text{mod}}$ 對應的連續整合發放次採樣序列通過編碼器層之後計算自監督式損失函數。由於改動後的權重會根據取樣到的 λ 值隨機改動，因此分割學習引導的兩個額外的損失函數：音框層級和分割層級損失函數，均是使用在未改動的權重 $\alpha_{1:T}$ 上。

3.1.5 模型驗證方式

在模型驗證階段，不同於在訓練階段中每筆批次資料採樣不同的 λ 值，在驗證階段使用給定或是自適應調節的方式選取 λ，在本章的實驗中均使用給定的驗證方式；自適應調節的驗證方式及結果在第四章中探討。本章中所實驗的下游任務，λ 值在驗證階段皆保持不變，而 λ 值的大小則根據所需要的壓縮率在

doi:10.6342/NTU202304502

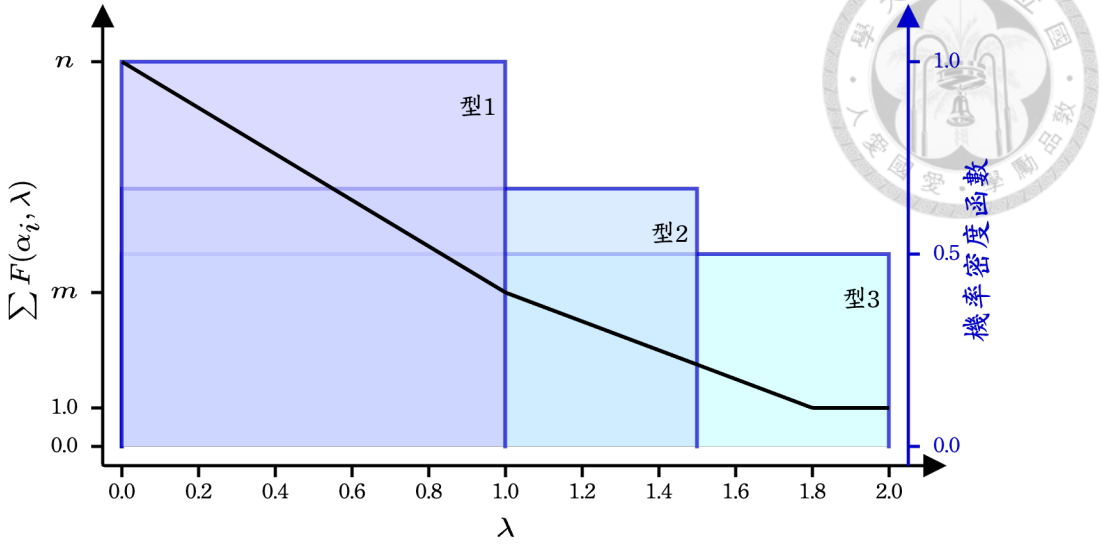


圖 3.7: 本論文所使用在預訓練階段 λ 機率分佈函數示意圖，共三種機率分佈：型 1、型 2、型 3，分別為 $\lambda \in [0, 1]$、$\lambda \in [1, 1.5]$、$\lambda \in [0, 2)$ 的連續型均勻分佈機率函數。

驗證任務開始前事先訂好。λ 值在預訓練完成後會以一對一的關係對應到一個序列壓縮率，而實際對應關係則會根據在預訓練階段使用不同的分割引導訓練（segmentation guidance）以及預訓練參數有所不同。

3.2 各項任務泛用序列壓縮法用於知識蒸餾模型

3.2.1 簡介

本節將所提出的各項任務泛用序列壓縮法用於知識蒸餾自監督式語音模型 DistilHuBERT [4] 上，此模型僅使用兩層編碼器層，有模型參數量小、預訓練速度快、等優勢，在同樣的時間及運算限制下能夠進行較多實驗，本節實驗以 DistilHuBERT 為基礎驗證不同參數設定對於各項任務泛用序列壓縮預訓練模型的影響。

doi:10.6342/NTU202304502

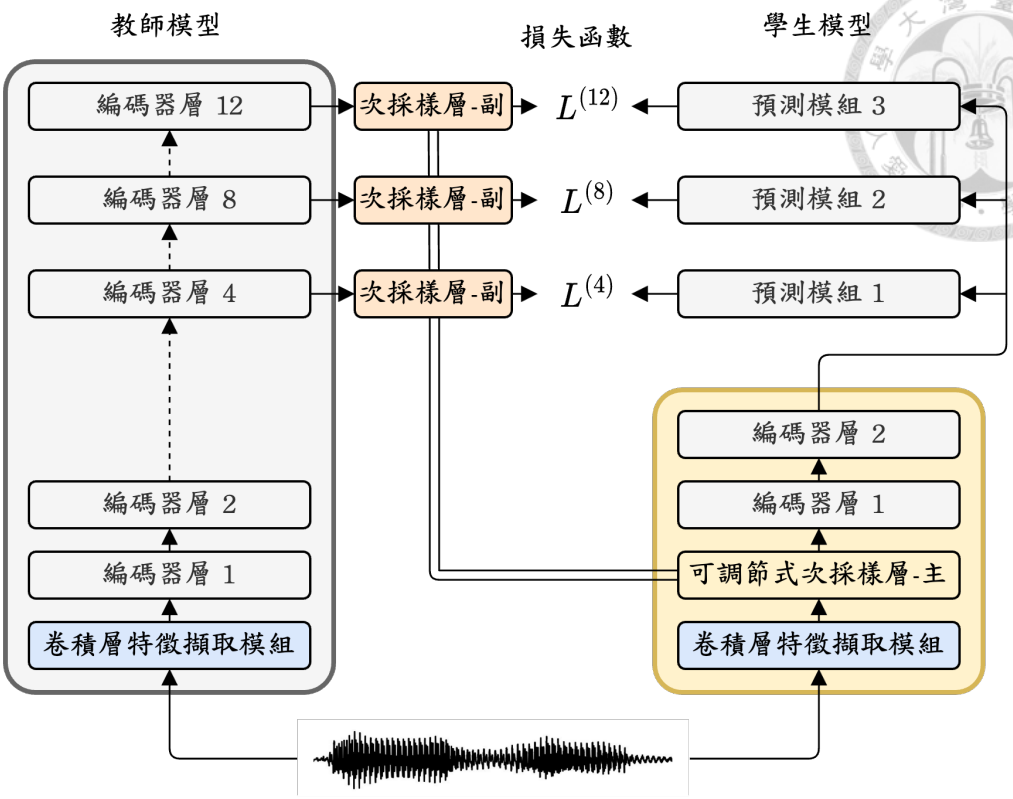


圖 3.8: 各項任務泛用序列壓縮法用於 DistilHuBERT 的模型架構示意圖。

3.2.2 模型架構

模型架構如圖3.8所示，可調節式次採樣層加入在學生模型的卷積層特徵截取模組及編碼器層之間，而 DistilHuBERT 的自監督式學習對象為教師模型的第 4、8、12 層編碼器層的表徵輸出，因此次採樣層（-副）會接在對應這幾層教師模型的編碼器層，將對應這幾層教師模型的輸出表徵做和學生模型相同的次採樣。

3.2.3 模型參數設置

本節實驗的各項任務泛用序列壓縮預訓練模型，除了額外的可調節式次採樣層之外，和 DistilHuBERT 所使用的模型參數和訓練參數均相同，設定細節如下：學生模型的卷積層及編碼器層分別使用教師模型的卷積層及編碼器層前兩層做初始化，總共使用三個預測模組（prediction head）分別預測教師模型的第 4、8、12

doi:10.6342/NTU202304502

層的表徵序列，訓練更新次數為 200,000 次、批次大小（batch size）為 24。學習率（learning rate）在前百分之七的更新步數線型遞增到 2e-4 接著線型遞減到 0。所使用的預訓練資料為 960 小時的 LibriSpeech[15] 資料集。

本論文使用和孟氏 [13] 同樣的分割引導訓練，並且使用其中表現運算量權衡（performance-efficiency trade-off）最好的非監督式語音辨識（unsupervised ASR）所產生的分割。其產生方式如下，由一個簡化版本劉氏（Alexander H. Liu）所提出的非監督式語音辨識系統 Wav2Vec-U 2.0 [12]，簡化其 k-means 群集損失，並訓練在 100 小時的 LibriSpeech 的語音資料及 LibriSpeech 語言模型的文字資料上。在訓練結束之後，非監督式語音辨識模型被當作是音框層級的分類器，最後，分割的產生來自於分類器所預測重複的音框範圍，所使用的分割引導權重分別為 $\gamma_{\text{frame}} = 0.25, \gamma_{\text{seg}} = 5e-3$。

此小節的實驗總共實驗三個不同的 λ 機率分佈，分別為 $\lambda \in [0, 1]$、$\lambda \in [0, 1.5]$、$\lambda \in [0, 2)$，SUPERB 基準預設使用的驗證方法為將各層的表徵序列做加權和（weighted sum），然而在使用兩層編碼器層預訓練模型的研究中，張氏 [4] 及孟氏 [13] 均是使用最後一層輸出表徵作為下游任務的輸入，為了讓實驗設定一致，本節實驗結果亦是僅使用最後一層輸出表徵序列的驗證結果。

3.2.4 實驗結果與分析

本小節分析實驗結果，以任務類別作為分類分成五個類別：內容相關、語者相關、語義相關、副語言相關、語義和生成相關。同時比較前作 [4, 13] 中單一序列壓縮率預訓練模型的結果，以連續曲線代表本論文提出的各項任務泛用序列壓縮法的結果，其中同一條曲線為單一個預訓練模型僅在驗證階段改變不同的序列壓縮程度，三條曲線分別代表三種不同預訓練階段使用的 λ 機率分佈。每一條

doi:10.6342/NTU202304502

曲線根據不同的任務特性，取樣序列壓縮率進行網格搜尋（grid search），並依據 SUPERB 基準的規範進行驗證任務最終連線而得到。

3.2.4.1 內容相關任務

首先分析內容相關任務，內容相關任務包含了音素辨識、語音辨識、關鍵詞辨識、按例查詢，結果如圖3.9所示。其中音素辨識及語音辨識為鏈接式時序分類任務，下游模型在預測時會先輸出和輸入序列長度一致的序列，其中可能包含了多個連續重複的字符以及空字符（null token），接著再經過塌縮函數將重複字符及空字符去除，因此這一類型的任務有一個基本的要求是輸入序列長度必須要大於或等於輸出序列長度，而音素辨識所使用的字符為音素單元（phone unit）其平均字符長度約為 100 毫秒，語音辨識則是使用字母單元（character unit）其平均字符長度約為 60 毫秒，由結果觀察到在接近這兩個任務的物理極限時兩任務的錯誤率大幅增加。關鍵詞辨識為序列層級合計任務，也就是輸入語音所對應到的輸出序列表徵會在序列的方向進行合計之後輸入給下游任務，這一類的任務對於輸出序列的長度並沒有物理上的限制，但是以壓縮過後的序列通過自監督式語音模型編碼器層亦會造成表現遞減，可以觀察到其正確率隨著壓縮率增加而遞減，然而其遞減的幅度小於使用鏈結式時序分類的任務：音素辨識及語音辨識。案例查詢為序列層級比對任務，使用動態時間校正演算法（dynamic time wraping, DTW）來比對給定的兩個時間序列的相似性（目標語音對應到的序列表徵以及受測語音對應到的序列表徵），在此框架下目標語音及受測語音均會受到同樣幅度的序列壓縮，對於序列長度並無如鏈結式時序分類任務有嚴格的限制，但是壓縮的程度亦會影響到其辨識相似度的準確性，實驗結果觀察到在大於 100 毫秒平均採樣間距之後的表現會明顯遞減。

doi:10.6342/NTU202304502

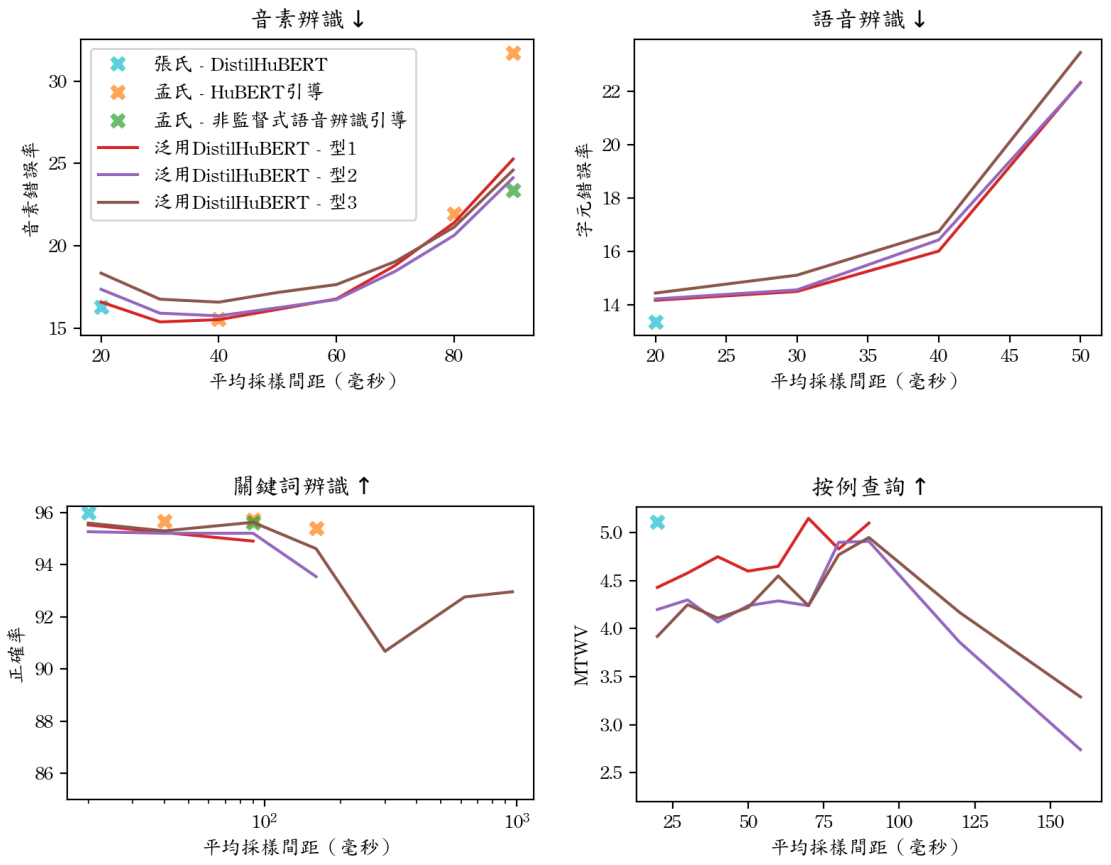


圖 3.9: 內容相關任務驗證結果，其中關鍵詞辨識的橫坐標為對數座標。評斷指標包含了音素錯誤率（phone error rate, PER）、字元錯誤率（character error rate, CER）、正確率（accuracy）、MTWV 分數（maximum term weighted value, MTWV）。

3.2.4.2 語者相關任務

語者相關任務包含了語者辨識，結果如圖3.10所示。語者辨識和關鍵詞辨識一樣為序列層級合計任務，對於序列壓縮的容忍程度相對較大，另外比較三種機率分佈函數（圖3.7中的型 1、型 2、型 3 機率分佈）預訓練模型的表現差異，可以觀察到在此任務上，機率分佈較小（也就是需要涵蓋的壓縮率範圍較小；型 1 機率分佈）的預訓練模型表現相對較好，而機率分佈範圍大的預訓練模型（型 3 機率分佈）表現相對較差但可以涵蓋的壓縮率範圍較廣。

doi:10.6342/NTU202304502

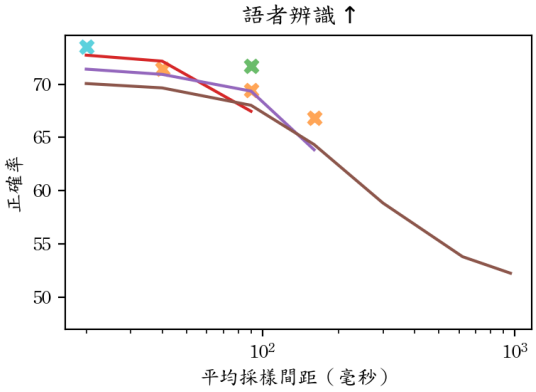


圖 3.10: 語者相關任務驗證結果，語者辨識的橫坐標為對數座標，評斷指標為正確率（accuracy）。

3.2.4.3 語義相關任務

語義相關任務包含了意圖辨識及填空任務，結果如圖3.11所示。其中意圖辨識為序列層級合計任務，隨著壓縮率增加，表現緩慢遞減，而填空任務為鏈結式時序分類任務，使用的字符為字母單元（character unit），可以觀察到在 60 毫秒平均採樣間距附近表現開始急遽下降。在這兩個任務中，三種不同的機率分佈對預訓練模型表現的影響並不大。

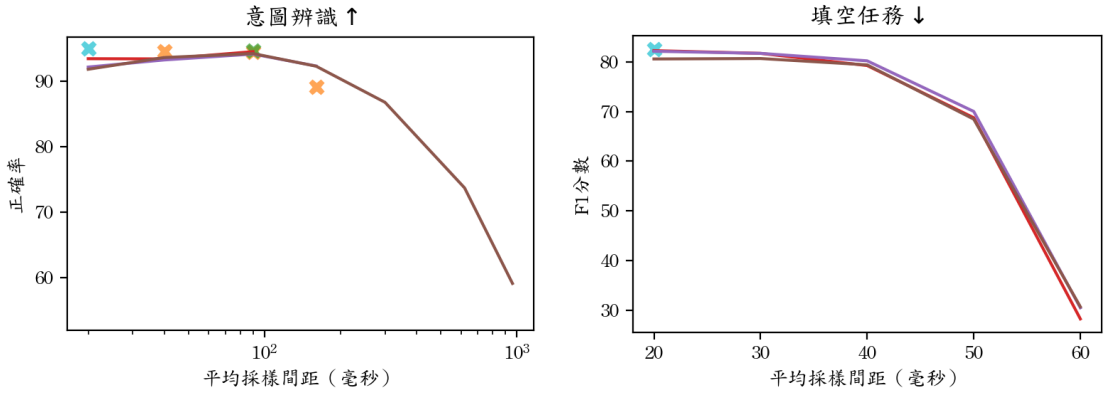


圖 3.11: 語義相關任務驗證結果，其中意圖辨識的橫坐標為對數座標，評斷指標包含了正確率（accuracy）及 F1 分數（F1 score）。

doi:10.6342/NTU202304502

3.2.4.4 副語言相關任務

副語言相關任務包含了情感辨識，結果如圖3.12所示。情感辨識為序列層級合計任務，序列壓縮對於此任務表現影響不大，在 8 倍壓縮率下（平均採樣間距 160 毫秒），相較於未壓縮模型（採樣間距 20 毫秒）僅有 1.6 個百分點的正確率下降；在最大 48 倍壓縮率下（平均採樣間距 960 毫秒），相較於未壓縮模型僅有 4.0 個百分點的正確率下降。

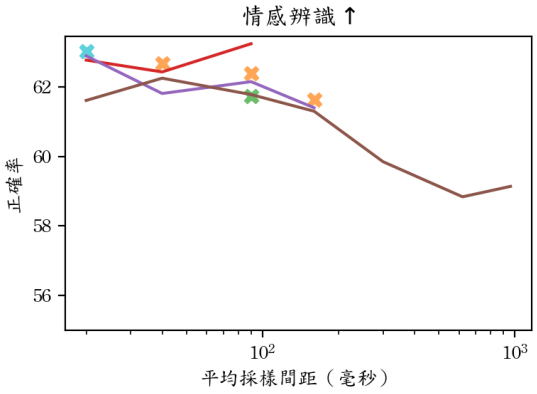


圖 3.12: 副語言相關任務驗證結果，情感辨識的橫坐標為對數座標，評斷指標為正確率（accuracy）。

3.2.4.5 語義和生成相關任務

語義和生成相關任務包含了語音翻譯，結果如圖3.13所示。語音翻譯為序列至序列任務（sequence-to-sequence），由於輸出序列是以自回歸的方式生成，對於輸入序列長度並沒有像鏈結式時序分類任務有嚴格限制，但是由實驗結果可以發現，在語音翻譯這個任務上，表現儘管沒有在接近特定序列壓縮率時急遽下降，壓縮序列程度也會大幅的影響語音翻譯的結果。

doi:10.6342/NTU202304502

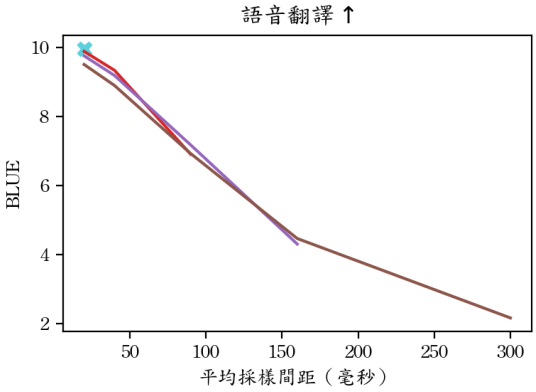


圖 3.13: 語義和生成相關任務驗證結果，評斷指標為 BLUE 分數（BLUE score）。

3.2.5 運算成本分析

在運算成本上，本小節分析自監督式語音模型所需要的乘積累加運算量（Multiply-accumulate operation, MACs），分析的方式使用 SUPERB 基準中評斷乘積累加運算量的標準[1]：將 LibriSpeech 資料集中給定的 32 筆音訊（音訊長度介於 1 秒到 20 秒之間），做為自監督式語音模型的輸入，計算模型所需的總乘積累加運算量。本小節分析各項任務泛用序列壓縮 DistilHuBERT（以型 3 機率分布預訓練）在不同序列壓縮率下所需要的運算量，同時，為了分析序列壓縮對於自監督語音模型各個模組的影響，結果呈現三個模組：卷積層特擷取模組、可調節式次採樣層、編碼器層分別的運算量占比，分析結果如圖3.14所示，以下分別根據不同的模組進行討論：

- 卷積層特徵擷取模組：在 DistilHuBERT 僅使用兩層編碼器層的架構中，卷積層特徵擷取模組所占之運算量比例較大，在未壓縮模型中占了接近 70 個百分點的運算量。

- 可調節式次採樣層：可調節式次採樣層中的連續整合發放機制會造成額外的運算量，但是占比極小，在原始未壓縮模型中僅占了 1.9 個百分點的運算

[1] https://superbbenchmark.org/challenge-slt2022/metrics

doi:10.6342/NTU202304502

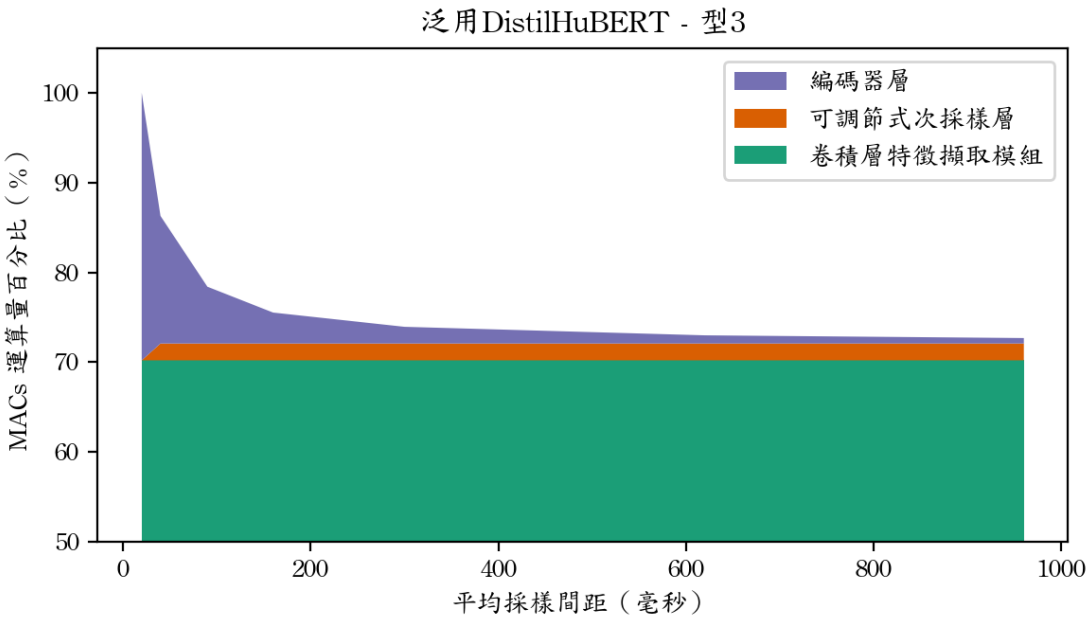


圖 3.14: 以型 3 機率分布預訓練之各項任務泛用序列壓縮 DistilHuBERT，所需之乘積累加運算量（MACs）對平均採樣間距關係圖。運算量以相對與原始 DistilHuBERT 運算量所占百分比呈現，結果中一併呈現三個模組分別所占的比例。

量，此外，當各項任務泛用序列壓縮模型運作在未壓縮的情況下（20 毫秒採樣間距），可調節式次採樣層可以直接被略過而不需要額外的運算量。

- 編碼器層：隨著序列壓縮率增加，編碼器層中所需的運算量明顯下降。

總結來說，在同時考慮可調節式次採樣層及編碼器層所需運算量的情況下，相較於原始未壓縮模型（20 毫秒採樣間距），壓縮至 90 毫秒及 960 毫秒平均採樣間距的序列分別減少了 72.5 及 91.7 個百分點的乘積累加運算量。由於本章所實驗的模型為小型的知識蒸餾模型，編碼器層所占的運算量比例較低，使用序列壓縮能夠減少的運算量占比相對有限，在下一節的實驗中，本論文將各項任務泛用序列壓縮法應用於大型的對比式預訓練模型中。

doi:10.6342/NTU202304502

3.3 各項任務泛用序列壓縮法用於對比式預訓練模型

3.3.1 簡介

前小節使用的知識蒸餾模型 DistilHuBERT 僅包含了兩層的編碼器層，有預訓練速度快的優勢，並能夠在有限的運算限制下達到和教師模型接近的表現，然而在整體的表現上，小型知識蒸餾模型往往還是會和大型預訓練模型有一段差異。在本節中，本論文將各項任務泛用序列壓縮法用於大型對比式預訓練模型（共十二層編碼器層）Wav2Vec 2.0 [1] 上，以回答下列兩個問題：

- 各項任務泛用序列壓縮法是否適用於使用對比式損失函數的大型自監督式語音模型上？
- 同樣使用各項任務泛用序列壓縮的情況下，大型預訓練模型相較於小型知識蒸餾模型所帶來的優勢為何？

3.3.2 模型架構

各項任務泛用序列壓縮法用於 Wav2Vec 2.0 模型架構如圖3.15所示，可調節式次採樣層加入在自監督式語音模型的卷積層特徵截取模組和多層編碼器層之間，自監督式語音模型中的表徵序列經過次採樣後，接著通過多層編碼器層。Wav2Vec 2.0 中所使用的對比式學習對象為經過向量量化後的表徵序列，在此架構中，自監督式學習對象同時會被可調節式次採樣層所次採樣，因此將各項任務泛用序列壓縮法應用於 Wav2Vec 2.0 時，不需要額外的次採樣層來次採樣自監督式學習對象。

doi:10.6342/NTU202304502

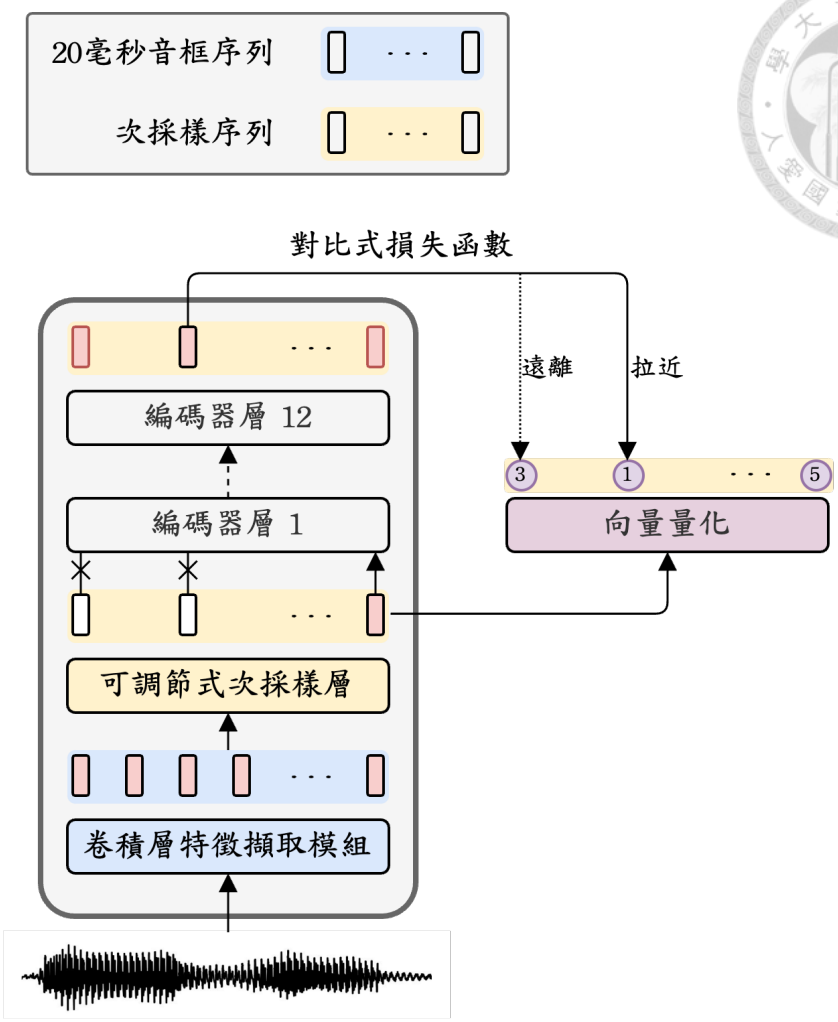


圖 3.15: 各項任務泛用序列壓縮法用於 Wav2Vec 2.0 的模型架構示意圖。

3.3.3 模型參數設置

Wav2Vec 2.0 為一個大型的自監督式語音模型，其隨機初始化開始訓練至收斂的運算量非常龐大，在貝氏（Alexei Baevski）[1] 的實驗設置中實驗了兩種模型架構：基礎（base）及大型（large）模型，分別包含了十二層及二十四層的編碼器層，在運算資源上分別需要使用 64 個 V100 型 GPU 訓練了 1.6 天及 128 個 V100 型 GPU 訓練 2.3 天，在進行各項任務泛用序列壓縮預訓練時，使用隨機初始化訓練的運算成本過於龐大，因此在本節的實驗中，預訓練模型架構採用較小的基礎（base）模型，並以由貝氏釋出的預訓練模型參數[2]作為模型初始化參數，同樣為

[2] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt

doi:10.6342/NTU202304502

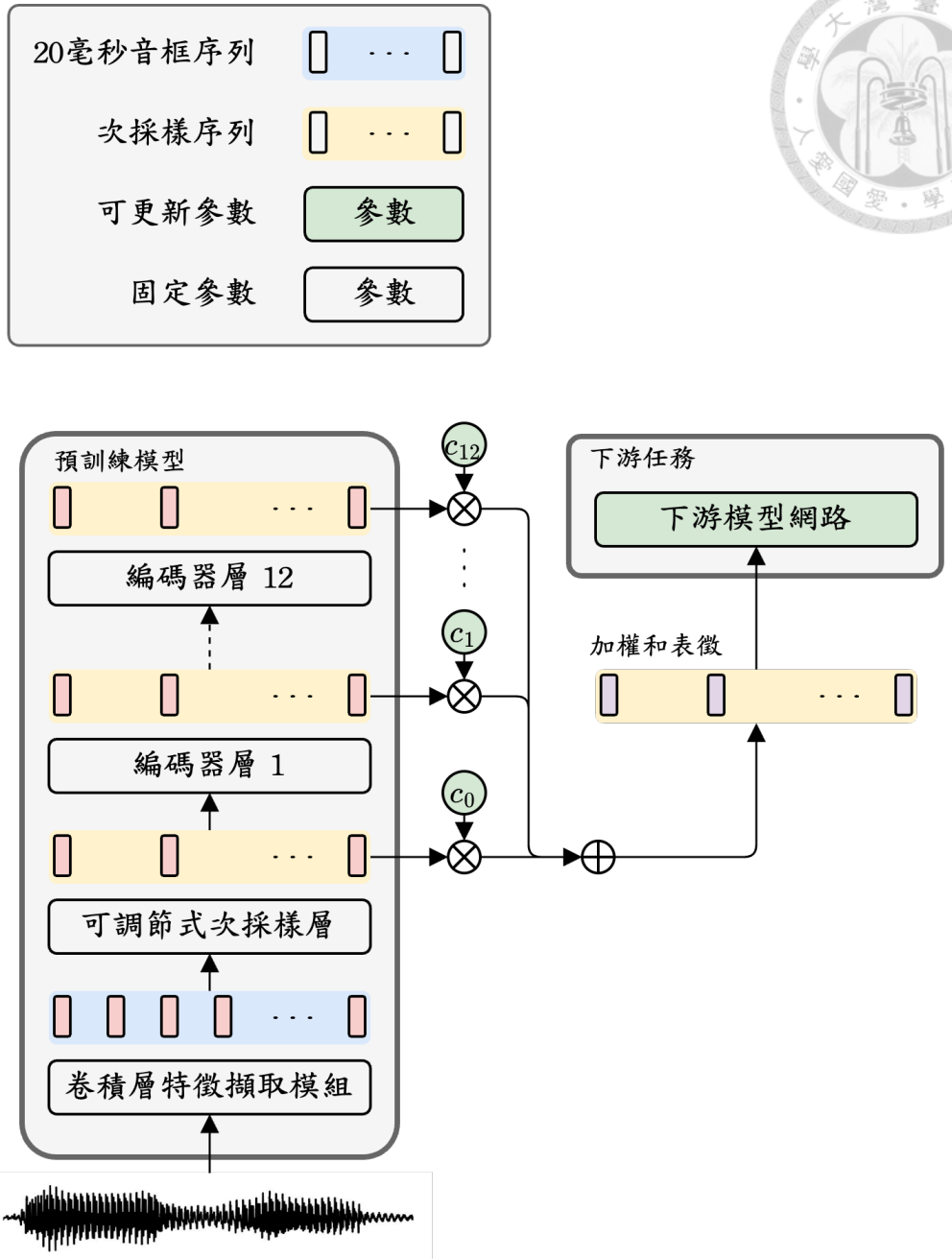


圖 3.16: Wav2Vec 2.0 模型下游任務驗證階段所使用的模型架構示意圖，其中 $\{c_0, c_1, \ldots, c_{12}\}$ 為各層表徵序列加權和之權重，為可更新至參數和下游任務模型網路之參數一起更新。

了節省運算成本，本節的實驗只訓練在其中一個機率分佈：型 1 機率分佈上。

在模型參數設置上，除了額外的可調節式次採樣層之外，所使用的模型參數和訓練參數均和預設的 Wav2Vec 2.0 一樣，包含了遮罩個數上限、損失函數比重、等等。在分割引導訓練上，本節的模型採用和前一小節相同由非監督式語音辨識所產生的分割作為引導，詳細的分割取得方式可以參考第3.2.3節的設定，使用的

doi:10.6342/NTU202304502

分割引導權重分別為 $\gamma_{\text{frame}} = 0.5, \gamma_{\text{seg}} = 5e-2$，訓練更新的步數為 5,000。

3.3.4 實驗結果與分析

本小節分析實驗結果，以任務類別作為分類分成五個類別：內容相關、語者相關、語義相關、副語言相關、語義和生成相關。在進行驗證任務的時候，本小節的實驗採用 SUPERB 基準規範，使用卷積層特徵截取模組後及各層編碼器層的輸出表徵序列，在各層之間做加權和（weighted-sum）後做為下游任務的輸入。下游驗證任務架構如圖3.16所示：每一層輸出的表徵序列會對應到一個權重，分別將每一個音框以對應權重做加權和，得到最終的加權和表徵序列，接著將此加權和表徵序列做為下游任務的輸入，其中權重為可更新參數和下游模型網路的參數一起更新。實驗結果根據不同任務的特性，取樣適當的序列壓縮率進行網格搜尋，並依據 SUPERB 基準規範進行驗證任務最終連線而得。

3.3.4.1 內容相關任務

內容相關任務包含了音素辨識、語音辨識、關鍵詞辨識、案例查詢，結果如圖3.17所示。由結果可以觀察到，在兩個使用鏈接式時序分類的任務：音素辨識及語音辨識中，對比式預訓練模型的表現明顯優於知識蒸餾模型，在屬於序列層級合計任務的關鍵詞辨識中，兩者的表現差異不大，然而在屬於序列層級比對任務的案例查詢中，未經壓縮的 DistilHuBERT 表現就優於 Wav2Vec 2.0，在這個任務中，序列壓縮對於對比式預訓練模型的影響更加明顯。

doi:10.6342/NTU202304502

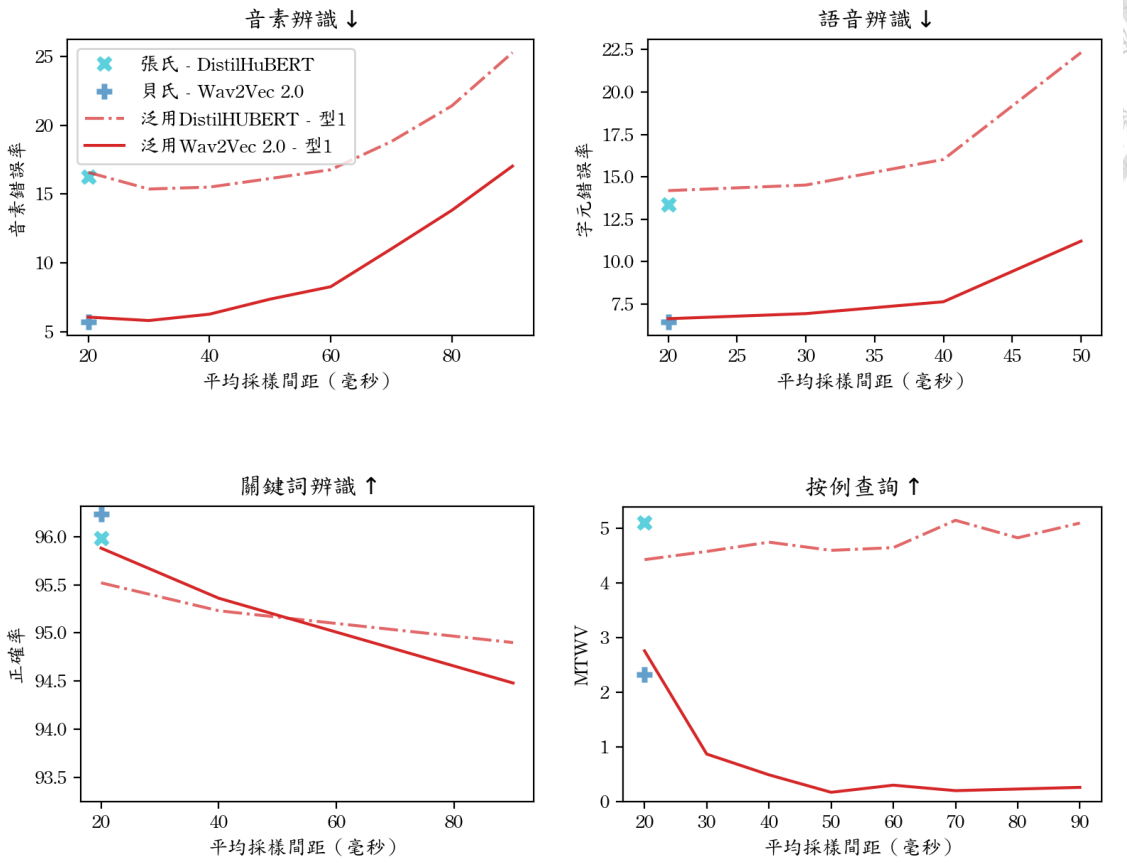


圖 3.17: 內容相關任務驗證結果，其中關鍵詞辨識的橫坐標為對數座標。評價指標包含了音素錯誤率（phone error rate, PER）、字元錯誤率（character error rate, CER）、正確率（accuracy）、MTWV 分數（maximum term weighted value, MTWV）。

3.3.4.2 語者相關任務

語者相關任務包含了語者辨識，結果如圖3.18所示。語者辨識為序列層級合計任務，序列壓縮對於表現的影響相對較小，然而在這個任務中，儘管未經壓縮的 Wav2Vec 2.0 的表現優於 DistilHuBERT，經過各項任務泛用序列壓縮預訓練後的對比式預訓練模型的整體表現明顯下降。

3.3.4.3 語義相關任務

語義相關任務包含了意圖辨識及填空任務，結果如圖3.19所示。意圖辨識為序列層級合計任務，序列壓縮對其表現影響不大，在小壓縮率的情況下，對比式

doi:10.6342/NTU202304502

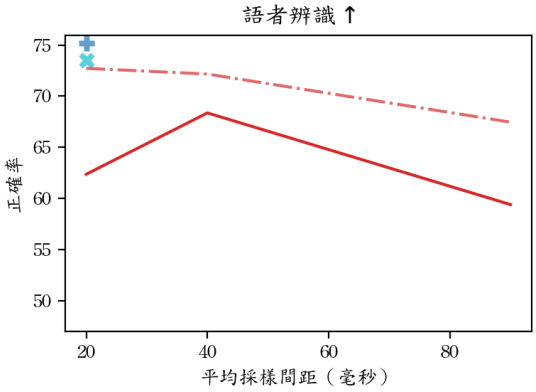


圖 3.18: 語者相關任務驗證結果，語者辨識的橫坐標為對數座標，評斷指標為正確率（accuracy）。

預訓練模型稍有優勢，而填空任務為鏈接式分類任務，使用對比式預訓練模型的結果表現較佳，尤其是在較大壓縮率的情況下（60 毫秒平均採樣間距），表現差異最為明顯。

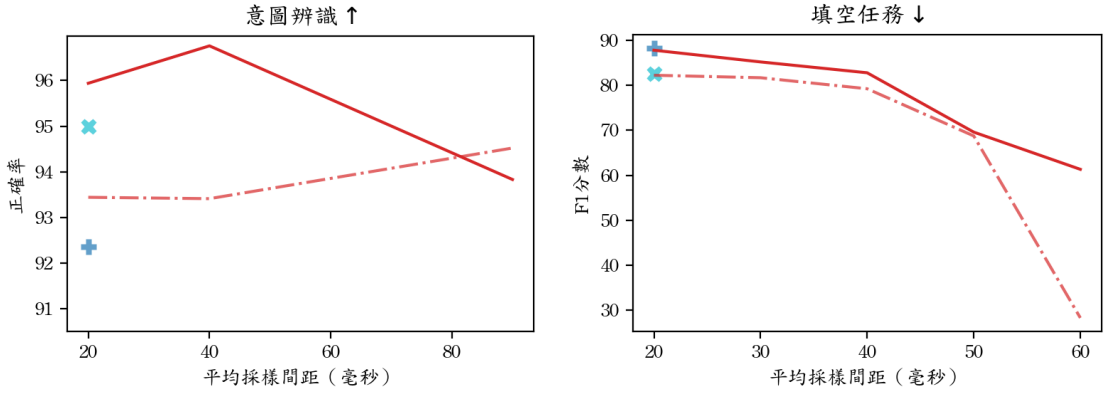


圖 3.19: 語義相關任務驗證結果，其中意圖辨識的橫坐標為對數座標，評斷指標包含了正確率（accuracy）及 F1 分數（F1 score）。

3.3.4.4 副語言相關任務

副語言相關任務包含了情感辨識，結果如圖3.20所示。情感辨識為序列層級合計任務，其結果對於序列壓縮率並不敏感，在此任務中，原始模型 Wav2Vec 2.0 和 DistilHuBERT 結果差異不大，序列壓縮對於對比式預訓練模型的表現影響相對

doi:10.6342/NTU202304502

較大，而知識蒸餾模型在此壓縮率範圍內幾乎不受影響。

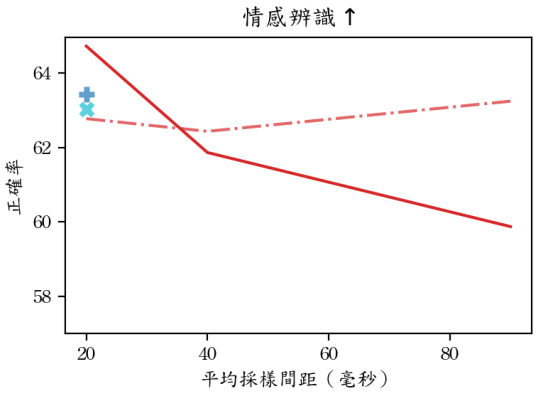


圖 3.20: 副語言相關任務驗證結果，情感辨識的橫坐標為對數座標，評斷指標為正確率（accuracy）。

3.3.4.5 語義和生成相關任務

語義和生成相關任務包含了語音翻譯，結果如圖3.21所示。語音翻譯為序列至序列任務，從對比式預訓練模型的結果可以觀察到和知識蒸餾模型有同樣的趨勢，對於序列壓縮沒有一個明顯表現變差的轉折點，而是隨著壓縮率增加，表現穩定下降。在此任務中，對比式預訓練模型的表現明顯優於知識蒸餾模型。

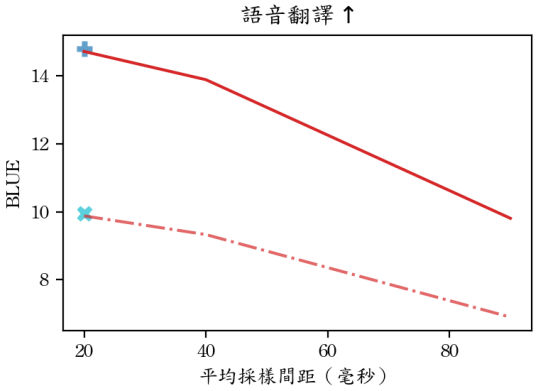


圖 3.21: 語義和生成相關任務驗證結果，評斷指標為 BLUE 分數（BLUE score）。

doi:10.6342/NTU202304502

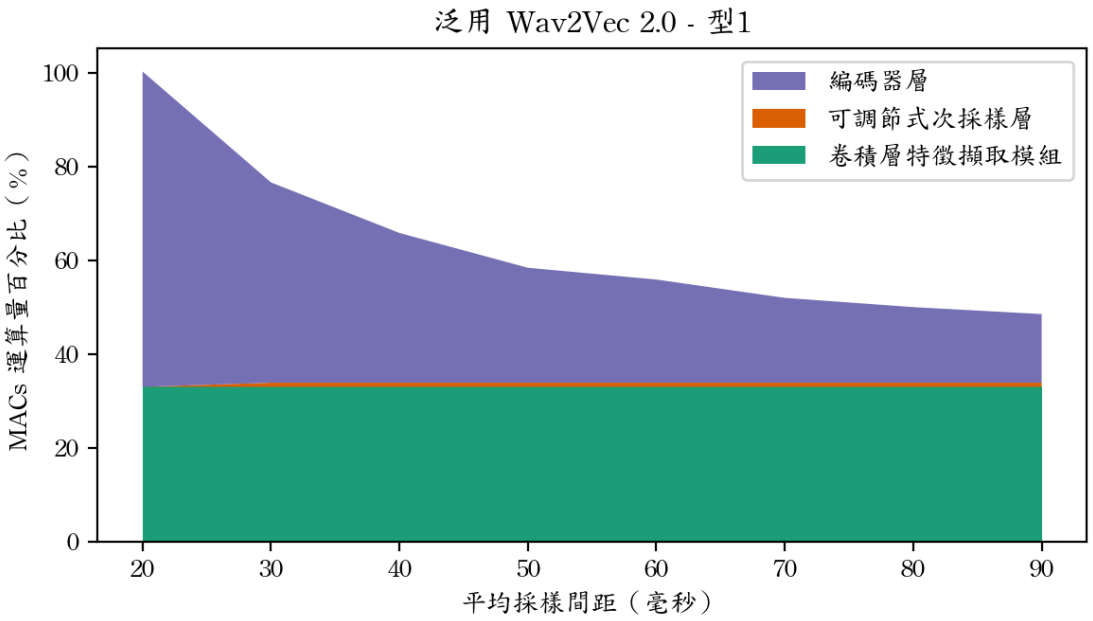


圖 3.22: 以型 1 機率分布預訓練之各項任務泛用序列壓縮 Wav2Vec 2.0，所需之乘積累加運算量（MACs）對平均採樣間距關係圖。運算量以相對與原始 Wav2Vec 2.0 運算量所占百分比呈現，結果中一併呈現三個模組分別所占的比例。

3.3.5 運算成本分析

本小節使用和第3.2.5節一樣的方式分析本節所使用的各項任務泛用序列壓縮 Wav2Vec 2.0（以型 1 機率分布預訓練）所需要之乘積累加運算量（MACs），結果如圖3.22所示，以下分別根據不同的模組進行討論：

- 卷積層特徵擷取模組：在大型對比式預訓練模型中，卷積層特徵擷取模組所占之運算量相對較低，占了未壓縮模型的 33 個百分點的運算量。

- 可調節式次採樣層：可調節式次採樣層中的連續整合發放機制所造成的額外運算量占總體壓縮量極低，僅占了未壓縮模型的 0.9 個百分點的運算量。

- 編碼器層：編碼器層所占的壓縮量較大，在未壓縮的情況下占了原始模型 76 個百分點的運算量，然而隨著序列壓縮率增加，編碼器層所需的運算量明顯下降。

doi:10.6342/NTU202304502

總結來說，大型預訓練模型中編碼器層占總體運算量比例大，序列壓縮所帶來的運算量降低顯著，同時，可調節式次採樣層所造成的額外運算量占比幾乎可忽略，相較於原始未壓縮模型（20 毫秒採樣間距），壓縮至 90 毫秒平均採樣間距的序列減少了 51.4 個百分點的乘積累加運算量。

3.4 本章結論

本章的實驗將所提出的各項任務泛用序列壓縮法應用於小型知識蒸餾模型及大型對比式預訓練模型中，實驗所得之結果可以總結為以下幾點：

- 使用各項任務泛用序列壓縮預訓練的模型，可以和單一序列壓縮率模型在同樣的序列壓縮率下達到接近的結果。
- 總體來說，預訓練使用機率分佈範圍較窄的模型表現相對較好，但是所實驗的三種機率分佈型態的預訓練模型之間差異並不大。
- 使用鏈接式時序分類的任務會根據所使用的輸出字符單元有嚴格的序列壓縮限制，在接近輸出字符序列長度的時候表現會大幅下降；序列至序列任務對於序列壓縮率沒有嚴格的限制，但序列長度對於表現的影響明顯；序列層級合計任務對於序列壓縮最不敏感，然而不同類型的序列層級合計任務因為序列壓縮造成的表現下降幅度不同；序列層級比對任務對於知識蒸餾模型的影響相對較小，然而當序列壓縮大於特定程度後也會開始影響其表現。
- 總體來說，知識蒸餾模型在序列層級合計的任務上，表現均接近或好過於對比式預訓練模型的結果，因為在這一類的任務中，知識蒸餾模型已經可以達到和對比式預訓練模型接近的結果，在經過序列壓縮之後兩者的表現差異不大；在序列層級比對的任務上，小型知識蒸餾模型的原始表現優於大型對比

doi:10.6342/NTU202304502

式模型，在經過序列壓縮之後的表現亦然是同樣的趨勢；而在序列至序列及鏈接式時序分類的任務中，大型對比式預訓練模型相較於小型知識蒸餾模型有明顯的表現優勢。

- 在運算成本上，序列壓縮法能夠有效的減少編碼器層中的運算量，在考量可調節式次採樣層中連續整合發放模組所造成額外運算量的情況下，仍然可以帶來明顯的運算量降幅。

doi:10.6342/NTU202304502

第四章　自適應序列壓縮率之探討

4.1　自適應序列壓縮率法

4.1.1　簡介

本章要探討的是各項任務泛用序列壓縮法的另一種應用情境，在第三章的實驗中，進行下游任務訓練之前預訓練模型的序列壓縮率會被選定（也就是 λ 值會根據網格搜尋採樣的序列壓縮率被事先決定，並在驗證下游任務的過程中均保持不變），本章實驗探討在下游任務訓練中，讓下游模型參數和各項任務泛用序列壓縮模型的壓縮率一併優化的情境。

第三章的實驗結果中發現不同的語音下游任務對於序列壓縮有不同的反應，舉例來說，在音素辨識中，未壓縮的模型並沒有最好的表現，反而是在接近 2 倍壓縮（40 毫秒平均採樣間距）的時候有最佳的表現。在第三章的實驗設定中，如果要找出針對不同任務的最佳壓縮率，使用網格搜尋（grid search）多次驗證下游任務是必要的，然而網格搜尋的結果雖然精確，但是所需要的運算量會隨著網格搜尋的次數增加而上升，本章的研究動機為：如果讓語音下游任務一併優化序列壓縮率，是否能夠找到對應任務的最佳壓縮率？進而避免網格搜尋所造成的額外運算量。

doi:10.6342/NTU202304502

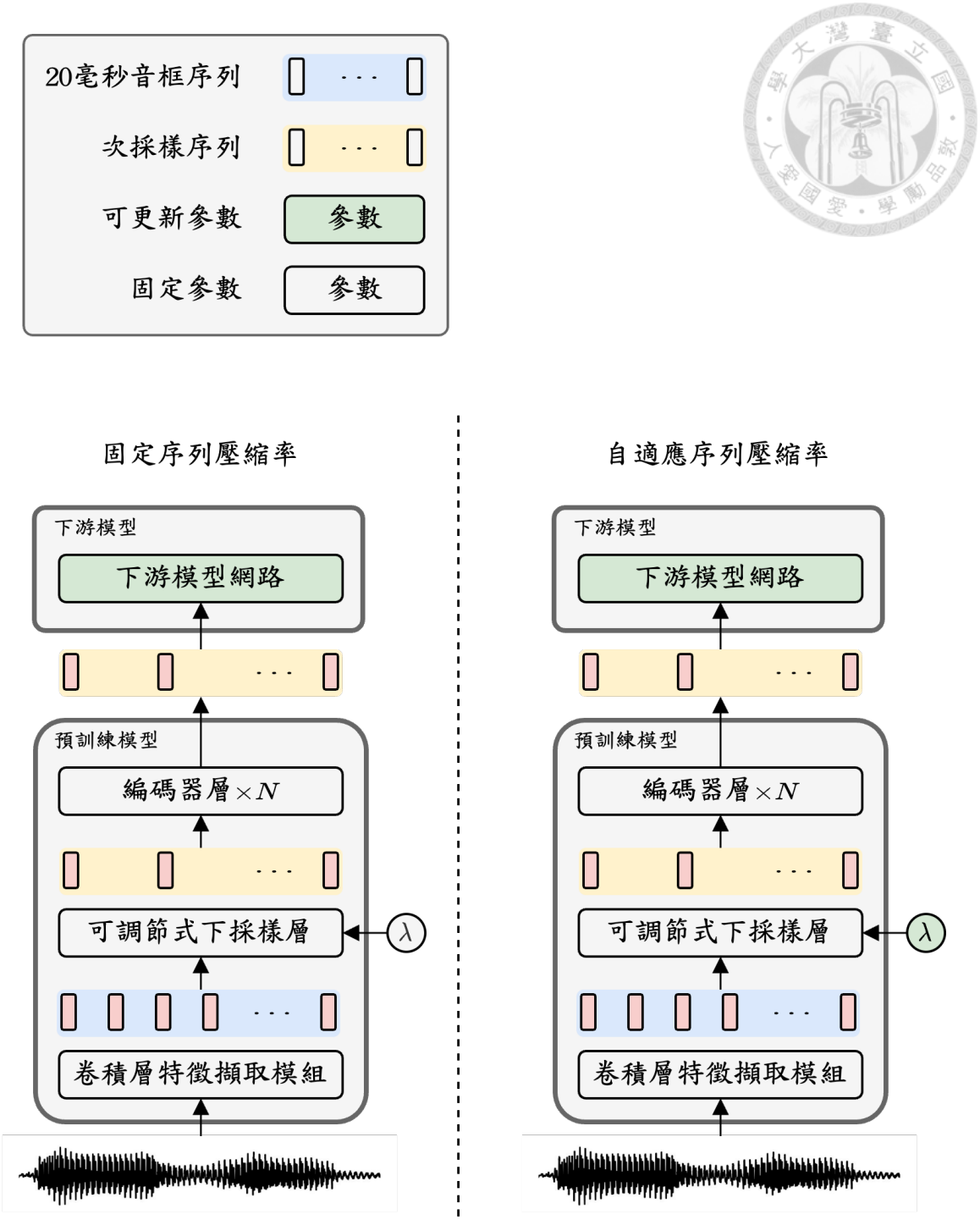


圖 4.1: 固定序列壓縮率及自適應序列壓縮率，下游模型串接預訓練模型架構和可更新參數示意圖。

4.1.2 同步最佳化下游任務以及序列壓縮率

在 SUPERB 基準中，下游任務訓練方法為將預訓練模型固定住僅更新下游任務模型網路中的參數，而用來調整序列壓縮率的參數 λ 是屬於預訓練模型的參數

doi:10.6342/NTU202304502

之一，在預設的下游任務驗證過程中，參數 λ 初始化之後即保持固定，也就是第三章中下游任務驗證方式，如圖4.1左半所示。為了讓下游模型有調整預訓練模型序列壓縮率的能力，自適應壓縮率的下游任務訓練方法為，將各項任務泛用預訓練模型中的參數 λ 視為下游任務可更新的參數之一，如圖4.1右半所示。為了避免不相稱（mismatch）的發生，一個 S 函數（Sigmoid function）被用來限制變數 λ 的範圍，使其最大可以變動的範圍和預訓練階段所使用的機率分佈範圍相同。舉例來說，對於使用型 2（$\lambda \in [0, 1.5]$）機率分佈的預訓練模型，首先引入一個實數域的參數 $p \in \mathbb{R}$，而 λ 則表示為，

$$\lambda = 1.5 \times \text{Sigmoid}(p) \tag{4.1}$$

，為了公平比較，除了多加入的參數 λ，其餘各項下游模型訓練所使用的參數、學習率、訓練步數皆和原始 SUPERB 基準所規範的一樣。

4.1.3 連續性與可微分性之分析

為了驗證優化參數 λ 的可行性，本小節探討參數 λ 對於模型輸出的連續性與可微分性。由於連續整合發放模組、自監督式預訓練模型、下游任務模型均是連續且可微分的，本小節著重探討權重改動模組中調整序列長度的函數（式3.4中的函數 F）之連續性及可微分性。函數 F 和變數 λ 的關係如圖3.5中所示，函數 F 在值域的範圍內為連續函數，在可微分性的部分，函數 F 為分段可微分函數，在情境一、情境二、情境三內部的範圍可微分，然而在情境交接處的轉折點為不可微分。

此種情況跟深層模型常見使用的激活函數：ReLU 函數（rectified linear unit, ReLU）非常相似，如圖4.2所示，ReLU 函數同樣為連續且分段可微分函數，僅在

doi:10.6342/NTU202304502

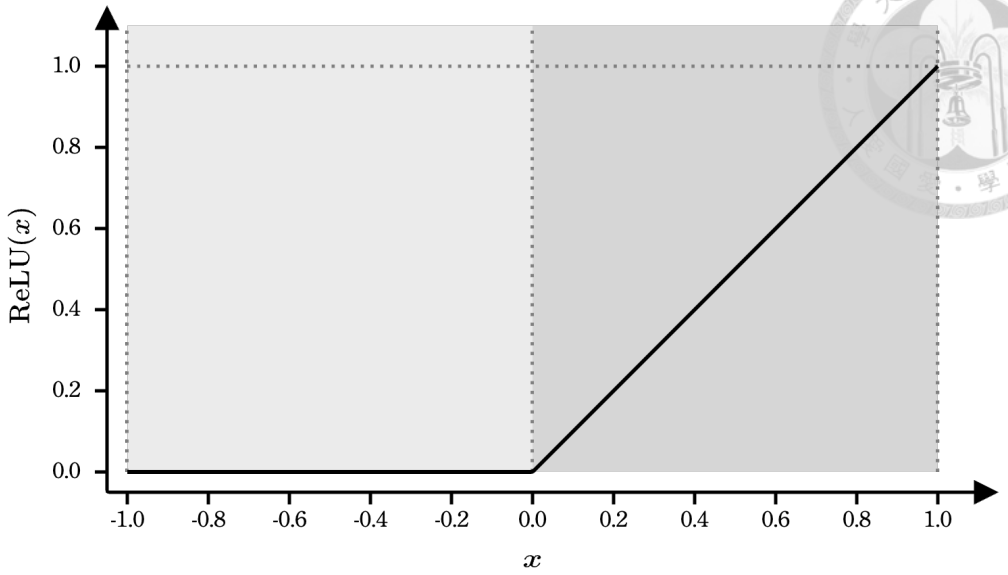


圖 4.2: ReLU 函數示意圖。

$x = 0$ 處微分不存在，在常用的深層學習套件 Pytorch 中，處理這的問題的方式為，直接指定 ReLU 函數在 $x = 0$ 的位置的微分值。在本章實驗中，以同樣的方式處理函數 F 在轉折處的微分值：指定邊界位置的微分值為從右邊逼近的微分值，如此一來，函數 F 即可視為在整段值域範圍內連續且可微分。

4.2 實驗設定及結果分析

4.2.1 模型參數設置與初始壓縮率設定

本節的實驗選用第3.2節中，以不同機率分佈預訓練在知識蒸餾模型上的三個各項任務泛用預訓練模型：各項任務泛用 DistilHuBERT - 型 1、型 2、型 3，分別以自適應壓縮率的方式驗證在音素辨識、填空任務、意圖辨識上。

在 SUPERB 基準中，個別下游任務使用獨立的最佳化設定，不同任務收斂速度不相同，因此本論文在優化 λ 的時候是使用獨立的優化器（optimizer）專門用來更新參數 λ，所使用的優化器為隨機梯度下降法優化器（Stochastic Gradient

doi:10.6342/NTU202304502

Decent optimizer, SGD optimizer），優化器的動量（momentum）設定為 0.9，音素辨識、填空任務、意圖辨識的學習率（learning rate）分別設定為 $1e-3$、$1e-2$、$1e-2$，其餘下游任務的模型參數更新方法則使用和 SUPERB 基準一致的優化器、學習率。為了探討不同初始序列壓縮率對於最終結果的影響，在這三個任務中，首先，從各自預訓練模型的機率分佈（圖3.7中的型 1、型 2、型 3 機率分佈）對應的壓縮率範圍中，隨機選取三個不同的初始壓縮率，接著以自適應序列壓縮率的方式同步優化預訓練模型的序列壓縮率及下游任務模型參數。

4.2.2 實驗結果與分析

首先，作為對照實驗，對於所選取的初始壓縮率先以固定壓縮率的方式驗證，其結果如圖4.3左半所示，模型最終的壓縮率和初始壓縮率相同，其任務驗證結果和以虛線代表的網格搜尋結果一致，而使用自適應序列壓縮率的驗證結果如圖4.3右半所示，結果以模型最終收斂的序列壓縮率及對應的評斷指標呈現，以下分析自適應序列壓縮率的結果。

在音素辨識中，由網格搜尋的結果可以發現在 30 到 40 毫秒平均採樣間距的序列壓縮模型表現最好，而在使用自適應壓縮率的情況下，原始分佈較廣的三個初始壓縮率在訓練結束時，均收斂到了 30 到 40 毫秒平均採樣間距的範圍內，最終音素辨識錯誤率的結果均優於使用固定壓縮率的實驗結果，也相當接近由網格搜尋出來的最佳結果。在填空任務中，由網格搜尋的結果可以發現，20 毫秒的平均採樣間距表現最好，自適應序列壓縮率的最終結果也和經由網格搜尋的結果一致。在意圖辨識中，網格搜尋出最好的表現在接近 100 毫秒採樣間距，且模型在 20 毫秒到 100 毫秒平均採樣間距範圍內表現差異不大，而自適應序列壓縮訓練收斂的壓縮率接近 20 毫秒採樣間距，雖然和網格搜尋的最佳壓縮率有一段差距，但是最終任務正確率表現相當接近。總結來說，自適應壓縮率在多個模型以及初始

doi:10.6342/NTU202304502

壓縮率的條件下均可以找到接近最佳解的序列壓縮率以及驗證結果。

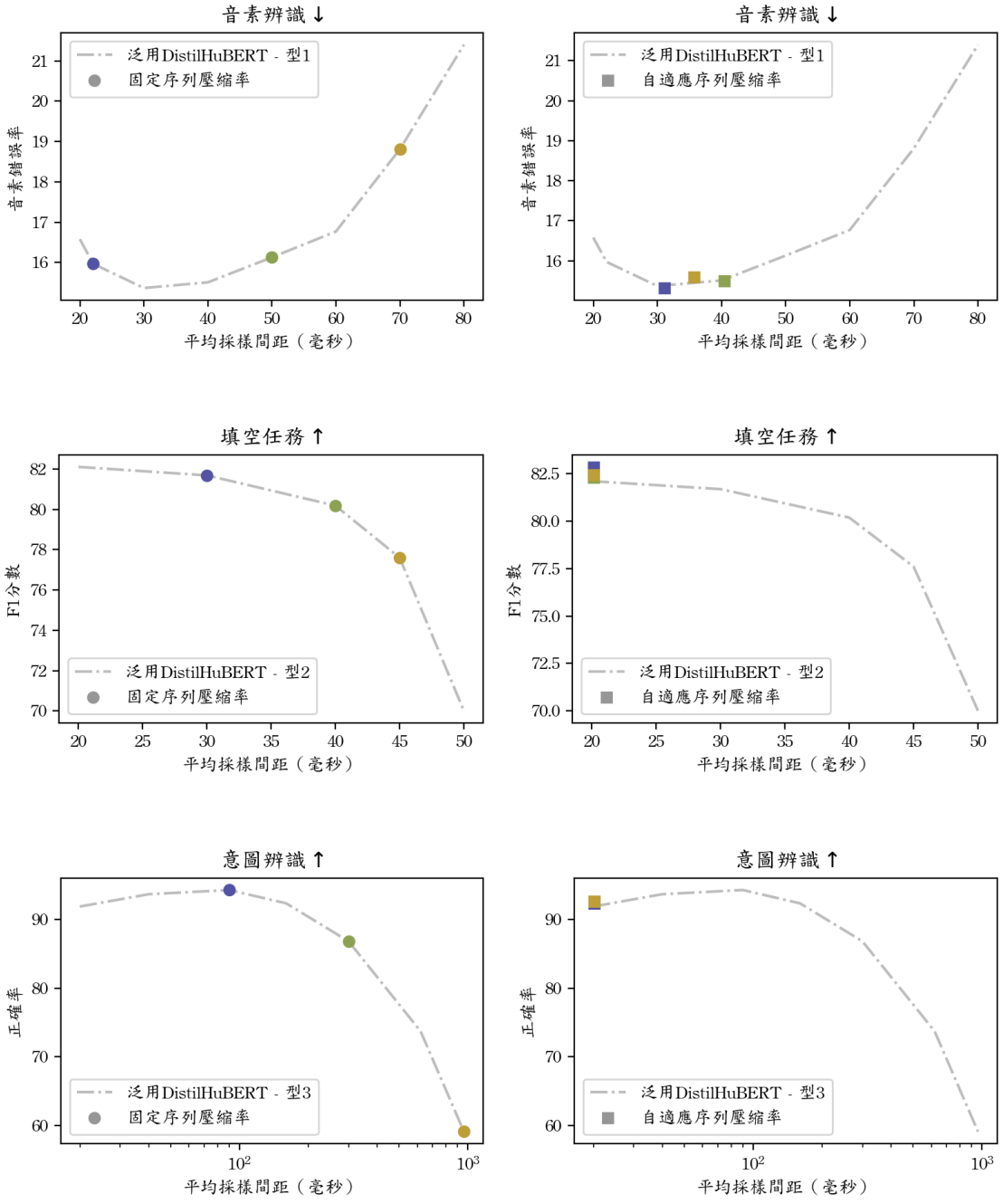


圖 4.3: 自適應序列壓縮率下游任務驗證結果，其中灰色的虛線為網格搜尋的驗證結果，左半邊圓形的資料點為使用固定序列壓縮率的結果，而右半邊方形的資料點為自適應序列壓縮率在訓練終點時收斂壓縮率及下游任務表現的結果。其中在同樣的下游任務中，相同的顏色代表相同的初始序列壓縮率。

doi:10.6342/NTU202304502

4.3 本章結論

本章探討各項任務泛用序列壓縮模型的另外一種應用，在前作 [13, 24] 使用序列壓縮法的時候，多是先選定其序列壓縮率，接著進行驗證任務，在支援多個序列壓縮率的情況下也是採樣多個序列壓縮率以網格搜尋的方式進行驗證，本章實驗讓下游模型一同調節預訓練模型序列壓縮率，在其中三個下游任務中，驗證了自適應調節壓縮率的可行性。自適應壓縮率法的優勢為：在不需要網格搜尋大量運算量的情況下，即能達到接近個別任務的最佳序列壓縮率結果。

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

第五章　結論與展望

5.1　研究貢獻與討論

本論文主要的貢獻為將各項任務泛用預訓練方式結合自監督式語音模型的序列壓縮法，前作 [23] 將各項任務泛用技術融合進語音自監督式學習是以沿用影像中常使用的子網路選取模式，然而自監督式語音模型中另外一個影響運算量的重要變因爲序列的長度，而自監督式語音模型的設計初衷是讓不同的下游任務均能使用同一個預訓練模型，加上不同下游任務對於壓縮序列長度的影響差異甚大，將各項任務泛用技術加入自監督式語音模型的序列壓縮中，對於自監督式語音模型的泛用性有重大的影響。

本論文第三章將各項任務泛用序列壓縮法應用於兩個常見的自監督式語音模型中：一個小型的知識蒸餾模型 DistilHuBERT 及大型的對比式預訓練模型 Wav2Vec 2.0，結果顯示在不論是大型、小型或是所使用的自監督式損失函數，在所實驗的兩個模型中，可以觀察到此技術對於不同自監督式模型的大小、損失函數的泛用性。再者，當所涵蓋的壓縮率範圍較小的情況下，使用此技術的預訓練模型能夠在原始未壓縮的情況下和未使用序列壓縮的單一序列壓縮率模型達到接近的結果，同時能夠在對於序列壓縮不敏感的應用中，動態的調整預訓練模型的序列壓縮率。總結來說，此技術能夠以成本極小的方式，融入現有的自監督式語音模型，並在各個不同的壓縮率（包含原始未壓縮的情況）和由同一個架構及自

doi:10.6342/NTU202304502

監督式損失模型訓練的結果相比擬，同時在部署時能根據下游任務特性或是裝置端預算限制，以更有彈性的方式調整序列壓縮率。

本論文第四章探討在序列壓縮領域未被探討的驗證方式：自適應序列壓縮率，以往，特定下游任務對於序列壓縮的反應，需要不同序列壓縮率的模型以多次驗證下游任務的方式搜尋出最佳之序列壓縮率，然而這種以網格搜尋多次驗證下游任務的方式同樣會造成運算成本的上升，同時在語音技術快速發展的情況下，不同的下游任務不斷的推陳出新，如果對於新的任務或應用在新領域的情況下，每次都進行網格搜尋的運算成本相當可觀，在第四章中提出了自適應序列壓縮率的方法，讓各項任務泛用序列壓縮預訓練模型能夠和下游任務同步優化壓縮率，在挑選的驗證任務中，自適應序列壓縮率均達到了和以網格搜尋到的最佳壓縮率的下游任務表現相近，驗證了連續可用序列壓縮率預訓練模型的另一優勢：以自適應序列壓縮法，在優化下游任務的同時，選取最佳的預訓練模型序列壓縮率。

5.2 未來展望

本論文提出用於自監督式語音模型的各項任務泛用序列壓縮法，經過本文分析之後有以下兩點未來的發展方向：

1. 隨著自監督式語音模型及語音下游任務日新月異，未來可以將各項任務泛用序列壓縮法應用於更多最先進的自監督式語音模型中，並將其結果應用在更多語音下游任務中。

2. 進一步改善序列壓縮法的效率，舉例來說，選用品質更好的分割來引導自監督式模型，或是引入額外的損失函數讓自監督式模型自行學習分割，以進一步提升序列壓縮法的表現運算量權衡曲線。

doi:10.6342/NTU202304502

總結而言，本論文提出了一個將各項任務泛用性融入自監督式語音模型序列壓縮法的架構，並驗證了其可行性及對於不同下游任務所帶來的影響，對於此技術的更多應用及進一步的改進其表現仍是有潛力可以繼續探索的方向。

doi:10.6342/NTU202304502

doi:10.6342/NTU202304502

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. In Advances in Neural Information Processing Systems, 2020.

[2] H. Cai, C. Gan, T. Wang, Z. Zhang, et al. Once-for-all: Train one network and specialize it for efficient deployment. In ICLR, 2020.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In ICASSP, pages 4960–4964. IEEE, 2016.

[4] H.-J. Chang, S.-w. Yang, and H.-y. Lee. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. In ICASSP, pages 7087–7091. IEEE, 2022.

[5] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee, et al. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 228–235. IEEE, 2021.

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech

doi:10.6342/NTU202304502

processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518, 2022.

[7] L. Dong and B. Xu. CIF: Continuous integrate-and-fire for end-to-end speech recognition. In ICASSP, pages 6079–6083. IEEE, 2020.

[8] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015.

[9] 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.

[10] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass. Parp: Prune, adjust and re-prune for self-supervised speech recognition. Advances in Neural Information Processing Systems, 34:21256–21272, 2021.

[11] Y. Lee, K. Jang, J. Goo, Y. Jung, and H. R. Kim. FitHuBERT: Going thinner and deeper for knowledge distillation of speech self-supervised models. In Interspeech, pages 3588–3592, 2022.

[12] A. H. Liu, W.-N. Hsu, M. Auli, and A. Baevski. Towards end-to-end unsupervised speech recognition. arXiv preprint arXiv:2204.02492, 2022.

[13] Y. Meng, H.-J. Chen, J. Shi, S. Watanabe, P. Garcia, H.-y. Lee, and H. Tang. On compressing sequences for self-supervised speech models. In 2022 IEEE Spoken Language Technology Workshop (SLT), pages 1128–1135, 2023.

doi:10.6342/NTU202304502

[14] Y. Miao, M. Gowayyed, and F. Metze. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In ASRU, pages 167–174. IEEE, 2015.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In ICASSP, pages 5206–5210. IEEE, 2015.

[16] A. Pasad, J.-C. Chou, and K. Livescu. Layer-wise analysis of a self-supervised speech representation model. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 914–921. IEEE, 2021.

[17] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386, 1958.

[18] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers: A survey. ACM Computing Surveys (CSUR), 2020.

[19] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, et al. SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In ACL, 2021.

[20] V. Vanhoucke, M. Devin, and G. Heigold. Multiframe deep neural networks for acoustic modeling. In ICASSP, pages 7582–7585. IEEE, 2013.

[21] 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.

[22] A. Vyas, W.-N. Hsu, M. Auli, and A. Baevski. On-demand compute reduction with stochastic wav2vec 2.0. In Interspeech, pages 3048–3052, 2022.

doi:10.6342/NTU202304502

[23] R. Wang, Q. Bai, J. Ao, L. Zhou, et al. LightHuBERT: Lightweight and configurable speech representation learning with once-for-all hidden-unit BERT. In Interspeech, pages 1686–1690, 2022.

[24] F. Wu, K. Kim, J. Pan, K. J. Han, K. Q. Weinberger, and Y. Artzi. Performance-efficiency trade-offs in unsupervised pre-training for speech recognition. In ICASSP, pages 7667–7671. IEEE, 2022.

[25] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, et al. SUPERB: Speech processing universal performance benchmark. In Interspeech, pages 1194–1198, 2021.

[26] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, et al. Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33:17283–17297, 2020.

doi:10.6342/NTU202304502