

摘要

我們這個研究是嘗試探究離散單元與音位之間的關係。

目錄

中文摘要	i
一、導論	1
1.1 研究動機	1
1.2 研究方向	4
1.3 主要貢獻	5
1.4 章節安排	5
二、背景知識	7
2.1 深層類神經網路	7
2.1.1 簡介	7
2.1.2 卷積式類神經網路	11
2.1.3 遷迴式類神經網路與序列至序列模型	12
2.1.4 專注機制與轉換器類神經網路	13
2.2 表徵與自監督式學習	17
2.2.1 特徵抽取與表徵學習	17
2.2.2 自監督學習	18
2.2.3 向量量化與離散單元	20
2.2.4 無文字 (Textless) 架構	21
2.3 本章節總結	22
三、單一語音離散表徵與音位的關係	23
3.1 相關研究	23
3.1.1 無文字與離散語音表徵	23
3.1.2 語音學分析	24
3.2 衡量指標	24
3.2.1 純度	26
3.2.2 熵和相互資訊	27
3.3 語音學的音位分類 (Phoneme Type)	28
3.4 實驗集與分析模型	30
3.5 分析方式	33
3.6 分析結果	36
3.6.1 綜觀分析	36
3.6.2 以離散單元角度切入	36
3.6.3 以音位角度切入	47
3.6.4 整體熱圖驗證	50
3.7 本章總結	54
四、多個語音離散表徵與音位的關係	55
4.1 動機	55
4.2 相關研究	55

4.3 文字處理中的分詞演算法	56
4.3.1 常見演算法	57
4.3.2 「句片段 (SentencePiece)」套件	57
4.4 分析方式	58
4.5 分析結果	58
4.5.1 由聲學片段角度探討	59
4.5.2 由音位角度探討	71
4.5.3 分析結論	74
4.6 本章總結	74
五、結論與展望	76
5.1 研究貢獻與討論	76
5.2 未來展望	77
參考文獻	78

第一章 導論

1.1 研究動機

語言是人與人彼此交流最主要的橋樑，而人們互相溝通最自然的方式便是透過說話的語音（Speech）達成。人類往往是自幼就牙牙學語開始說話，直到已屆學齡左右才開始學習認字與書寫。雖然在這個資訊爆炸的時代，人們已經習慣以文字呈現的語言作為獲取資訊的主要媒介，但不論如何，任何書寫系統所承載的語言必定有對應的語音形式。更何況世界上現存大約七千多種[1]語言中，絕大多數不見得存在成熟且普及的文字系統，卻無礙於這些語言被人們所熟悉和使用。因此，「語音」作為語言不可或缺的存在方式，了解它和研究它的價值和重要性不言而喻。

然而，相對於穩定、易於處理和保存的文字文本，語音訊號不但是變化萬千，而且蘊藏了大量從語者風格、表達內容到抑揚頓挫（韻律，Prosody）等不同層次的訊息，使得對它的處理、研究相比之下，其複雜度與難度高得多。由於語音的這種特性，過往對於語言最有興趣的語言學家們，即便明白語音作為多數語言主體的事實，也不得不藉文字符號為依託來進行探索。進入資訊化時代後，藉助電腦硬體等計算設備的幫助，從語料庫、計算語言學到自然語言處理等透過科技的力量發展語言處理技術的領域，頗長一段時間也是專注於文字的處理與分析。而嘗試結合訊號處理發展的語音技術領域，當時則是透過語言學家對語言的領域知識，例如從音位（Phoneme）、構詞（Morphology）、語法（Syntax）等等用以刻劃人類語音和語言特性的概念，將之結合機器學習建立模型，開發技術以方便人們能以語音這種更靈活的媒介，讓電腦、手機等科技工具可以更接近「直接溝通」的使用方式，便利人們的日常生活。

近年來，由於圖形處理器（Graphics Processing Unit，GPU）等硬體平行運算技術的進步，深層學習（Deep Learning）快速崛起成為人工智慧的主流，有了此項機器學習的技術，模型的彈性能夠更好的萃取資料、更貼近的尋找資料背後的機制並進行預測，使得人們不再非得依賴大量費時費工的人類標註過程，進而使得利用大量語料庫發展語言技術，進一步推進語言科技發展成為可能。尤其在自監督學習（Self-supervised Learning）技術出現之後，深層學習模型可以依照人們給定的方向，更細緻的從大量未標註、較易取得的語音或文字的語料中，找出其中的語音、語法及語義等等結構，形成可以達成對人類語言作到前所未見的效能的基石模型（Foundation Model），是這個領域的一大里程碑。尤其在以處理文字為主體的自然語言處理領域，甚至出現了幾乎使人類真偽難辨的生成式模型，改變了人們生活的方方面面。

借鏡文字方面的成功經驗，語音處理領域的研究者們也開始嘗試將語言模型（Language Model）的概念套用於變化莫測的語音訊號之上，原先人們藉助訊號處理知識一直使用的各種語音訊號特徵（Feature）也在自監督學習的架構之下，出現了許多模型從大量語音資料中得到的「語音表徵（Speech Representation）」，作為精煉語音資訊的另外一種新選擇，並開始廣泛被採用。然而，相比於文字符號的穩定與單純，語音訊號的複雜性使得它處理起來會需要更大量的資料和運算資源來擷取其中不同層次的細節；而且作為物理訊號，語音還必須處理掉環境中的雜訊等干擾。為了從紛亂的聲音中提取出最重要的訊息，向量量化（Vector Quantization）的技巧因而經常被使用在語音 [2, 3, 4] 或影像的領域中。爾後，拉氏（Lakhotia）[5]（位置不好不知 [5] 是用來說明什麼？）基於模仿人類學習語言的過程，人們藉助諸如 CPC [6]、HuBERT [7]、Wav2vec 2.0 [8] 等自監督學習模型的幫助，引入向量量化的技術，提出了「無文字（Textless）」的學習架構，轉而

以語音表徵量化後的「離散單元（Discrete Unit）」作為操作對象，企圖單純以大量的語音資料，訓練出不依賴文字的語言模型。此種學習架構的優勢，在於能保有利用大量未標註文字轉寫語音資料的同時，與連續表徵相比資訊的位元率（Bit Rate）[[[?]]][[[不通]]]利用更有效率、容易儲存、處理與傳輸，以及形式上更像文字的特性，因而可以將其視為一種「機器自己學習出來的文字 [[?]]」。接下來就可以借用長久以來只能在自然語言處理（Natural Language Processing，NLP）領域中各種語言模型的相關技術和任務的解決方法，套用在語音處理的領域中，期望可以像文字那樣從大量的語音資料中，找尋出「語音訊號版本的文字」。自此之後有一系列如應用於英語和閩南語之間的語音到語音翻譯 [9] 等等使用「離散單元」進行任務訓練的研究，一定程度的印證了這些離散單元捕捉語音內容的效果。

儘管離散單元在編碼語音之上固然有不錯的效果，並有相關研究展現了離散單元具有一定程度上與文字的相似性，然而如想將其作為「完全文字的替代者」仍然有相當的距離。借鑑過往有人[<補上?>]在自監督學習的語音表徵出來之後，便嘗試重新從語言學（Linguistics）的概念汲取靈感，對其進行語音學（Phonetics）層面的分析；本論文期望初步結合原先 HuBERT 中基於資訊理論（Information Theory）求得的統計數據，結合語音學分析的視角，對於離散表徵（Discrete Representation）本身與音位（Phoneme）和語音類別（Phoneme Type）之間的關係進行相關性的統計與分析，期望可以對 HuBERT 等自監督學習表徵進行量化（Quantization）後所得的離散單元所編碼、擷取到的資訊 [[[??]]] 是什麼有較為深入程度的了解。

1.2 研究方向

本研究論文為了探究離散單元本身是否具有潛力可以單純透過大量語音資料的自監督學習與統計過程，從文本中找尋出語音中更精細的結構，乃至於類似文字或是從語言學等人類知識領域定義出的「離散單位」—如音素（Phone）、音位（Phoneme）、字符（Character）、「詞綴與字根」（即「詞素（Morpheme）」）或單詞（Word）等等。因此，本研究取法自 HuBERT 本身為了證明其離散單元具有一定「聲學單元（Acoustic Unit）」特性的「純度（Purity）」和「相互資訊（Mutual Information，MI）」的分析數據作為分析離散語音表徵和「音位」—作為人類知識理解語音中最基礎的單位—之間相關性（Correlation）的參考。[[[???]]][[[太長句子，不知所云]]]

此外，基於訊號速率（如序列的長度）的考量，結合在文字處理中如位元組對編碼（Byte-Pair Encoding，BPE）等常見的次詞單位（Subword Unit）分詞（Tokenization）演算法，基於形式上的相似性，因而也可以套用在像是 HuBERT 離散單元這種離散的符號上，將離散單元序列中相似的類型（Pattern）發掘出來。近期如 Wav2Seq [10]、[11]、[12] 等作品也先進行了類似的嘗試。本論文則是在除了經驗上（Empirically）將其用於大量資料訓練的視角以外，也從「將其 [[什麼是”其”？]] 視為另一種離散單位」的觀點進行統計數據的量化分析（Quantitative Analysis），作為在計算資源有限的前提下決策數據編碼的一個判斷標準 [[（不知所云？）]]。

1.3 主要貢獻

有鑑於離散單元在當前語音處理領域中愈來愈廣為使用，本論文將以更細緻的方式，對離散單元與音位之間的關係進行分析與探討，比較不同離散語音表徵之間的差異，透過結合語音學的知識，給出這些離散單元之間可能關係的觀察。隨後藉助文字處理作法提供的靈感，將多個離散單元分組結合形成「次詞單位」，將語音訊號以不同的方式重新編碼，並與未分組的離散單元進行比較。透過模型表徵與標註資料提供的現象分析，嘗試初步探索模型從機器學習演算法找出的離散單元和人們應用知識提供的標註，兩者之間的相似程度高低。並以此對這些離散單元的動機——為語音提供類似文字的表徵——進行驗證，為往後研究語音語言模型（Spoken Language Model）中的「對語音編碼」提供先期參考。

1.4 章節安排

本論文將以如下的方式進行章節安排：

- 第二章：介紹後面章節所需要的與深層學習、表徵學習與自監督學習相關的基本背景知識。
- 第三章：從介紹離散單元本身提出後，「無文字」的相關前作文獻開始，帶出對從「無文字」系列作品用到的各種自監督學習模型所抽取之離散單元本身的純度（Purity）和相互資訊（Mutual Information，MI）等統計數據，進行比較與分析。
- 第四章：探討為何單一離散單元本身不是以發掘出類似音位的單位，並進而對應到文字，以及近年人們嘗試以離散單元為基礎，透過分詞演算法發展之「聲學片段（Acoustic Piece）[11]」的進展，接著我們在將單元進行分詞法重

新編碼處理前後，觀察數據上與第三章結果間的差異，以討論對離散單元進行分詞是否可以找出更接近音位的單位，以及「離散單元可否被文字化」或「離散單元學到的是否為更精細的語音訊號規律或結構」等疑問。

- 第五章：總結前面的觀察結果，並進一步探討本研究還可以如何延伸，並怎麼幫助語音語言模型的發展。

第二章 背景知識

2.1 深層類神經網路

2.1.1 簡介

深層類神經網路（Deep Neural Network，DNN）是由神經科學家麥氏（McCulloch）與皮氏（Pitts）於 1943 年提出 [13] 的計算模型，靈感取自連結主義（Connectionism）的核心主張——以模仿生物神經網路的連結方式模擬複雜的心智活動。

為模擬神經細胞處理訊號的過程，深層類神經網路最基本的單位稱為「神經元（Neuron）」，其本質為線性分類器。每個神經元接收的輸入數值 $x = (x_1, x_2, \dots, x_N)$ 是一個 N 維向量，每一維會被賦予一個權重（Weight） $w = (w_1, w_2, \dots, w_N)$ ，加權後總和再加上偏差值（Bias） b ，得到線性輸出值。為了模擬神經細胞的觸發過程，該分類器常被加上非線性的激發函數（Activation Function） σ 的轉換，才得到最終輸出值 y 。如圖 2.1 所示，神經元的運算規則以下列數學式描述：

$$y = \sigma(w^T x + b) \quad (2.1)$$

常見的激發函數包含線性整流單元（Rectified Linear Unit，ReLU）、S 函數（Sigmoid Function）或雙曲正切函數（Hyperbolic Tangent Function，tanh）等等。

結合數個神經元的運算，羅氏（Rosenblatt）[14] 於 1958 年提出感知器（Perceptron）模型。根據通用近似定理（Universal Approximation Theorem）[15]，感知器理論上可逼近任意函數。然而，後續研究發現單層的感知器具有如「線性

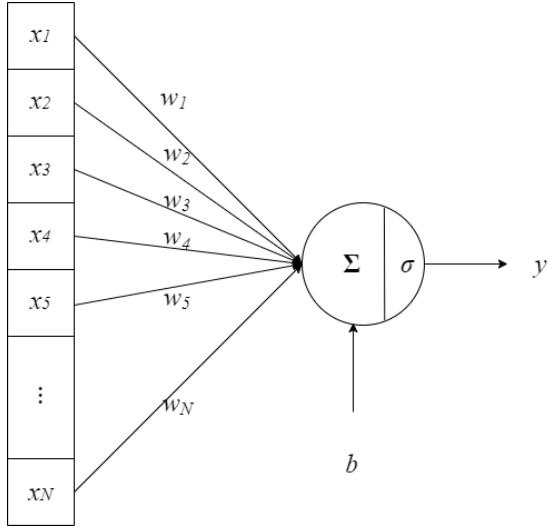


圖 2.1: 神經元示意圖

不可分」¹ 等先天限制，使其曾經一度不被看好。

為了突破該缺陷，人們嘗試在輸入與輸出層之間增加「隱藏層（Hidden Layer）」，成為「多層感知器（Multilayer Perceptron，MLP）」，如圖 2.2 所示。藉助隱藏層的幫助，多層感知器可對輸入進行多次非線性轉換，大大拓展了模型的適用範圍。此模型是透過「加深隱藏層」得來，現今為人們熟知的「深層類神經網路」即由此得名。

藉助深層類神經網路的彈性，我們可以透過大量訓練資料來訓練模型，藉此逼近應用任務中欲近似的函數 f ，該函數蘊藏在資料集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ 中，其中每個資料點 (x_i, y_i) 為輸入與輸出間的配對，即對於 N 個資料點都有

$$y_i = f(x_i) \quad \forall i \in \{1, \dots, N\} \quad (2.2)$$

之關係。為了使這個函數更加逼近目標函數 f ，類神經網路會構建一個逼近中的函數 $f_{\theta_t}(\cdot)$ 。透過不停的迭代，模型對資料集 \mathcal{D} 的每一筆資料 x 紿出預測 $f_{\theta_t}(x)$ 。透過某個減損函數（Loss Function） \mathcal{L} 計算出誤差（Error），此誤差對參數 θ_t 求

¹ 例如無法貼合異或（Exclusive OR，XOR）運算等函數

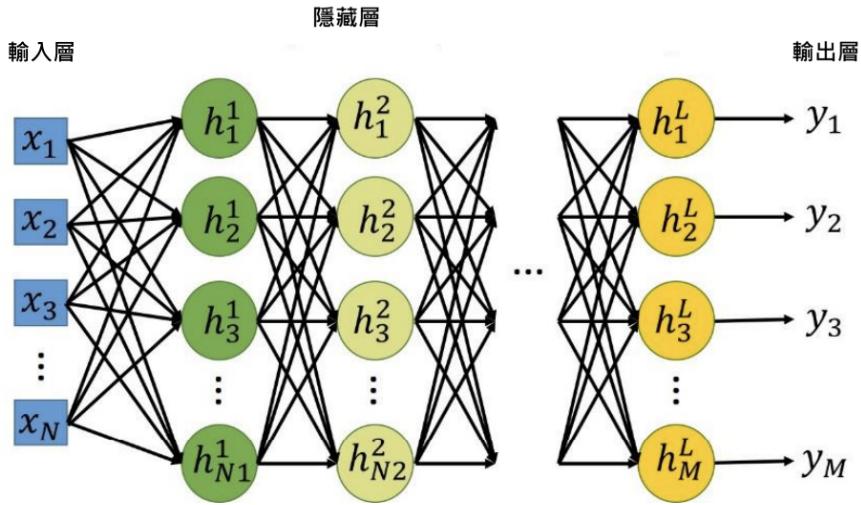


圖 2.2: 多層感知器／深層類神經網路示意圖

出梯度 (Gradient) 後將指示模型更新的方向，以此乘上學習率 (Learning Rate) η 後從參數 θ_t 減去，便能對整個模型進行更新，使之更有機會接近目標函數 f 。由於此過程是依照梯度使得函數 \mathcal{L} 逐步降低，以此獲名「梯度下降法 (Gradient Descent)」，其公式如下：

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta} \mathcal{L}(\mathcal{D}, f_{\theta_t}(\cdot)) \quad (2.3)$$

其中， t 為當前的迭代數， θ_t 為當前模型參數， θ_{t+1} 為更新後的模型參數。

在此模型更新的過程中，減損函數承擔著指引模型逼近的角色，因此根據應用的任務不同，常見的減損函數包括

- 均方誤差 (Mean Squared Error, MSE)：一般用於迴歸 (Regression) 問題，直接計算兩數值之間的差距的平方和

$$\mathcal{L}_{\text{MSE}}(y_i, \hat{y}_i) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.4)$$

式 (2.4) 中的 \hat{y}_i 是模型預測的輸出值，理想上希望愈接近資料標註 y_i 愈好。

- 交叉熵 (Cross-entropy, CE)：一般用於分類 (Classification) 問題，著重計算兩個機率分佈之間的差異

$$\mathcal{L}_{\text{CE}}(y_i, \hat{y}_i) = - \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2.5)$$

式中的 \hat{y}_i 與 y_i 所代表的意義和式 (2.4) 相同，分別表示模型預測值與資料標註，只是計算兩者差距的方式有異。

透過上述的訓練方式可以得知，類神經網路的訓練需要相當龐大且複雜的運算過程，因此剛提出時仍舊難以應用於現實應用中。

為了提高函數貼合的效率，魯氏 (Rumelhart) [16] 與辛氏 (Hinton) [17] 等人提出了反向傳播 (Back-Propagation) 演算法，旨在將上述的更新過程，藉助鏈鎖率 (Chain Rule) 的幫助，由隱藏層逐層反向傳播至輸入層，對整個類神經網路進行修正。

反向傳播演算法的設計，正好能配合圖形處理器 (Graphics Processing Unit, GPU) 等硬體裝置的優勢，以平行運算能力加速函數貼合 (Fit) 的效率。由此開始，這種透過深層類神經網路，從大量資料集中發掘函數關係的機器學習演算法，被稱為「深層學習 (Deep Learning)」。類神經網路在各個領域的泛化能力 (Generalizability) 已經得到前所未有的效能，包含電腦視覺、語音處理和自然語言處理，因此深層學習在近年成為人工智慧發展的主流。

然而，根據資料特性的不同，並不是所有的資料都適用簡單的「輸入與輸出配對」的模式。研究者根據任務需求，發展出了不同架構的類神經網路以適應資料特性。前述最基本的深層類神經網路，由於資料是直接由輸入層，通過逐層的矩陣運算得到輸出，因此被稱之為「前饋式類神經網路 (Feed-Forward Network, FFN)」。

藉由調整各神經元之間的連接關係，發展出卷積式 (Convolutional)、遞迴式

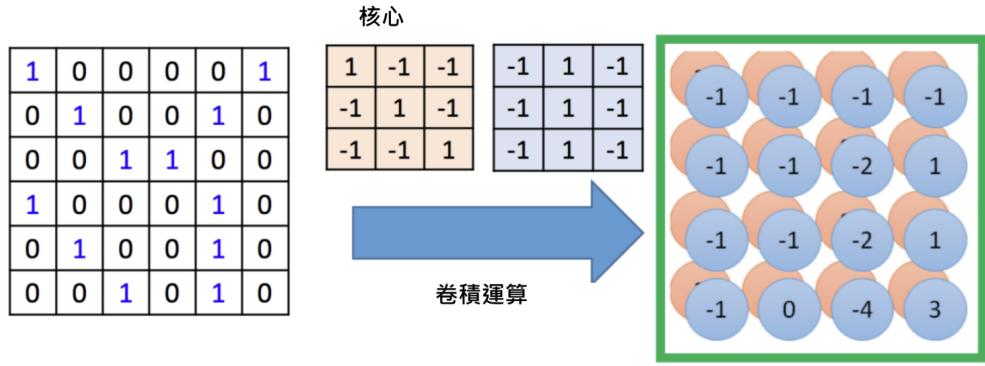


圖 2.3: 卷積式類神經網路示意圖，取自李宏毅教授的課程投影片

(Recurrent) 與轉換器 (Transformer) 類神經網路等架構變體，以適應如影像、語音和文字等不同型態的資料。這些架構在語音與文字處理被普遍使用，接下來將逐一分別介紹：

2.1.2 卷積式類神經網路

卷積式類神經網路 (Convolutional Neural Network, CNN) 為 1998 年由楊氏 (LeCun) [18] 提出，旨在以訊號處理的卷積 (Convolution) 運算，模擬生物的視覺皮質感知 [19]。

如圖 2.3 所示，卷積式類神經網路透過核心 (Kernel)，對輸入的資料 — 如圖中的二維矩陣 — 進行卷積運算，獲得該輸入的特徵圖 (Feature Map)。核心帶來的移動不變性 (Shift-invariance) 非常適用於捕捉二維影像中的局部特徵，以作為類神經網路分辨資料的依據。

有別於影像處理中，資料多以二維矩陣表示像素 (Pixel) 三原色的亮度數值，因此以二維的卷積運算為主；由於語音時常處理時間軸之上的訊號，包含聲波波形 (Waveform)、時頻譜 (Spectrogram) 或聲學特徵，因此一維的卷積式模型也時常出現，以模仿人耳聽覺對時變訊號的窗框 (Window) 的效應，進而觀察到語音

中在不同解析度（Resolution）的資訊。

2.1.3 遞迴式類神經網路與序列至序列模型

遞迴式類神經網路

遞迴式類神經網路（Recurrent Neural Network，RNN）常用於處理隨時間變化的序列資料，特別是語音與文字等等，順序資訊相當關鍵的各種語言任務。為了處理需要記憶和狀態的資料類型，遞迴式類神經網路的輸出會重新接回輸入層，使得前一個時間點（Timestep）的資料與內部狀態會繼續影響後續的時間點。常用的遞迴式類神經網路類型有長短期記憶（Long Short-term Memory，LSTM）[20] 和閘門循環單元（Gated Recurrent Unit，GRU）[21] 等。

遞迴式類神經網路通常用在處理序列至序列的應用，例如語音辨識、語音合成或機器翻譯等和語言密切相關的任務中。

序列至序列模型

由於許多語言資料通常以兩個序列互相配對的形式呈現，因此專門處理這類資料的模型被稱為「序列至序列模型（Sequence-to-sequence，Seq2seq）」[22]。此類模型的典型架構由編碼器（Encoder）和解碼器（Decoder）組成，旨在模擬輸入與輸出序列之間的變化與相依關係（Dependency）。

序列到序列模型一般有兩種模式：其一是每個時間點都生成一個輸出的向量，適用於輸入與輸出序列等長的任務，這種模式被稱為「符記分類（Token Classification）」；但更常見的情況是，輸入與輸出序列的長度並不相同。處理後者的典型作法是讓編碼器將輸入序列依據時間，一步一步輸入編碼器，將序列編碼為內部表徵（Latent Representation）。完成編碼後，編碼器將最後一個時間點的表

徵用以代表整個序列，稱為「語境向量（Context Vector）」。該向量接著被傳遞給解碼器，依序生成輸出序列。

2.1.4 專注機制與轉換器類神經網路

專注機制（Attention Mechanism）

由於遞迴式類神經網路需要處理整個序列的編碼和解碼資訊，對時間點距離較遠的輸入容易被遺忘，亦即難以處理長期相依性（Long-term Dependency）問題。為了解決這種困境，巴氏（Bahdanau）等人 [23] 提出了「專注機制」。該機制讓解碼器將輸入序列的每個訊號都視作「部分的」語境向量，由對不同時間點的向量加權合計獲得，使得在生成輸出序列時能依據當時的需求從輸入序列中提取所需的訊息。專注機制的引入，使得序列至序列模型在處理如語音辨識、機器翻譯等任務時效能大大改善。

轉換器類神經網路

儘管遞迴式類神經網路善於處理時序資料，但其難以平行化的架構限制了其在訓練和推理（Inference）時的效率。2017 年，瓦氏（Vaswani）等人 [24] 提出了完全由專注機制構成、不依賴遞迴運算的序列至序列模型，並稱之為「轉換器（Transformer）」，以解決機器翻譯等任務。

轉換器類神經網路一般包含編碼器和解碼器兩部分，均為多層架構。圖 2.4 展示完整的轉換器架構圖，以下分別介紹其主要元件：

位置編碼（Positional Encoding）

對於編碼器或解碼器的輸入序列，模型先對序列中不同位置的時間點進行編碼，取代遞迴式類神經網路逐步運算的過程，使其能在平行計算的同時考慮不同

時間點的影響。編碼的函數可依照需求變換，如原始的轉換器採用三角函數進行位置編碼，而在語音模型中，有時也會採用卷積式網路以捕捉輸入的細微資訊。

經過位置編碼後，向量會通過每一個轉換器層（Transformer Layer），進行以「多頭專注」為主的一連串運算：

多頭專注（Multi-head Attention）

轉換器層中的專注機制涉及三個輸入向量：詢向量（Query） Q 、鑰向量（Key） K 和值向量（Value） V 。專注機制運算如下：

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V \quad (2.6)$$

其中 softmax 為正規化指數函數， d_k 為鑰向量 K 的維度。這一運算首先通過鑰向量和詢向量的內積計算專注權重，而後為避免受維度過大影響而縮小為 $\sqrt{d_k}$ 分之一，最後通過正規化指數函數使得權重總和為 1，以此分配給值向量進行加權。

為應對多樣的輸入訊號，每個轉換器層具備多個獨立的專注機制，對三組輸入向量先進行各自不同的 W^Q 、 W^K 、 W^V 線性轉換，稱為「多頭專注」。對於第 i 個專注頭（Head）有

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2.7)$$

也就是每個專注頭分別透過前述式 (2.6) 的 $\text{Attention}(\cdot, \cdot, \cdot)$ 運算，使模型針對不同輸入訊號可以進行不一樣的運算處理。最後，若有 h 個專注頭，多頭專注模組會將多個頭的結果進行串接（Concatenate，即式 (2.8) 描述之 Concat），經過線性轉換 W^O 作為模組輸出，運算式表示為

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.8)$$

最終將 $\text{MultiHead}(Q, K, V)$ 作為整個多頭專注模組的輸出值。

其他層內運算

每層轉換器層在經過多頭專注運算後，會依序進行以下三個步驟：

1. 與輸入向量透過殘差連接（Residual Connection）相加，隨後進行層正規化（Layer Normalization）以穩定訓練。
2. 將此結果通過一個簡單的前饋式類神經網路對向量做線性轉換。
3. 再將前饋網路的輸入與輸出再次計算殘差總和後，進行層正規化輸出。

以上為轉換器被提出時的最原始模型，其後對殘差連接、層正規化的安排也存在各類變體。

跨專注機制（Cross-attention）

為了輸出結果，解碼器需要編碼器提供輸入序列的資訊。因此，原本在編碼器層中的自專注機制，在解碼器中會再經過一次跨專注機制的運算，使用編碼器提供的詢向量和鑰向量對解碼器的值向量進行專注運算。由於轉換器不需要對每個時間點逐一運算，使此過程能被高度平行化，類神經網路得以透過專注機制同時進行序列資料的大量訓練。這種可擴展性（Scalability）使其在自然語言和語音處理上取得了巨大的進展，幾乎取代了原先遞迴式類神經網路的應用場景，近年來甚至被應用在圖像類的資料上[25]，展現了此種模型架構的彈性與泛用性，成為目前最前沿的人工智慧主流架構。

除了模型架構，機器學習中不可或缺的另一大部分是對資料的編碼過程。如何更有效率的讓機器理解、處理和輸出資料，是機器學習乃至深層學習的一大課題。面對捉摸不定、抽象且變化萬千的人類語言，語音和文字處理中的表徵學習尤為重要。

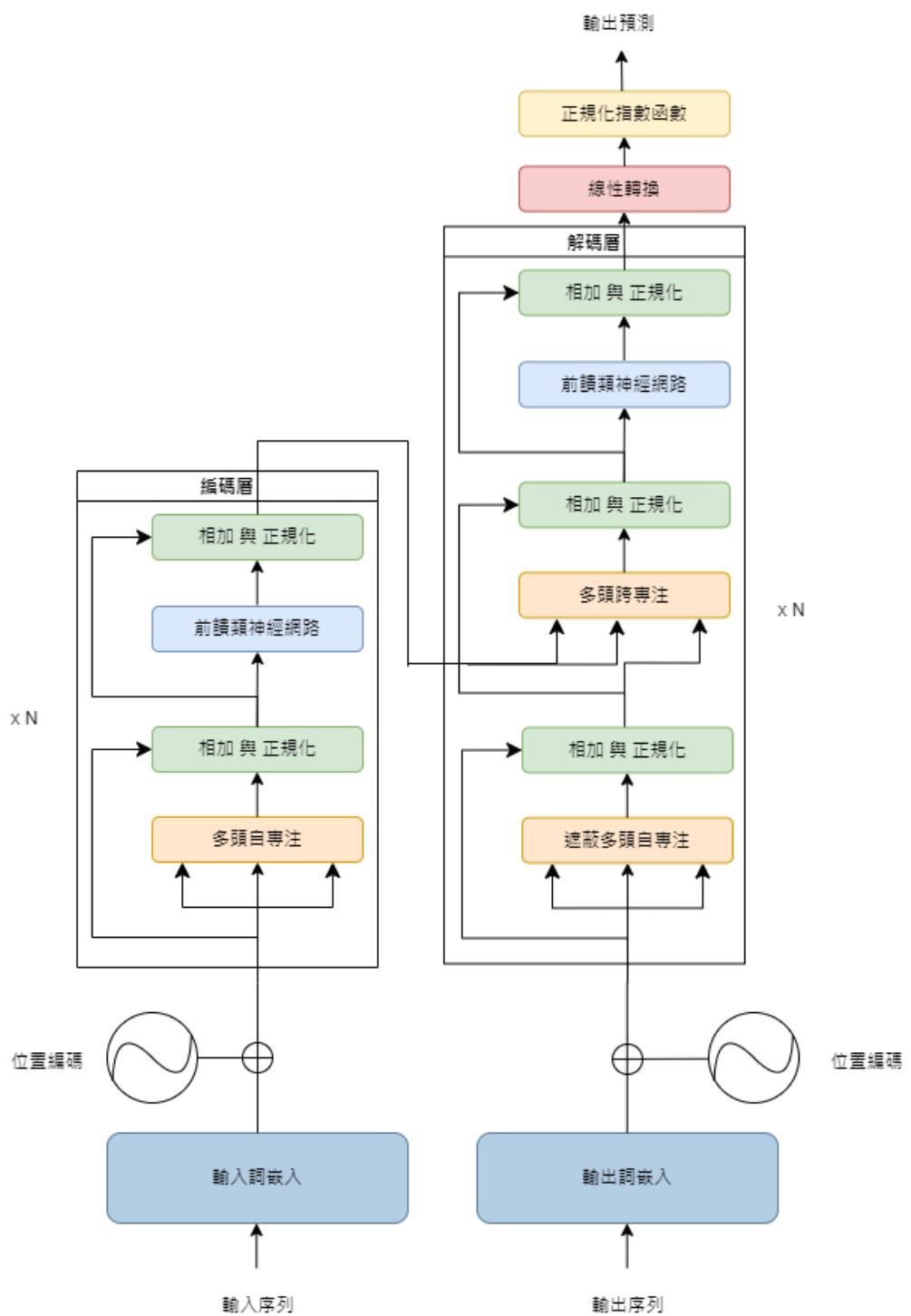


圖 2.4: 轉換器架構圖

2.2 表徵與自監督式學習

2.2.1 特徵抽取與表徵學習

不論採用何種模型，為了讓機器可以處理並捕捉輸入資料中的訊號與規律，包括如何對資料編碼和運算的步驟，在機器學習中稱之為特徵抽取（Feature Extraction）或表徵學習（Representation Learning），這是模型建構中不可或缺的重要步驟。

對於抽象的語言概念，早期工程領域根據對語音和文字的理解，分別進行了不同的處理。對於離散且可計數的文字，人們使用詞頻統計衍生出如 n 連詞 (n -gram)、TF-IDF (詞頻-倒數文件頻率，Term-Frequency Inverse Document Frequency) 等特徵作為模型學習的前處理步驟；而對於連續且複雜的語音，工程師則透過聲學原理與訊號處理的知識，使用如濾波器組 (Filter Bank)、梅爾倒頻譜係數 (Mel-Frequency Cepstrum Coefficient，MFCC) 等特徵，類比人耳捕捉語音訊號的過程。

在深層學習逐漸發展的過程中，自然語言處理領域的一大里程碑是米氏 (Mikolov) 提出的「Word2vec」模型 [26]，該模型以連續的向量表徵 (Vector Representation) 取代稀疏 (Sparse) 的統計數據，對離散的文字單詞進行「詞嵌入 (Word Embedding)」編碼。通過大量文本運算，將各單詞之間的共現 (Collocation) 以跳躍詞 (Skip-gram，SG)、連續詞袋 (Continuous Bag-of-Word，CBOW) 等演算法轉換成高維向量空間中的點，找出每個單詞最適合的語義表徵。爾後，為了更細緻地捕捉同一單詞在不同句子中的脈絡變化，ELMo (來自語言模型的詞嵌入，Embeddings From Language Model) [27] 提出了「含上下文詞嵌入 (Contextualized Embedding)」的概念，使得各單詞在運算表徵的過程中可以根據

上下文進行些微調整。

2.2.2 自監督學習

隨著轉換器模型的提出，BERT（來自轉換器的雙向編碼器表徵，Bidirectional Encoder Representations From Transformers）[28] 被提出。通過自專注機制，工程師們無需依賴人工標記，透過預先設定任務（Pretext Task）引導模型從大量文本中自行找出更細緻且考量前後文的語義關係，並在許多文字任務上獲得了優異的成績。

自此，楊氏（LeCun）將這種以特定任務作為引導、藉助資料本身的結構替代標註，從大量未標註資料中學習資料規律的訓練方式，稱之為「自監督學習（Self-supervised Learning，SSL）」。BERT 的成功使自監督學習得以大行其道，並出現了許多由巨量資料進行預訓練（Pre-train）的基石模型（Foundation Model），有效解決了語言處理領域中的標註資料稀缺的問題。人們在解決語言相關任務時，不需從頭蒐集資料與進行耗時耗能的訓練過程，而是可以利用基石模型優良的泛化（Generalization）能力，解決各種應用任務的需求。相比於預訓練的任務，這些更貼近日常現實的任務被稱為「下游任務（Downstream Task）」，能應對廣泛的下游任務種類，這是基石模型最大的優勢。

有鑑於文字處理方面的成功，語音領域的研究者嘗試將相似模式應用於語音，眾多語音基石模型隨之出現。大量的語音資料庫幫助模型萃取出有助於下游任務的語音表徵（Speech Representation），在各種任務上獲得了優於傳統聲學特徵的表現。語音表徵具備的無窮潛力，逐漸成為聲學特徵之外的新選擇。

依照這些語音自監督模型的預訓練學習模式，可大致分為重建式、預測式與對比式模型。以下分別介紹這三類模式：

重建式學習 (Reconstruction Learning)

此類模型通過對輸入訊號進行擾動 (Perturb) 後，期望模型將被更動的輸入重新預測回原始資料，通常減損函數表示為：

$$\mathcal{L}_{recon} = \mathbb{E}_x[|f_\theta(\tilde{x}) - x|] \quad (2.9)$$

其中 \tilde{x} 為擾動後的資料， $f_\theta(\cdot)$ 為模型函數。擾動方式通常以遮蔽為主，在文字處理中以 BERT 為代表，稱為「遮蔽語言模型 (Masked Language Model, MLM)」。在語音中，採用此方式學習的有 Mockingjay [29]、TERA [30] 等模型。

預測式學習 (Predictive Learning)

此類模型通過預訂一些學習目標函數，製造類似輸入與輸出的配對資料，讓模型預測該函數的結果來學習資料中的特定結構。其訓練減損函數可表示為：

$$\mathcal{L}_{pred} = \mathbb{E}_x[\text{eval}(f_\theta(x), \hat{f}(x))] \quad (2.10)$$

其中 \hat{f} 是期望模型學習的目標函數， $f_\theta(\cdot)$ 為模型函數，eval 是用來評估預測好壞的標準。

目標函數的典型代表是自迴歸 (Autoregressive)，期望模型預測未來時間點的輸入表徵。文字方面以「GPT (生成式預訓練轉換器，Generative Pretrained Transformer)」系列 [31, 32] 為代表，語音上的「自迴歸預測編碼 (Autoregressive Predictive Coding, APC)」[33] 也是採用此種模式。此外，語音基石模型還可以使用其他訓練目標，如 PASE+ [34] 預測其他模型的表徵，而本文著重探究的「HuBERT (隱藏單元 BERT，Hidden-unit BERT)」[7, 35] 則以預測分群 (Cluster) 後的輸入表徵為目標，這些預測目標又被視為虛擬標註 (Pseudo-label)，後文將著重探討。

對比式學習 (Contrastive Learning)

此學習方式的訓練目標是要求模型區分正樣本 (Positive Sample) 與負樣本 (Negative Sample) 的差異，減損函數通常定義為：

$$\mathcal{L}_{contr} = -\mathbb{E}_x \left[\log \left(\frac{\sum_{\tilde{x} \in x_{pos}} \exp(\text{sim}(x, \tilde{x}))}{\sum_{\tilde{x} \in \mathcal{X}} \exp(\text{sim}(x, \tilde{x}))} \right) \right] \quad (2.11)$$

其中 x 為輸入， x_{pos} 為正樣本， \mathcal{X} 為包含正負樣本的資料集， $\text{sim}(\cdot, \cdot)$ 是評估兩個樣本相似程度的函數，常用的相似度函數為內積運算得出的餘弦相似度 (Cosine Similarity)。語音上最早使用對比式學習的模型為「對比預測編碼 (Contrastive Predictive Coding, CPC)」[36]，之後如 Wav2vec [37]、Modified CPC [38]、Wav2vec 2.0 [39] 等模型亦是以對比正負樣本的模式訓練，但訓練時正負樣本的定義有所差異，如 Wav2vec 僅以時間維度上相同的向量為正樣本，其餘則將固定時間內的向量皆視為正樣本。

對比式學習通過正負樣本的定義，將預訓練任務形塑為分類問題，因此減損函數本質上為交叉熵，使模型能夠判斷訓練資料中的結構差異。

2.2.3 向量量化與離散單元

語音訊號雖然記錄語言資訊，卻與影像資料一樣都是連續數值資料，不像離散的文字較易處理，因此發展出了許多應用廣泛的模型。為了使語音模型訓練可以套用自然語言處理領域的演算法，從連續語音中找出離散表徵逐漸成為研究趨勢，這類研究被稱為「聲學單元發掘 (Acoustic Unit Discovery, AUD)」。

由於語言概念本質上是離散符號，向量量化技術常用於涉及語言標註的情境，如電腦視覺經典的量化向量變分自編碼器 (Vector-Quantized Variational Autoencoder, VQ-VAE) [40]，利用影像標註的離散語言單詞特性，使模型學習的表徵向量被約束在編碼簿 (Codebook) 的幾個向量中。

在語音領域，基於 Wav2vec 之上的 Vq-wav2vec [41] 和 Wav2vec 2.0 將連續的語音特徵量化加入訓練目標中，在語音辨識等任務上取得了顯著進步。

HuBERT [35] 則先對連續的 MFCC 特徵進行 K-平均（K-Means）演算法分群，以所得的群心（Centroid）或碼字（Code Word）編號作為訓練目標，實施類似 BERT 的遮蔽語言模型訓練，並改以此次訓練得到的語音表徵為目標，再次分群後實施第二次訓練。這些經過兩輪訓練後，從模型表徵分群得到的群心，被視為「隱藏單元（Hidden Unit）」，呈現了語音訊號中的代表性聲學特徵。透過找出隱藏單元的過程，HuBERT 在低資源情況下達到與 Wav2vec 2.0 相近的語音辨識成績。

2.2.4 無文字（Textless）架構

奠基於 HuBERT 等語音基石模型的成功，利用隱藏單元的概念，將大量語音資料表徵進行 K-平均演算法，作為這些語音訊號的虛擬標註。如此得到的大量離散隱藏單元形成了「虛擬文字（Pseudo-text）」的語料庫，基於這些離散單元訓練的語言模型，稱為「生成式口語語言模型（Generative Spoken Language Model，GSLM）」[5]。配合反向語音合成訓練基於離散單元的語音生成模型，整體架構完全不依賴文字標註，訓練出純語音語言模型，稱為「無文字（Textless）架構」[42]。

無文字模式在語音問答（Spoken Question Answering）[43] 和語音到語音翻譯（Speech-to-speech Translation）[9] 中取得了前所未有的進展。這些「離散單元（Discrete Unit）」被視為類似文字卻不依賴人類文字標記的語音表徵，具有儲存位元率低和可套用文字語言模型訓練模式的優勢，受到語音社群的廣泛借鑑，後續也帶出了許多如 [44] 等將語音以離散表徵編碼的研究。

雖然在系統與應用任務上取得了成功，但這些離散單元本身與文字的差異，

及其對語音語言模型訓練的幫助，仍是領域內探討的焦點。有鑑於此，本論文基於語言知識，從最接近文字且與語音訊號最相關的「音位（Phoneme）」開始探討，期望了解離散單元能帶來的特徵及其對後續應用的幫助。

2.3 本章節總結

本章節首先介紹了深層學習模型的核心部件—類神經網路的基本原理，隨後對本論文研究的核心—「語音表徵」與「離散單元」的發展與歷史進行了梳理。接下來的章節將緊扣這些基石模型得到的離散特徵，對其與「音位」這類語音學標記之間的統計關係進行更深入分析。

第三章 單一語音離散表徵與音位的關係

HuBERT [7, 35] 和 Wav2vec 2.0 [39] 等語音基石模型的成功，不僅在語音任務上達到了前所未有的表現，還促進了語音表徵離散化的發展。由此產生的「無文字（Textless）」架構 [42, 45, 5]，讓人們在處理語音訊號時，有了連續表徵以外的新選擇。離散形式的表徵可以直接應用文字領域發展的技術，如機器翻譯、生成式模型等，為語音技術帶來新的突破。另一方面，基於離散「符記（Token）」的共同形式，離散語音表徵可以更好的整合文字資料，促成多模態領域的發展。跨模態離散表徵的成功，甚至驅使影像領域也開始發展離散表徵，如探討唇語的 AV-HuBERT [46] 等等，展現了離散表徵在資料處理上的優勢。

此外，除了技術的角度切入，這樣的技術也可以探討離散語音表徵成功背後的可能因素，以及它們與語言學對人類語音理解之間的差異，甚至是進一步利用這些技術協助更細緻的探討人類的語音現象。因此，原先在連續語音表徵上的語言學分析，也開始關注離散表徵在多大程度上能描述語音現象，將其列入考量，成為除了連續語音特徵和時頻譜之外的另一個選擇。

3.1 相關研究

3.1.1 無文字與離散語音表徵

自 HuBERT 帶起的研究之後，出現了愈來愈多離散表徵相關的研究 [47, 48, 49, 50, 44, 51]。它們在提出自己的離散表徵時，也會採取 HuBERT 的衡量方式，來驗證這些離散單元與語音中的內容及人類對語音的詮釋之間，具有一定程度的相關性，並從資訊理論（Information Theory）的角度，證明這些離散單元

確實具備區分不同語音資訊的能力。

3.1.2 語音學分析

由於語音處理本身最終是針對人類語音，因此有一群研究者通過對人類語音的理解，將這些知識應用在分析模型如何對語音訊號建構表徵之上 [52, 53, 47, 48]。基於這些作品對語音離散表徵的興趣和探討，本論文也先透過過往幾個常用來分析語音表徵的方式，特別是 HuBERT [35] 提出的標準進行初步的分析。

3.2 衡量指標

本次研究主要探討純度（Purity）、熵（Entropy）和相互資訊（Mutual Information，MI）等指標，這些指標在 HuBERT 中被採用 [7, 35]，用以比對機器學習過程中得到的虛擬標註與人類標註之間的相關性（Correlation），接下來將詳細解釋這些指標的定義。

包含聲學特徵與語音基石模型，不論採用何種方式獲得語音表徵，語音訊號皆是以音框（Frame）為基本單位進行處理。具體而言，給予一段聲音訊號，語音處理系統會將這段訊號按照固定時間切割成多個片段分別處理，這些片段的長度被稱之為時間解析度（Time Resolution）。因此，對於任意一段語句（Utterance），系統會將訊號轉換成一連串的向量 $\mathbf{x} = [x_1, \dots, x_T]$ 作為語音表徵，其中 T 是該段語句的音框總數，與該語句的時長成比例。其中，第 t 個向量 x_t 表示第 t 個音框的語音訊號內容。在離散表徵的研究中，每個語音表徵向量 x_t 透過向量量化（Vector Quantization）程序，對應到編碼簿中的某個碼字 e_{z_t} 。因此，該段語句將被表示為 $z = [z_1, \dots, z_T]$ 的離散單元序列。

與此對應，藉由強迫對齊器（Forced-Aligner）或人工標註，可以獲得該段

語句的音素標註（Phonetic Label）。然而，通常音素標註是以每個音位的起始至終止的時間點配上此時間段的音位類別呈現。因此，為了配合語音表徵對語句的處理方式，這段音素標註會被依照時間點對應的範圍在音框上對齊，成為 $\mathbf{y} = [y_1, \dots, y_T]$ 的形式以便分析與後續處理。

為方便具體說明，吾人從語音常用的 LibriSpeech [54] 公開資料集中取一段音檔¹，放上波形與音框的對照在圖 3.1 呈現。該段語句內容為”... what means could it...”，上方兩個橫列為單詞標註、音位標註²。接下來四個橫列中，可以看見第三與第五個橫列將語句切割成以 20 毫秒為單位的片段，此即前面所述之音框。第三列為 HuBERT 模型分群數 100 所得之離散單元序列，而第五列則是由第二列的音位標註片段按照所對應的時間段，分別對齊到音框上的音位標註。由於音位的長度通常長於一個音框，因此在離散單元和音框音位標註在呈現上習慣將標註類別相同的音框合在一起成為長短不一但更接近時間發音的時間段，分別標在第四與第六列之上。

此時若將整個待分析資料集的語音訊號全部蒐集起來，一共有 T' 個音框，如此可分別獲得一個離散單元序列 $\mathbf{z} = \{z_t\}_{t=1}^{T'}$ 與音位標註序列 $\mathbf{y} = \{y_t\}_{t=1}^{T'}$ 進行統計分析。我們可以根據離散單元與標註之間配對的出現次數，寫為一個雙變數的共同分佈（Joint Distribution）

$$p_{yz} = \frac{\sum_{t=1}^{T'} [y_t = i \wedge z_t = j]}{T'} \quad (3.1)$$

其中 i 是第 i 個音位類別，而 j 指編號為 j 的離散單元。兩個變數的邊際機率

¹取自 train-clean-100 訓練子集，編號 89-218-0056，即編號 89 語者在章節編號 218 中第 56 句。

²ARPABet 表示法，是以純字母表示的音位表示法。介紹音位分類的章節會對此詳細描述。音位中的數字表示重音。

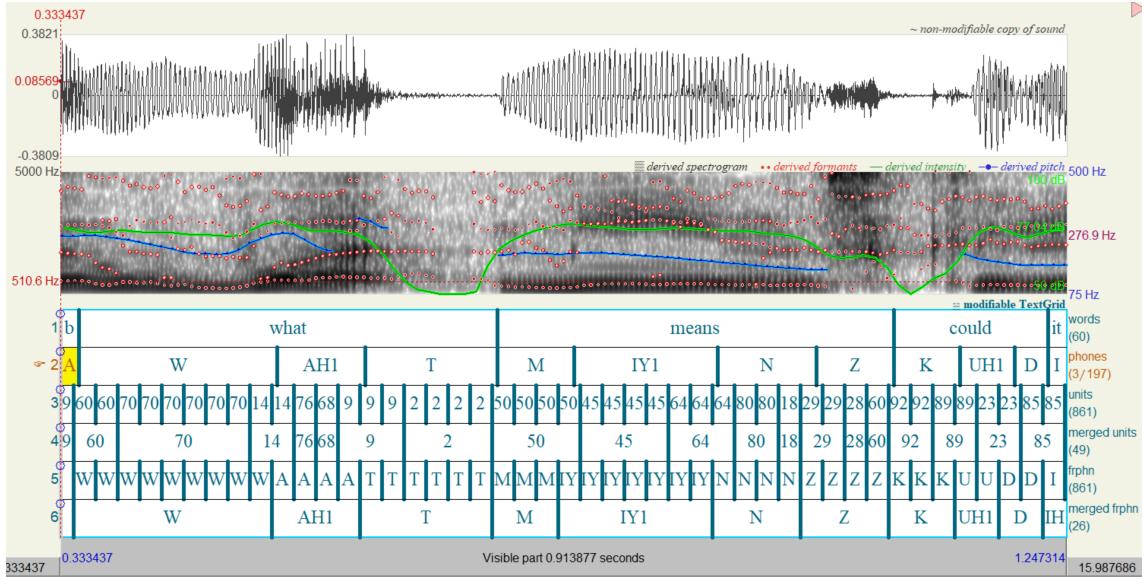


圖 3.1: 以音框對齊的離散單元與音位標註範例

(Marginal Probability) 分別為

$$p_z(j) = \sum_i p_{yz}(i, j) \quad (3.2)$$

$$p_y(i) = \sum_j p_{yz}(i, j) \quad (3.3)$$

因此，對於每一個音位 i 而言，這個音位最可能的對應離散單元為

$$z^*(i) = \arg \max_j p_{yz}(i, j) \quad (3.4)$$

與之相對應的，對於每一個離散單元的類別 j 則可以找到機率最高的音位

$$y^*(j) = \arg \max_i p_{yz}(i, j) \quad (3.5)$$

透過這些定義，以下分節介紹將要用來分析的指標。

3.2.1 純度

本指標考慮音位和離散單元兩個序列之間對應的最高機率，因此從音位與離散單元的角度出發，可以得到以下兩項數據：

音位純度 (Phoneme Purity)

考慮每個離散單元對應的音位中，最高機率音位的機率，表示為

$$\mathbb{E}_{p_z(j)} [p_{y|z}(y^*(j)|j)] \quad (3.6)$$

此指標表示該單元是否對其對應的音位有足夠的代表性。

分群純度 (Cluster Purity)

與音位純度相對，改以每個音位的角度，考慮對應單元類別的機率

$$\mathbb{E}_{p_y(i)} [p_{z|y}(z^*(i)|i)] \quad (3.7)$$

由於離散表徵進行分群演算法時的類別數是一項超參數 (Hyperparameter)，且通常離散單元的分群數量會比音位多，因此該統計數據本身不直接具有語音學的解釋意義，而且在分群數量很多時其數值會顯著下降。然而該指標在考量音位純度時必須一併考慮，因為當分群數非常多時，分群純度過低暗示離散單元做不到歸納音位類別的效果，使得音位純度失去其意義。一個極端的情形是每一個音框都給予不同的離散單元編號，如此音位純度可以達到 100%。

3.2.2 熵和相互資訊

除了純度提供「最高機率」的對應關係，根據 HuBERT 論文 [35] 中的分析方式，我們也可以從資訊理論的角度，觀察兩個序列的熵和相互資訊。

熵 (Entropy)

熵的定義按照資訊理論，衡量兩個序列中標籤類別出現機率的不確定性

(Uncertainty)，公式寫作：

$$H(y) = \sum_i p_y(i) \log p_y(i) \quad (3.8)$$

$$H(z) = \sum_j p_z(j) \log p_z(j) \quad (3.9)$$

其中 $H(y)$ 和 $H(z)$ 分別為音位和離散單元的熵，數值愈高分別表示各種音位和離散單元出現的機率愈平均。

以音位標準化之相互資訊 (Phone-normalized Mutual Information, PNMI)

本數據以「觀察到某一個離散單元，能降低多少音位標註的不確定性」，定義該離散單元的出現背後提供了多少音位的資訊。公式寫為：

$$\frac{I(y; z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i, j) \log \frac{p_{yz}(i, j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \quad (3.10)$$

$$= \frac{H(y) - H(y|z)}{H(y)} \quad (3.11)$$

$$= 1 - \frac{H(y|z)}{H(y)} \quad (3.12)$$

該項數據愈高，表示離散單元的分群愈能提供語音音位的資訊，是一個品質更好的分群結果。由於離散單元是否能夠正確對應到音位才是人們所關心的問題，因此與純度不同，只以音位的角度出發，而不考慮以離散單元分群的角度。

3.3 語音學的音位分類 (Phoneme Type)

除了單一音位本身的特性以外，由於音位之間存在相似的特徵，可以分成幾個組別。這裡依照希氏 (Sicherman) [47]、阿氏 (Abdullah) [48] 等前作的分組方式，對英語的音位進行分類。如此一來，除了單純把音位標註以約 40 類完全獨立

的標籤看待，還能夠觀察這些離散單元是否有擷取到相似的發聲特徵。首先，按照發音過程氣流是否受到阻礙，因此可否形成獨立的音節，音位可以分為輔音與元音兩大類，而後再根據發音的細部特性共分成七組。

輔音 (Consonant)

輔音是指透過阻擋氣流發聲的音位，因此通常不單獨構成音節，按照發音方式可分為以下五個類別：

- 塞音 (Plosive)：以完全阻塞氣流的方式發音的音位，包含 /p/、/b/、/t/、/d/、/k/、/g/ 六種。
- 擦音 (Fricative)：藉由在口腔中形成的縫隙，使氣流通過時摩擦形成的發音，包含 /f/、/v/、/s/、/z/、/ʃ/ (sh)、/ʒ/ (如「garage」的「-ge」)、/θ/ (無聲的 th)、/ð/ (有聲的 th)、/h/ 九種。
- 塞擦音 (Affricate)：由塞音和同部位的擦音同時發出的輔音，英語中只有 /tʃ/ 和 /dʒ/ 兩種，即 ch 和 j 的發音。
- 鼻音 (Nasal)：使氣流通過鼻腔形成的聲音，有 /m/、/n/、/ŋ/ (ng) 三種。
- 近音 (Approximant)：又稱半元音，為介於元音和輔音之間的聲音，有 /j/ (為 y 作為輔音時的發音)、/r/、/l/、/w/ 四種。

元音 (Vowel)

與之相對，元音則是不阻礙氣流通過，因此可自成音節的音位。其中又可分為發音位置固定的單元音 (Monophthong) 和會移動發音位置的雙元音 (Diphthong) 兩類。通常以 a、e、i、o、u 字母產生的聲音皆屬於此類別。

透過將音位分成以上七組後，並重新分析統計指標，以觀察這些分組的規律如何在離散單元的出現機率上呈現，進而顯示離散單元是否與語音的發音方式具有一定的關聯性。

另外，為了方便統計與作圖，這些音位在圖中並非以語言學慣用之國際音標（International Phonetic Alphabet，IPA）[55]，而是參考語音處理領域常用的「卡內基梅隆大學發音辭典（Carnegie Mellon University Pronouncing Dictionary，CMUDict）[56]」，取用其中的 ARPABet 表示法 [57]，以避免字母以外的符號在處理上的困難。表 3.1 中列有更詳細的音位資訊³。

3.4 實驗集與分析模型

本研究的分析對象參考無文字架構 [42, 45, 5] 的研究，採用論文中提及的四種語音表徵，簡述如下：

- HuBERT [35]：卷積式編碼器 + 轉換器預測器，以預測式學習訓練，其訓練目標為 K-平均分群演算法的結果，透過遮蔽語言模型的方式訓練。表徵來自轉換器第 6 層，每 20 毫秒作為一個音框
- CPC [38]：卷積式編碼器 + 遞迴式預測器，以對比式學習訓練。表徵來自預測器的中間層，每 10 毫秒提取一個向量表徵作為音框
- Wav2vec 2.0 [39]：卷積式編碼器 + 轉換器預測器，以對比式學習訓練。表徵來自轉換器第 14 層，每 20 毫秒作為一個音框
- LogMel：為 80 維對數梅爾時頻譜的聲學特徵，在此作為比較基線（Baseline）。音框寬度為 10 毫秒

³範例單詞取自 CMUDict 官網 (<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>) 說明。

音位	ARPABet 表示法	音位分類	範例單詞	範例單詞的音位
/a/	AA	單元音	odd	AA D
/æ/	AE	單元音	at	AE T
/ʌ/	AH	單元音	hut	HH AH T
/ɔ/	AO	單元音	ought	AO T
/aʊ/	AW	雙元音	cow	K AW
/aɪ/	AY	雙元音	hide	HH AY D
/b/	B	塞音	be	B IY
/tʃ/	CH	塞擦音	cheese	CH IY Z
/d/	D	塞音	dee	D IY
/ð/	DH	擦音	thee	DH IY
/ɛ/	EH	單元音	Ed	EH D
/ɜ/	ER	單元音	hurt	HH ER T
/eɪ/	EY	雙元音	ate	EY T
/f/	F	擦音	fee	F IY
/g/	G	塞音	green	G R IY N
/h/	HH	擦音	he	HH IY
/ɪ/	IH	單元音	it	IH T
/i/	IY	單元音	eat	IY T
/dʒ/	JH	塞擦音	gee	JH IY
/k/	K	塞音	key	K IY

表 3.1: 英語音位的 ARPABet 表示法和音位分類資訊

音位	ARPABet 表示法	音位分類	範例單詞	範例單詞的音位
/l/	L	近音	lee	L IY
/m/	M	鼻音	me	M IY
/n/	N	鼻音	knee	N IY
/ŋ/	NG	鼻音	ping	P IH NG
/əʊ/	OW	雙元音	oat	OW T
/ɔɪ/	OY	雙元音	toy	T OY
/p/	P	塞音	pee	P IY
/r/	R	近音	read	R IY D
/s/	S	擦音	sea	S IY
/ʃ/	SH	擦音	she	SH IY
/t/	T	塞音	tea	T IY
/θ/	TH	擦音	theta	TH EY T AH
/ʊ/	UH	單元音	hood	HH UH D
/u/	UW	單元音	two	T UW
/v/	V	擦音	vee	V IY
/w/	W	近音	we	W IY
/j/	Y	近音	yield	Y IY L D
/z/	Z	擦音	zee	Z IY
/ʒ/	ZH	擦音	seizure	S IY ZH ER

表 3.1: 英語音位的 ARPABet 表示法和音位分類資訊（續）

我們跟隨拉氏等人所提出的無文字架構 [5]，使用該篇論文中釋出之預訓練模型與 K-平均量化模型，預訓練模型的設定細節於原論文有更詳細的描述，而量化模型則是拉氏等人透過公開的 LibriSpeech 資料集 [54] 中之 train-clean-100 訓練子集，獲取語音表徵後執行 K-平均分群演算法所得，並釋出分群數為 50、100 和 200 的三個版本。

本論文以 LibriSpeech 之 train-clean-100 訓練子集作為分析對象，將語音語料庫的語音資料經過四個模型得到連續表徵後，再經過量化模型得到完全由離散單元組成的「虛擬文字」語料。至於音位標註的取得，則是透過強迫對齊器⁴的英語預訓練模型，將語料庫的文字轉寫轉換為帶有對應時間範圍的音位標註資料，並依據各自語音表徵的時間解析度，生成以音框對齊的音位標註語料，隨後進行相關性的分析。

3.5 分析方式

在前述章節中，我們為了討論離散單元和音位標註之間的關係，介紹了相關的研究、衡量指標與語音學的分類方式。接下來，我們將詳細描述分析相關性的具體方法，並以圖表展示分析結果，希望可以藉此對離散表徵獲得更深刻的理解。為了更直觀解釋這些指標的意義並看清這些數字背後所代表的現象與細部特徵，我們使用熱圖（Heatmap）來呈現音位與離散單元的共同機率分佈 p_{yz} 。這樣的可視化（Visualization）方式有助於深入探討這些指標的意義。

首先，圖 3.2 以 HuBERT 為基石模型，離散單元分群數為 50 的統計數據為例，說明我們如何分析語音離散表徵與音位標註的關係。

圖中的縱軸表示各個音位，橫軸表示各個離散單元。在這張圖中，縱軸的音

⁴<https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner>

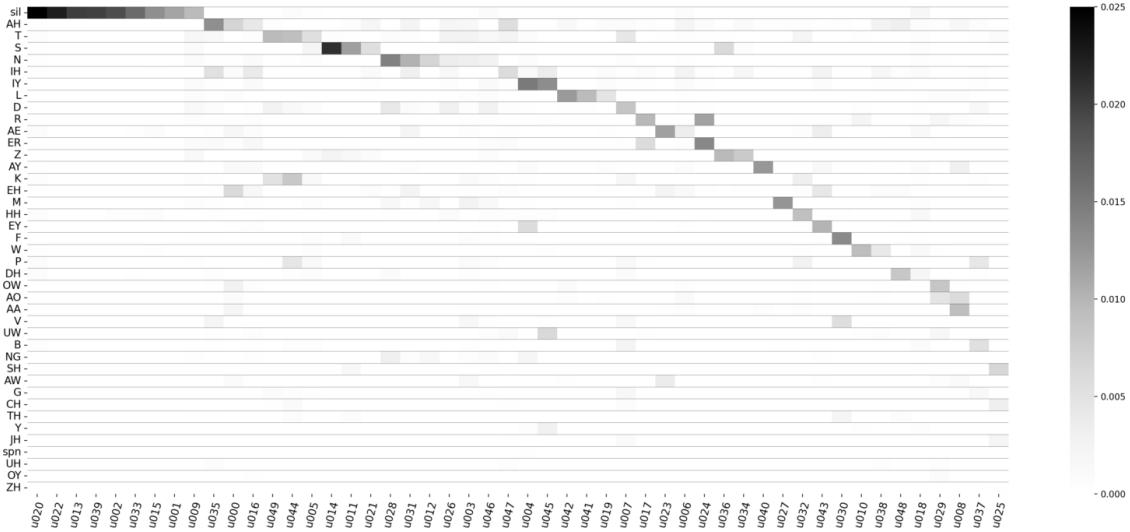


圖 3.2: HuBERT 模型、分群數為 50 之離散單元與音位標註的共同機率分佈圖

位是按照其邊際機率 $p_y(i)$ 由高至低排序；橫軸的離散單元則是依據其對應的最高機率音位 $y^*(j)$ 的縱軸排序位置進行排列。⁵ 這樣可以在熱圖上顯示由左上至右下的對應關係。

藉由熱圖的幫助，我們不僅可以更加完整清晰的觀察離散單元與音位標註之間的關係，對於純度大小的意義也可以從此圖上有更具體的了解：

1. 將每個直行 (Column) 取最大值相加後的總和即為音位純度。如果每個離散單元與音位都相對集中，則可以得到較高的音位純度。且如同指標說明的小節所述，當分群數量增加時，音位純度能夠在每個直行上取到更多的機率值總和。最極致的情況是，當分群數量與音框數量相同，音位純度可以達到 100%。這些性質透過熱圖的可視化呈現，可以被更直觀的說明。

2. 將每個橫列 (Row) 取最大值相加後的總和則是分群純度。如果每個音位都恰好可以很集中的對應到少數幾個離散單元，則此數值將較高，每個橫列最

⁵如果兩個離散單元 j_1 和 j_2 對應到相同的音位 $y^* = y^*(j_1) = y^*(j_2)$ ，則依照機率值 $p_{yz}(y^*, j_1)$ 和 $p_{yz}(y^*, j_2)$ 由高到低進行排序，對於多個離散單元的情況以此類推。

高可以貢獻的值為該音位出現的機率 $p_y(i)$ 。同樣的，當分群數量增加時，隨著直行數目的增多，單看每一個音位對應的橫列，會發現每個格子的機率值隨之被稀釋。受到音位標註類別數的限制，分群純度最高只能取 41 個 p_{yz} 值的總和，使得單位純度因而明顯降低。

此外，比起只有音位與分群純度兩個數字，機率熱圖不但可以呈現純度指標的綜觀解釋性意義，我們還可以分門別類對個別的音位與離散單元進行細部探討。畢竟，模型的虛擬標註與實際人類給予的標註資料並不能總是完美而集中的互相對應。我們想知道的細部觀察可分為兩個層面：

1. 從離散單元的角度出發，每個單元 j 所對應的音位是如何的集中，因而多能夠代表這個單元中最高機率的音位 $y^*(j)$ ？如果恰巧該單元對應的音位條件機率分佈 $p_{y|z}(i|j)$ 較為分散，那與這個單元最相關，也就是條件機率前幾高的音位之間，又是否呈現特定關係？
2. 反之從音位標註考慮，對於每個音位 i ，觀察它所對應的離散單元集中程度，也就是離散單元條件機率分佈 $p_{z|y}(j|i)$ 得出的熵值 $H(z|y)$ ，可否觀察到特定一些音位較難或較易被離散單元集中歸類，進而推論模型是否善於辨認該音位的發音特性。

接下來，我們將從綜觀的角度比較來自不同語音表徵與分群數的離散單元的純度和相互資訊數據，並輔以對應的機率熱圖佐證，觀察離散表徵在捕捉發音資訊方面的強弱。此後，分別從離散單元和音位兩個面向，藉助音位分類知識的幫助，進行細部觀察。最後將細部觀察的結論，重新對應回機率熱圖上的深淺規律，以對這些觀察的進行驗證。

3.6 分析結果

3.6.1 綜觀分析

表 3.2 提供了不同語音表徵與分群數的純度和相互資訊的指標數據。

首先，我們先比較同樣是分群數為 50 時，四種語音表徵的共同機率分佈熱圖，呈現在圖 3.3 中。從圖中可以明顯觀察到，HuBERT 和 CPC 在熱圖上具備較多較深且清晰的方塊，這表示音位與離散單元之間的對應相對 Wav2vec 2.0 與 LogMel 更為明確。這反映出 HuBERT 和 CPC 的離散表徵更擅長捕捉並區分音位之間的關係。此觀察也對應到這兩個模型較高的音位純度與相互資訊數值。

接著考慮分群數的效應，我們進一步觀察分群表現最好的模型 HuBERT 在分群數為 50、100 和 200 的共同機率分佈熱圖。圖 3.4 是三者的比較結果，從圖中可以發現，在分群數愈多時，熱圖較深的區域愈是可以集中連成一條線，落在線外的色塊變得更少，但每個格子的機率值也隨之迅速降低。這個趨勢可以解釋為什麼表 3.2 中上升的音位純度與下降的分群純度，不過從表格中可以發現，其實相互資訊的數值仍是隨著分群數上升而提高的，也就是分群數多時，可以幫助提升離散單元與音位標註之間的相關性。

以上是不同離散表徵系統的離散單元對語音訊號給予虛擬標註時，對應音位標註是否明確的觀察。我們發現 HuBERT 是四種語音表徵之中效果最佳的，而分群數則是愈多愈好。

3.6.2 以離散單元角度切入

探討完綜觀機率分佈的比較，接著我們從離散單元的角度出發，基於離散單元進行統計觀察。首先，我們可以如同綜觀分析的探討方式，分別從模型與分群

	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.5256	0.3382	3.3152	3.8681	0.4993
CPC	0.5188	0.3812	3.3146	3.7918	0.4992
Wav2vec 2.0	0.4006	0.2676	3.3152	3.8215	0.3706
LogMel	0.3253	0.1473	3.3158	3.8630	0.2647

(a) 分群數 = 50

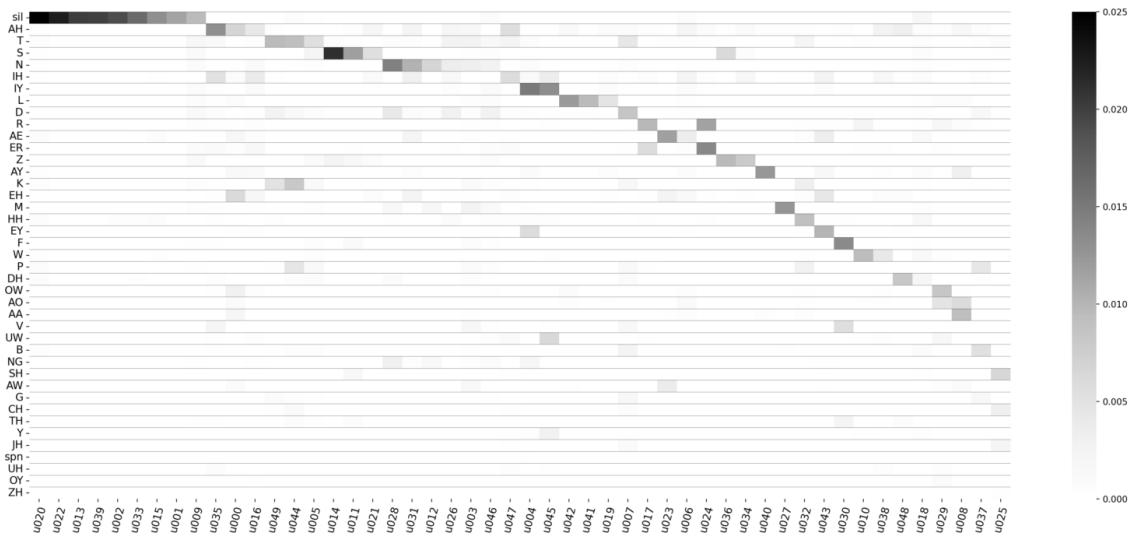
	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.6097	0.2553	3.3152	4.5704	0.5786
CPC	0.5895	0.2674	3.3146	4.5034	0.5557
Wav2vec 2.0	0.4877	0.2118	3.3152	4.5284	0.4596
LogMel	0.3348	0.0931	3.3158	4.5591	0.2789

(b) 分群數 = 100

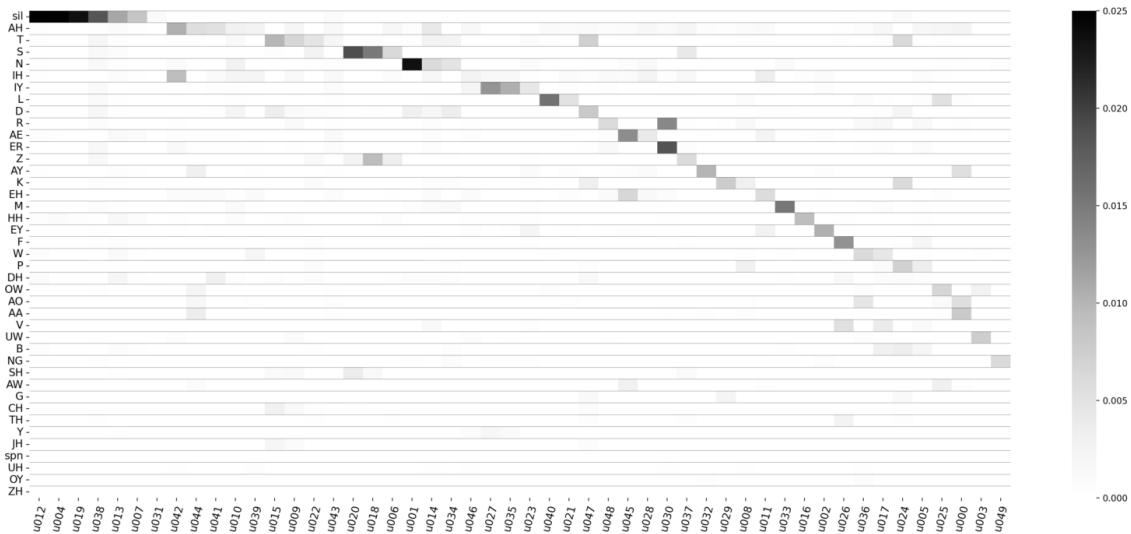
	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.6474	0.1644	3.3152	5.2681	0.6289
CPC	0.6098	0.1789	3.3146	5.1885	0.5882
Wav2vec 2.0	0.5427	0.1467	3.3152	5.2173	0.5188
LogMel	0.3474	0.0569	3.3158	5.2322	0.2955

(c) 分群數 = 200

表 3.2: 四種語音表徵在不同分群數的純度與相互資訊數據

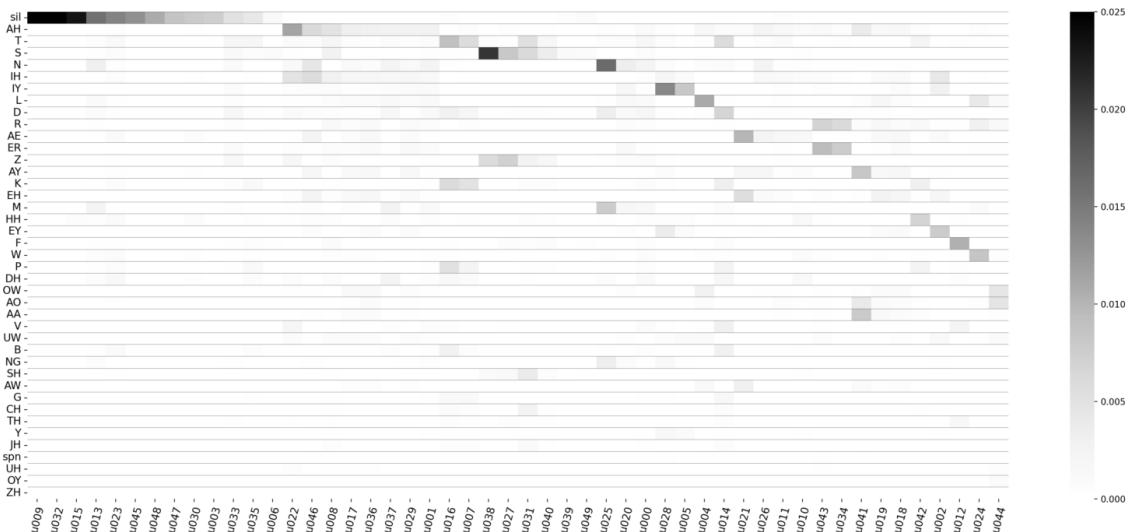


(a) HuBERT



(b) CPC

圖 3.3: 不同語音表徵在分群數為 50 的共同機率分佈熱圖



(c) Wav2vec 2.0



(d) LogMel

圖 3.3: 不同語音表徵在分群數為 50 的共同機率分佈熱圖（續）

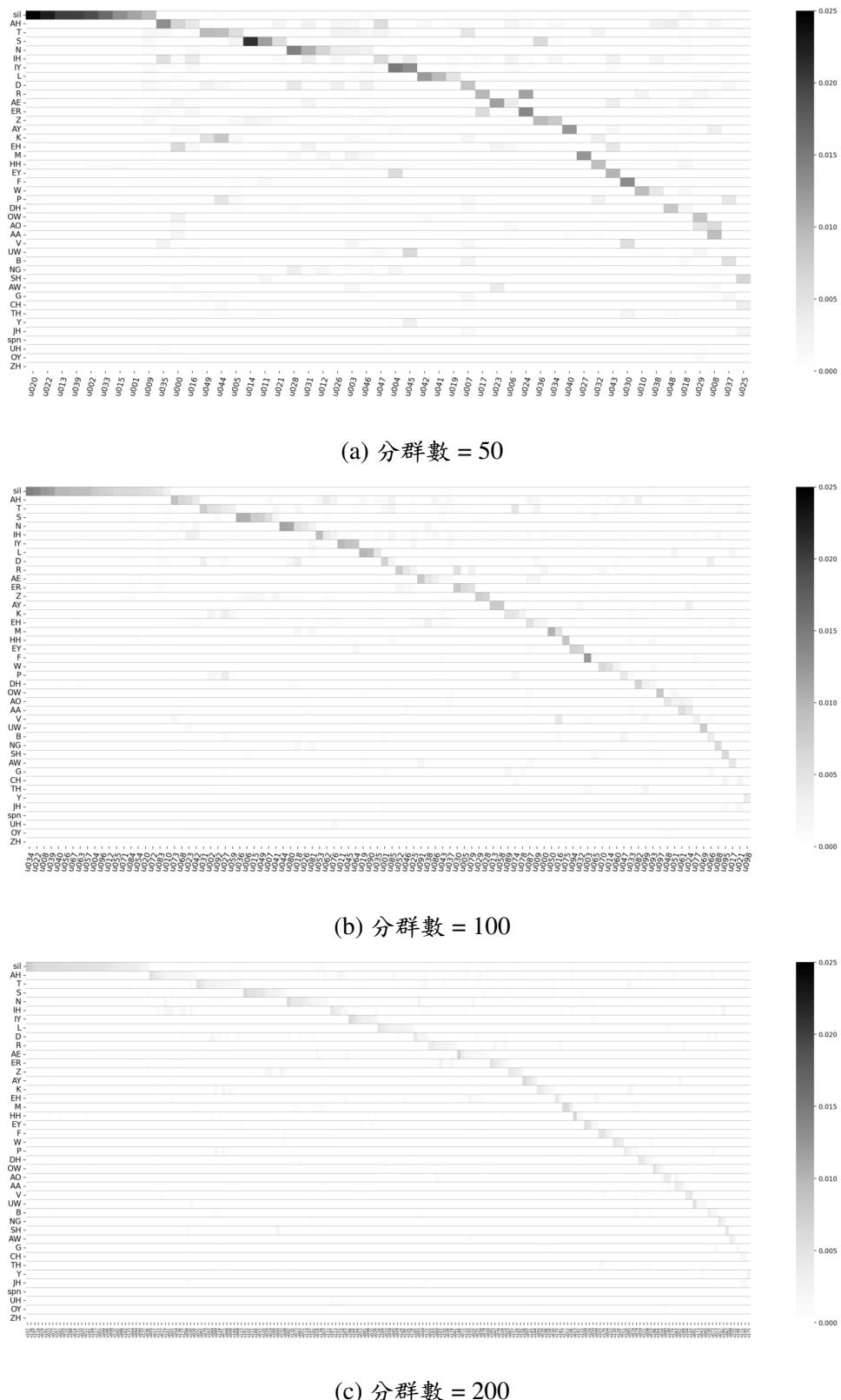


圖 3.4: HuBERT 模型在不同分群數的共同機率分佈熱圖

數量兩個變量切入，並對每個離散單元計算對應的條件音位熵 $H(y|z)$ 並畫出直方圖進行比較。

圖 3.5 可以觀察到四種模型在分群數都是 50 時的條件音位熵直方圖，從圖中可以發現 HuBERT 和 CPC 的音位熵相較於 Wav2vec 2.0 和 LogMel 偏低，也就是 HuBERT 和 CPC 每個離散單元對應的音位相較集中，與綜觀探討得到的觀察相符。

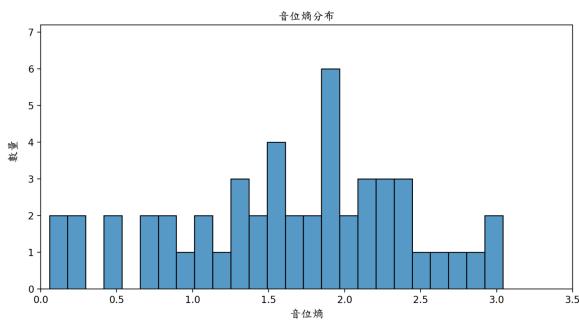
圖 3.6 則是比較 HuBERT 模型在分群數為 50、100 和 200 時的條件音位熵。需注意的是，由於此時離散單元數量不同，因此直方圖的縱軸改以比例數值呈現，亦即將數量分別除以 50、100 和 200 以進行公平的比較。從圖中可以觀察到，分群數愈多確實使整體條件音位熵降低，也與綜觀探討得到的小結一致。

接著，由於本小節基於離散單元的角度，我們仿照前作如 SpeechTokenizer [44]、DinoSR [50] 的作法，將熱圖改以 $p_{y|z}(i|j)$ 呈現，即對每個直行進行標準化得到條件機率，以顯示每個單位對應到哪個音位，探討這種對應分佈是如何的集中或分散。

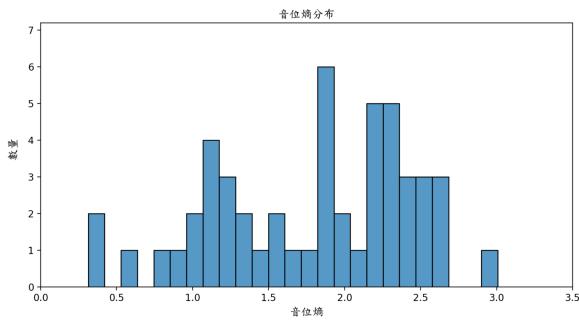
觀察圖 3.7 中由左上而右下角對應的連線區域，首先我們會在左上方觀察到一條明顯較深的區域，也就是模型會安排一定數量的離散單元用以對應實際上並非音位的音位標註 sil。此外，我們還可以在連線區域之外觀察到一些零星的色塊，在此指示存在不少離散單元，它們對應的音位是相比較為分散的，也因此使得音位純度無法到達 100%。

不過，如果我們嘗試觀察這些將離散單元對應機率分散出去的音位，藉助語音學的知識可以觀察到一些有趣的發現：這些音位彼此之間在發音上具有很強的關聯性，幾乎與語音學提供的分類是對應的。

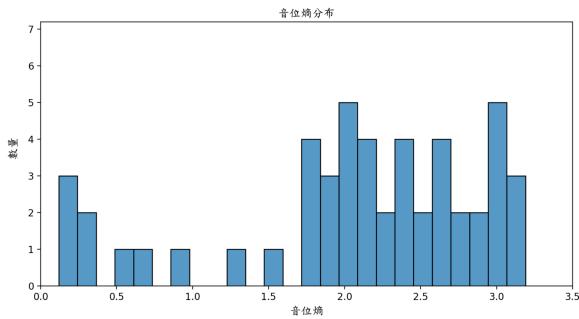
為了方便說明，我們將熱圖上各個離散單元排名前五高的對應音位另外列表



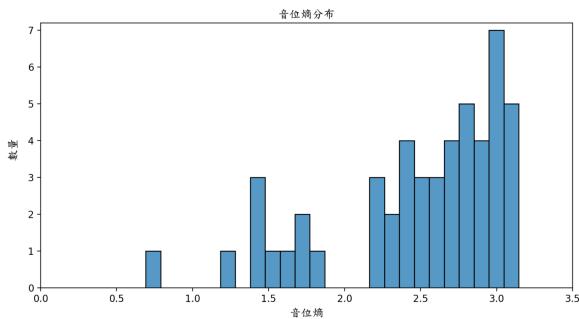
(a) HuBERT



(b) CPC

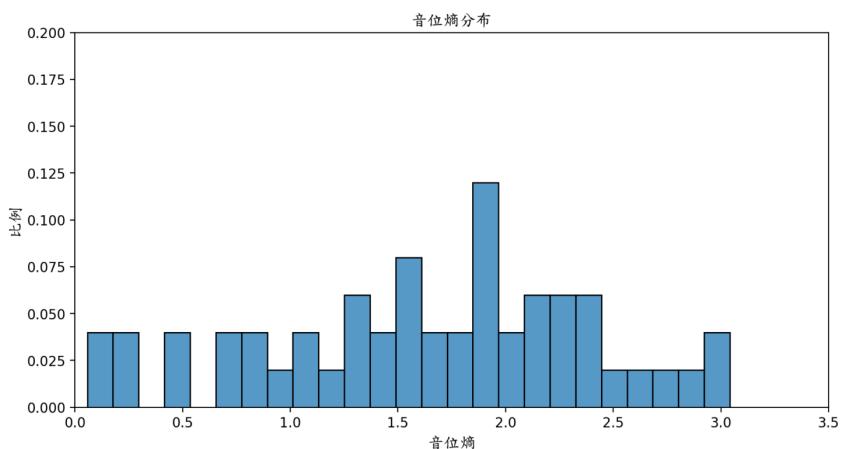


(c) Wav2vec 2.0

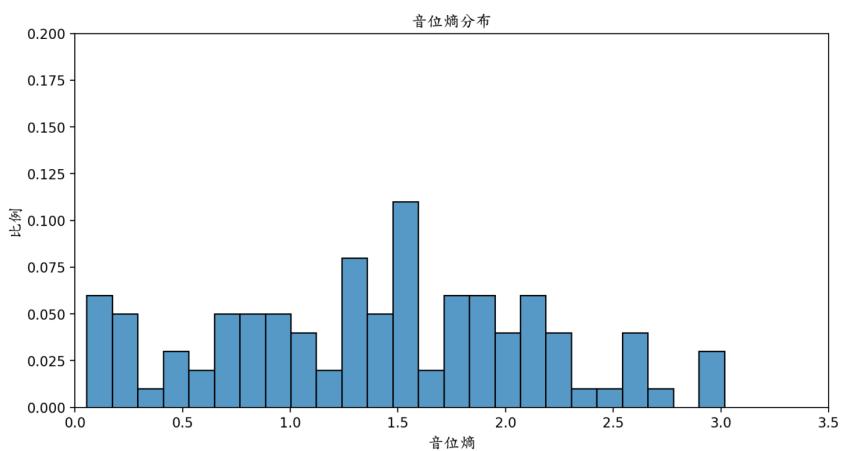


(d) LogMel

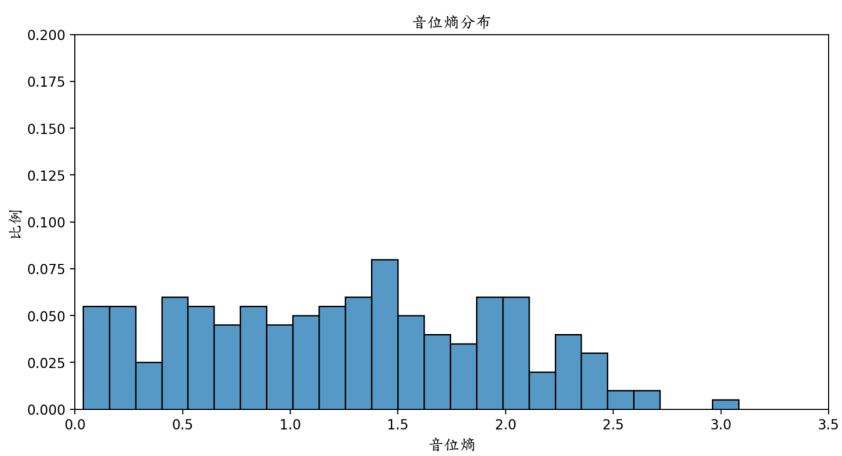
圖 3.5: 不同語音表徵在分群數為 50 的條件音位熵直方圖



(a) 分群數 = 50



(b) 分群數 = 100



(c) 分群數 = 200

圖 3.6: HuBERT 模型在不同分群數的條件音位熵直方圖

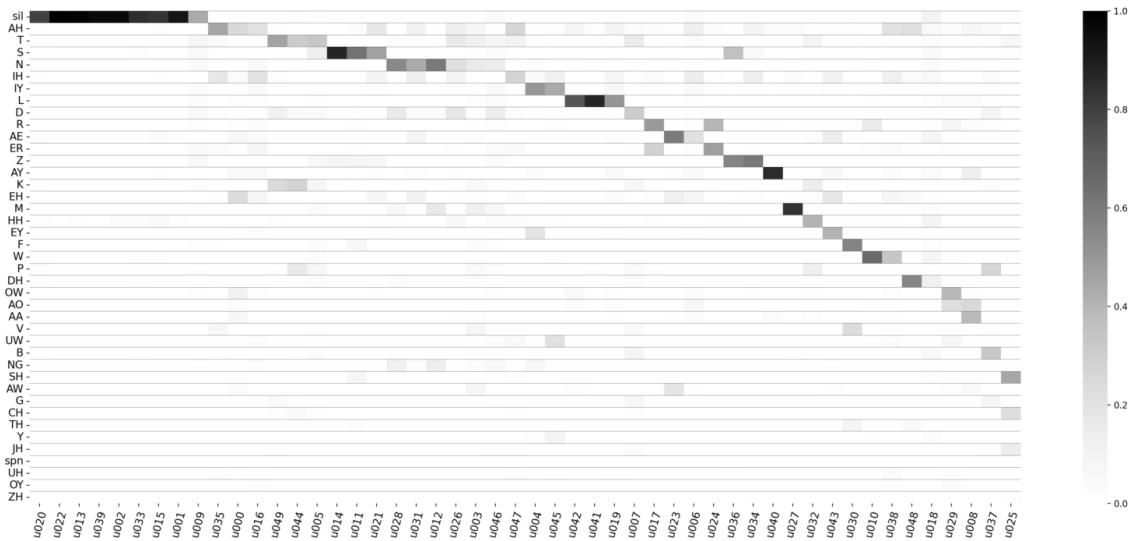


圖 3.7: HuBERT 模型、分群數為 50 之 $p_{y|z}(i|j)$ 條件機率分佈圖

	u035	u000	u016	u049	u044	u005	u014	u011	u021	u028	u031	u012	u020
rank 1	AH	AH	AH	T	T	T	S	S	S	N	N	N	
rank 2	IH	EH	IH	K	K	S	Z	SH	AH	D	IH	M	
rank 3	V	OW	EH	D	P	K	T	Z	IH	NG	EH	NG	
rank 4	T	AA	ER	G	D	P	spn	F	Z	M	AH	D	
rank 5	ER	AE	IY	AH	CH	Z	HH	TH	EH	DH	AE	UW	

圖 3.8: HuBERT 模型、分群數為 50 之部分離散單元所對應的前五高機率音位
(音位分類以顏色標示區分)

CONSONANTS (PULMONIC)												© 2020 IPA	
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal		
Plosive	p b			t d		t d	c ɟ	k g	q ɢ			? ʔ	
Nasal	m	m̪		n		n̪	n̪	ŋ	ɳ			N	
Trill	B			r						R			
Tap or Flap		v̪		r̪		t̪							
Fricative	f̪ β̪	f v̪	θ̪ ð̪	s̪ z̪	f̪ z̪	s̪ z̪	ç̪ j̪	x̪ y̪	χ̪ ʁ̪	h̪ ɸ̪	h̪ f̪		
Lateral fricative			ɬ̪ ɭ̪										
Approximant		v̪		ɹ̪		ɻ̪	j̪	w̪					
Lateral approximant				l̪		ɺ̪	ʎ̪	ɻ̪					

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

圖 3.9: 國際音標表的輔音表格，說明離散單元

對語音聲學特徵的捕捉並不僅限單一面向

呈現在圖 3.8 中，並用顏色標明各音位所屬的音位分類。從表中大致可以看出前幾高機率的音位所屬的類別確實是相近的。而且即便不是同一個音位分類，這些音位在語音學中，仍有其他層面——如發音部位和清濁音——的相似性，還是可以將各離散單元的前幾高音位中找出共通點。

事實上，為了作圖與統計方便，語音處理相關研究 [47, 48] 對音位的歸類是相對簡化的。根據語音學的知識，音位之間的分組方式並不只一種，而本研究著重的分類方式僅是以「發音方式」為主。例如 05 號單元對應的前兩名 /t/ 和 /s/ 雖然並不屬於同一個發聲方式，因而被分成兩個類別，但如果參考圖 3.9 的國際音標表⁶，會發現它們都屬於「齒音」，亦即它們的「發音部位」是相同的。換言之這些離散單元捕捉到的語音特性是多個面向的，並不僅限於單一的分類方式，而是可以對應到國際音標表上至少兩個維度以上的類型。

透過以上的觀察，因此我們有足夠的理由重新對熱圖的縱軸重新排列，並按照語音學分類進行分組，來觀察這些離散單元是如何指示出音位之間的相似性，

⁶表中的每個橫列約等於本論文與相關研究 [47, 48] 使用的「發音方式」分類法，而每個用直線隔開的直行則是指「發音部位」相同，同一個格子的左右則是呈現一對清濁音音位。

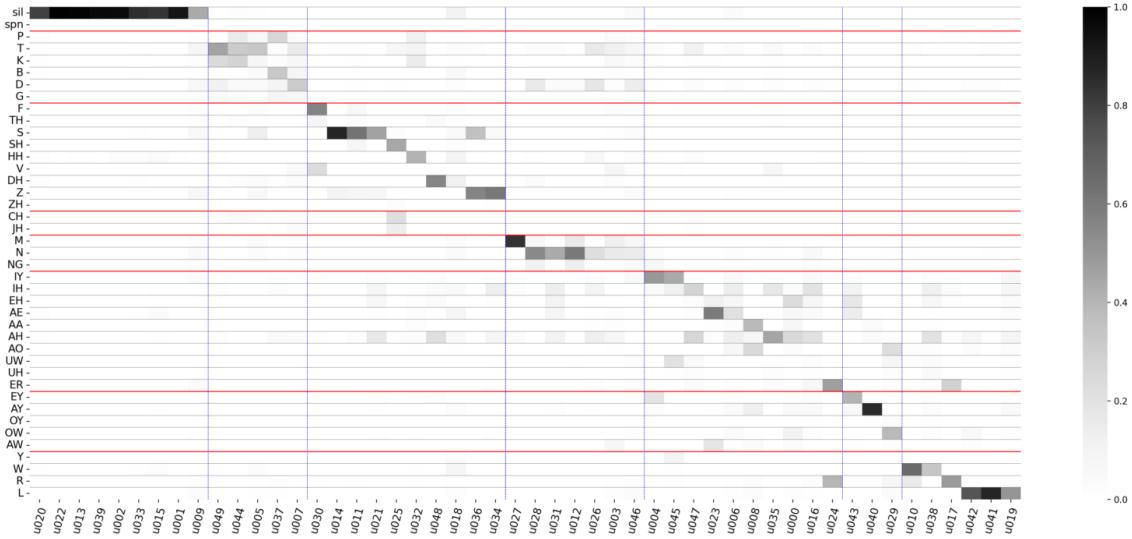


圖 3.10: HuBERT 模型、分群數為 50 之離散單元與音位標註的條件機率，
依照韋氏 (Wells) [53] 論文與音位類別排序的分佈圖

區分出同個音位、同類發音，或者如何被混淆為其他類別。

圖 3.10 的分組順序是依照韋氏 (Wells) [53] 論文中的出現順序排列，而組別內則是清音在上、濁音在下，而同樣清濁音則是以發音位置由前往後排列。除了縱軸上按照音位本身特性分組，依循純度中使用的「代表音位」 $y^*(j)$ 概念，我們同樣也對每個離散單元的代表音位排序，並且也依照這些代表音位進行分組觀察。

最後，對於每個離散單元 j 與對應的最高機率音位 $y^*(j)$ ，為了統計該單元 j 除了 $y^*(j)$ 以外，是否也給予與 $y^*(j)$ 同音位分類 $\kappa^*(j)$ 的其他音位較高的機率值，我們藉由調整純度的計算式，但將音位標註改為音位分類並重新統計，以方便我們比較離散單元「對音位分類歸類能力」的強弱。計算方式為：

如果 $y^*(j) \in \kappa^*(j)$ ， $\kappa^*(j) = \{y^*(j), y', y'', \dots\}$ 是所有與 $y^*(j)$ 同音位分類的音位，則將這些音位的標註改為 $\kappa^*(j)$ ，統計音位分類純度

$$\mathbb{E}_{p_z(j)} [p_{\kappa|z}(\kappa^*(j)|j)] \quad (3.13)$$

語音表徵	分群數	音位類別純度	以音位類別標註之分群純度
HuBERT	50	0.7006	0.1509
HuBERT	100	0.7584	0.0882
HuBERT	200	0.7793	0.0488
CPC	50	0.7019	0.1998
CPC	100	0.7417	0.1054
CPC	200	0.7564	0.0684
Wav2vec 2.0	50	0.6304	0.1570
Wav2vec 2.0	100	0.6856	0.0950
Wav2vec 2.0	200	0.7163	0.0564
LogMel	50	0.5382	0.0969
LogMel	100	0.5436	0.0581
LogMel	200	0.5510	0.0345

圖 3.11: 以音位分類為標註計算，四種語音表徵在不同分群數的純度數據與音位分類的分群純度

$$\mathbb{E}_{p_{\kappa}(\kappa^*)} [p_{z|\kappa}(z^*(\kappa^*)|\kappa^*)] \quad (3.14)$$

以此刻劃離散表徵是否能歸類語音學上相似的發音特徵。

圖 3.11 是不同離散表徵在音位分類純度與對應的分群純度結果。由表中可以再次確認，HuBERT 的離散單元不但能夠很好的區分出音位，即便某些離散單元沒有集中分到特定音位之上，也可以很不錯的給予同類別的音位較高的機率值，以得到較高的音位分類純度數值。

3.6.3 以音位角度切入

接著，我們改從音位的角度切入，觀察每個音位所對應的離散單元條件熵 $H(z|y)$ ，以探討不同音位之間是否有特定音位較容易或較難以被離散表徵歸類。表 3.12 分別呈現了不同模型在分群數為 50 和 100 時，離散單元熵最高與最低的幾個音位。雖然沒有特別明顯的趨勢，但可以大致看到以下幾點：

- 熵值較低的音位有 AA、EY、F、ZH、SH、S 等，其中 F、ZH、SH、S 皆屬於擦音。
- 熵值較高的音位有 spn、AH、IH、T、D 等，其中 T、D 屬於塞音。

整體而言，擦音的離散單元相對較為集中，而塞音則相對較為分散。至於其他如 AH、IH 等高熵值的元音音位，推測其原因可能是它們本身在不同發音情境下音色的變化相對於 AA、EY 等較大，因而較難集中於某幾個離散單元。這個趨勢在不同的語音表徵和分群數下約略存在，但以 HuBERT 和 CPC 較為明顯。

為了進一步驗證不同音位分類之間的差異，我們再度引用純度與相互資訊的計算公式，但將統計範圍限定為不同音位分類，分別計算針對每個音位分類的純度與相互資訊。換言之，我們共同機率分佈圖 p_{yz} 按照圖 3.13 的紅色水平線分成八塊後，重新標準化並各自計算純度與相互資訊，相當於將原本的語音音框按照音位分類分成八組各自統計這些指標。如此一來，我們既可以依據每個音位分類觀察在不同離散表徵對該類別的表現差異，也可以比較不同音位分類彼此的整體趨勢，歸納音位分類本身發音特徵被捕捉的難易程度。

表 3.14 中呈現了這些模型的比較數據。由圖中依然可以觀察到 HuBERT 優於其他模型，且分群數愈多時相互資訊與音位純度愈高，這些趨勢與前面所有的觀察結論一致。

從音位分類之間的比較，我們可以觀察到：撇除非音位（表格中的「XXX」）類別不考慮，⁷可以確定塞音和塞擦音的純度較低，確實是較難以集中歸類的音位分類；而近音、雙元音和擦音則純度相對較高，這也驗證為什麼它們的離散單元熵值較低且分佈較為集中。

⁷因其只有 sil 一類標註，因此相互資訊和純度相當高。

	離散單元熵最高 (分群數 = 50)					離散單元熵最低 (分群數 = 50)			
	HuBERT	CPC	Wav2vec 2.0	LogMel		HuBERT	CPC	Wav2vec 2.0	LogMel
# 1	spn	spn	spn	spn	# 1	ZH	E	AA	SH
# 2	AH	AH	AH	T	# 2	SH	AA	E	sil
# 3	IH	IH	UW	HH	# 3	E	sil	S	S
# 4	T	DH	IH	AH	# 4	EY	ER	EY	AA
# 5	D	UH	T	DH	# 5	AA	NG	SH	ZH
# 6	EH	T	D	G	# 6	W	S	AW	AO
# 7	TH	D	DH	D	# 7	S	M	IY	AW
# 8	HH	EH	HH	V	# 8	CH	CH	ER	EY
# 9	UH	HH	TH	UW	# 9	IY	EY	CH	IY
#10	G	TH	UH	JH	#10	Y	ZH	AO	Z
#11	sil	Y	R	IH	#11	OW	AW	OY	Y
#12	N	OY	V	UH	#12	Z	SH	ZH	EH
#13	K	AE	L	K	#13	AO	P	P	E
#14	AE	W	EH	OY	#14	L	UW	NG	CH
#15	P	G	JH	B	#15	ER	L	W	OW

(a) 分群數 = 50

	離散單元熵最高 (分群數 = 100)					離散單元熵最低 (分群數 = 100)			
	HuBERT	CPC	Wav2vec 2.0	LogMel		HuBERT	CPC	Wav2vec 2.0	LogMel
# 1	spn	spn	spn	spn	# 1	SH	SH	AA	SH
# 2	AH	AH	AH	AH	# 2	Y	E	SH	S
# 3	IH	IH	IH	HH	# 3	ZH	ZH	EY	sil
# 4	T	UH	T	T	# 4	E	NG	AW	ZH
# 5	D	EH	D	DH	# 5	NG	S	OY	AA
# 6	sil	DH	DH	G	# 6	EY	CH	AY	Z
# 7	EH	D	TH	D	# 7	UW	P	AO	AO
# 8	HH	T	UH	UW	# 8	W	K	E	AW
# 9	UH	HH	HH	IH	# 9	AW	V	OW	IY
#10	AE	R	V	V	#10	AY	EY	ZH	E
#11	K	AE	N	UH	#11	M	M	UW	CH
#12	N	AO	JH	OY	#12	AA	Z	sil	Y
#13	R	TH	EH	JH	#13	OW	IY	S	EY
#14	G	N	R	AE	#14	CH	UW	ER	L
#15	P	sil	G	N	#15	L	JH	CH	TH

(b) 分群數 = 100

圖 3.12: 不同語音表徵在分群數為 50 和 100 時，離散單元條件熵最高與最低的音位

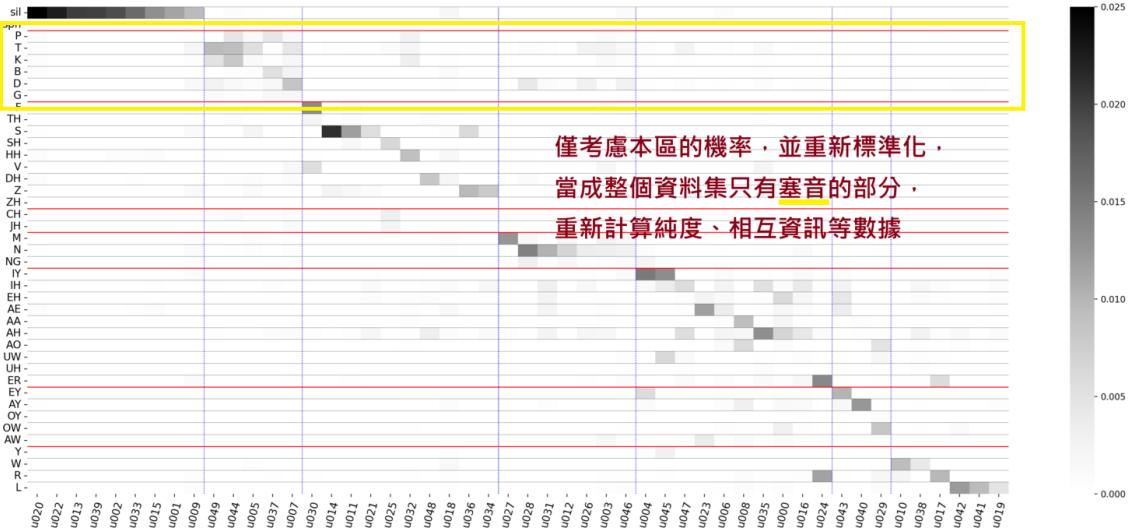


圖 3.13: 對共同機率分佈按照音位分類分別計算純度作法示意圖

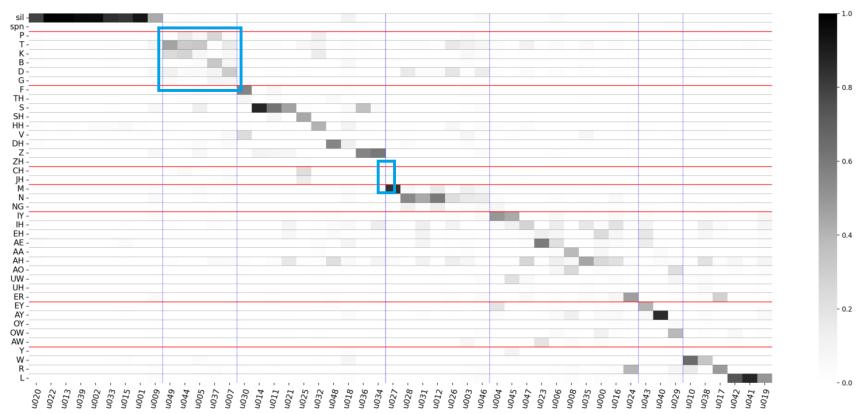
(以塞音舉例，同樣的作法對紅線分開的八塊區域分別計算)

3.6.4 整體熱圖驗證

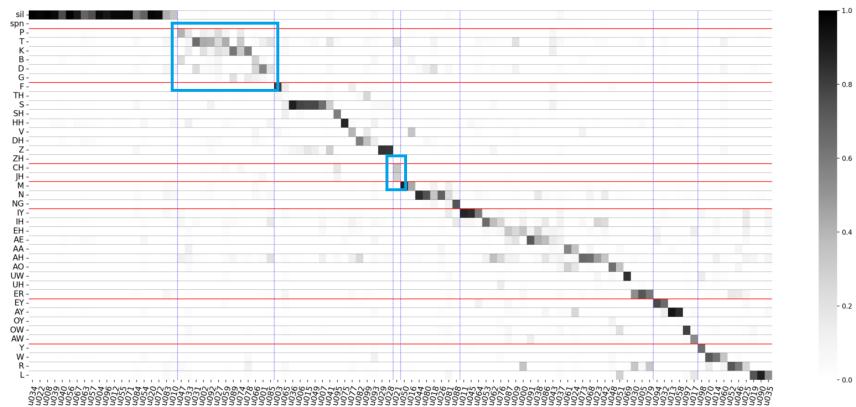
最後，參考韋氏（Wells）[53] 的研究方法，我們探討不同音位分類中音位與離散單元的對應關係。比起直接看離散單元的編號，我們改對機率熱圖進行分區觀察以確認趨勢。為了同時對比語音表徵與分群數兩個變因造成的差異，我們比較 HuBERT 分群數 50 和 100 以及 CPC 分群數 100 的機率熱圖，並參考 SpeechTokenizer [44] 和 DinoSR [50] 的作法，以 $p_{y|z}(i|j)$ 呈現，確認離散表徵對於音位的歸類效果，最終驗證前述觀察。圖 3.15 中框出的區域為前述觀察到較為分散的塞音與塞擦音，這些區域確實顏色較淺（塞擦音在 HuBERT + 分群數 = 50 和 CPC + 分群數 = 100 甚至可能沒有專門的離散單元以其為代表），證明其語音特徵歸類的確較為困難；而圖 3.16 中框出的區域則是離散表徵歸類較集中的擦音、雙元音與近音，這幾區的色塊如前面推論所預測的較為明顯，屬於容易區分的音位分類。

語音表徵	分群數	XXX 音位純度	塞音 音位純度	擦音 音位純度	塞擦音 音位純度	鼻音 音位純度	單元音 音位純度	雙元音 音位純度	近音 音位純度
HuBERT	50	0.9924	0.4744	0.7033	0.6616	0.7580	0.5222	0.7813	0.8658
HuBERT	100	0.9943	0.5480	0.7535	0.6917	0.8447	0.6306	0.8273	0.8952
HuBERT	200	0.9945	0.6330	0.7711	0.6656	0.8721	0.6663	0.8707	0.9287
CPC	50	0.9930	0.5276	0.6039	0.6057	0.8594	0.5489	0.7542	0.8426
CPC	100	0.9935	0.6385	0.7466	0.6107	0.8807	0.5683	0.7960	0.8649
CPC	200	0.9941	0.6819	0.7632	0.6748	0.8900	0.5906	0.8079	0.8792
Wav2vec 2.0	50	0.9921	0.4089	0.5314	0.6372	0.6156	0.4649	0.6458	0.6642
Wav2vec 2.0	100	0.9941	0.4890	0.6096	0.6822	0.6477	0.5205	0.7595	0.7933
Wav2vec 2.0	200	0.9946	0.4989	0.6716	0.7022	0.7024	0.6129	0.7970	0.8305
LogMel	50	0.9903	0.4149	0.4507	0.6728	0.5891	0.3600	0.5817	0.6221
LogMel	100	0.9904	0.4183	0.4766	0.6768	0.5907	0.3721	0.5959	0.6317
LogMel	200	0.9911	0.4255	0.4917	0.6862	0.5944	0.3932	0.6142	0.6709
語音表徵	分群數	XXX 分群純度	塞音 分群純度	擦音 分群純度	塞擦音 分群純度	鼻音 分群純度	單元音 分群純度	雙元音 分群純度	近音 分群純度
HuBERT	50	0.1733	0.2621	0.4688	0.4760	0.3633	0.3252	0.4996	0.4130
HuBERT	100	0.0855	0.1936	0.3591	0.3197	0.3325	0.2507	0.3978	0.3147
HuBERT	200	0.0461	0.1470	0.2220	0.2765	0.1927	0.1651	0.2872	0.1749
CPC	50	0.3927	0.2673	0.4025	0.4337	0.5423	0.3420	0.4327	0.4270
CPC	100	0.1389	0.2506	0.3742	0.3604	0.2951	0.2476	0.4118	0.2553
CPC	200	0.1051	0.1741	0.2612	0.3268	0.1911	0.1340	0.3122	0.1784
Wav2vec 2.0	50	0.1541	0.2152	0.3467	0.3245	0.3313	0.2611	0.3446	0.3197
Wav2vec 2.0	100	0.1253	0.1505	0.2788	0.3040	0.1855	0.2062	0.3919	0.2433
Wav2vec 2.0	200	0.0826	0.1150	0.1760	0.2547	0.1396	0.1370	0.2810	0.1847
LogMel	50	0.1660	0.0887	0.1657	0.1597	0.1614	0.1324	0.1682	0.1883
LogMel	100	0.1099	0.0555	0.1242	0.1381	0.0810	0.0773	0.0828	0.1271
LogMel	200	0.0802	0.0347	0.0693	0.0752	0.0476	0.0459	0.0527	0.0687
語音表徵	分群數	XXX PNMI	塞音 PNMI	擦音 PNMI	塞擦音 PNMI	鼻音 PNMI	單元音 PNMI	雙元音 PNMI	近音 PNMI
HuBERT	50	0.8295	0.2082	0.5596	0.1065	0.3512	0.3961	0.5683	0.7110
HuBERT	100	0.8672	0.3351	0.6318	0.1998	0.5609	0.5168	0.6667	0.7734
HuBERT	200	0.8737	0.4563	0.6691	0.1597	0.6223	0.5657	0.7396	0.8326
CPC	50	0.8288	0.2999	0.4508	0.0252	0.5635	0.4201	0.5438	0.6551
CPC	100	0.8454	0.4477	0.5982	0.0412	0.6136	0.4431	0.5936	0.7085
CPC	200	0.8588	0.5038	0.6404	0.1399	0.6520	0.4799	0.6298	0.7331
Wav2vec 2.0	50	0.8134	0.1086	0.3351	0.0706	0.0959	0.3019	0.3844	0.3738
Wav2vec 2.0	100	0.8457	0.2245	0.4279	0.1375	0.1419	0.3880	0.5779	0.5806
Wav2vec 2.0	200	0.8578	0.2479	0.5130	0.1743	0.2601	0.4956	0.6346	0.6495
LogMel	50	0.6971	0.0918	0.2402	0.0832	0.0272	0.1963	0.2483	0.2941
LogMel	100	0.7104	0.1014	0.2655	0.0982	0.0328	0.2160	0.2757	0.3239
LogMel	200	0.7314	0.1112	0.2888	0.1108	0.0394	0.2389	0.2998	0.3647

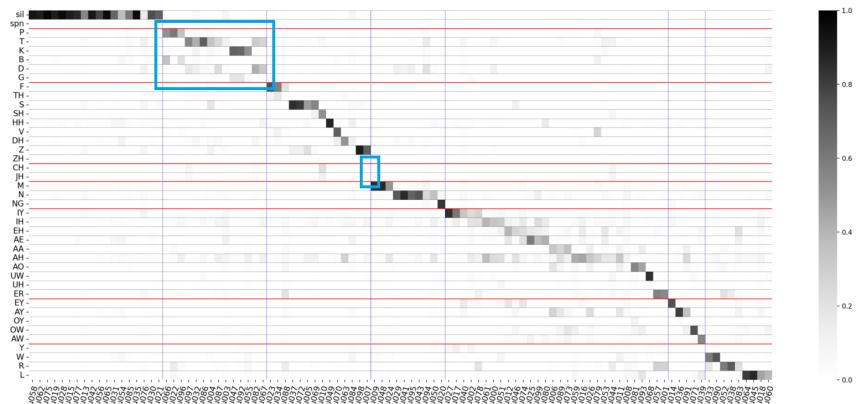
圖 3.14: 按照音位分類分開各自計算的純度與相互資訊



(a) HuBERT + 分群數 = 50



(b) HuBERT + 分群數 = 100

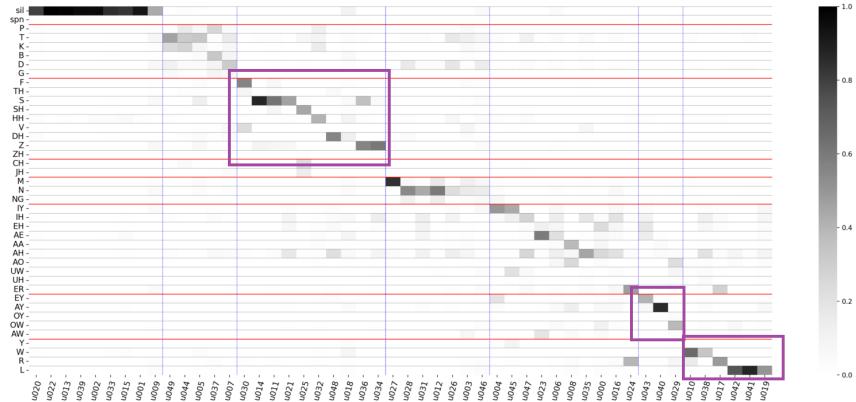


(c) CPC + 分群數 = 100

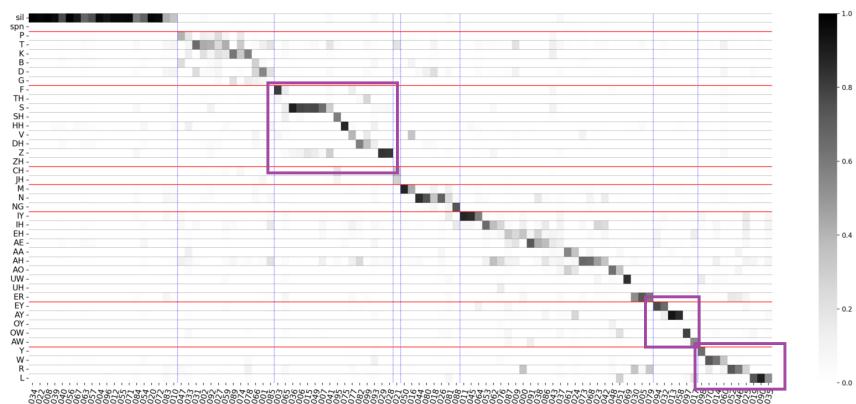
圖 3.15: 热圖驗證塞音、塞擦音較難以被離散單元歸類，

注意塞擦音在 HuBERT + 分群數 = 50 和 CPC + 分群數 = 100

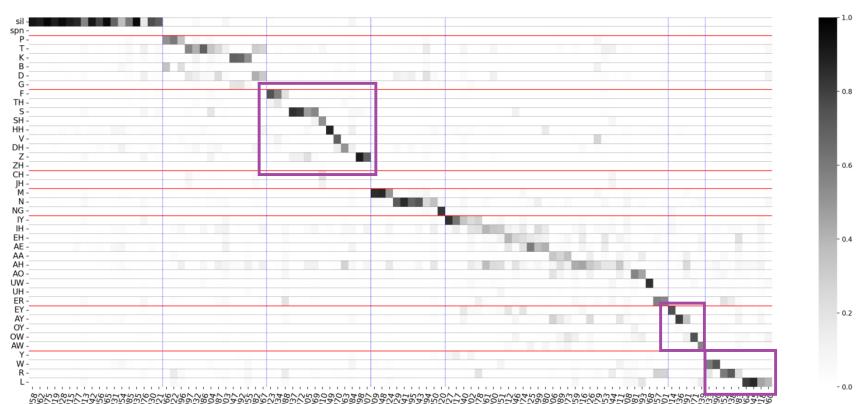
甚至可能沒有專門的離散單元以其為代表



(a) HuBERT + 分群數 = 50



(b) HuBERT + 分群數 = 100



(c) CPC + 分群數 = 100

圖 3.16: 热圖驗證擦音、雙元音與近音的特徵較明顯

3.7 本章總結

本章節探討以音框為單位取出的語音離散表徵與對應的音位標註之間的關係。我們從純度的計算開始，對整個機率熱圖進行可視化分析，並透過語音知識的協助，尤其是對音位的分類，將原先約 40 類的獨立標註進行更深入的特性分析。

透過這些探討，從語音學的角度，我們發現塞音的離散單元分佈較為分散，而擦音則較為集中。同時，藉由比對不同模型間的數據表現差異，也確認了 HuBERT 模型的離散表徵在各項數據中與音位之間的相似性最為明顯。因此，進一步印證了 HuBERT 為何是抽取語音離散表徵時最常使用的模型，並常被無文字架構所使用。

然而，單一離散表徵僅能代表 10 或 20 毫秒的語音訊號，而音位的長度經常佔據不只一個離散表徵。因此，下一章節將嘗試進一步組合多個離散表徵成為符記，分析它們與音位之間的關係。

第四章 多個語音離散表徵與音位的關係

4.1 動機

如第三章所述，一個文字或音位往往對應到上百毫秒的語音訊號，然而單一離散單元所對應的聲音訊號為 10 或 20 毫秒，亦即同一段語音所對應的離散單元數目將比音位或文字多出許多。本章節從自然語言處理中獲取靈感，將文字處理中的分詞演算法（Tokenization）應用於離散單元序列上，使得離散單元重新組合成次詞單位（Subword Units），稱之為「聲學片段（Acoustic Piece）」，以這些由多個離散單元組成的符記（Token）¹作為新的基本單位重新編碼語音訊號，取代原先的離散單元。為了分析聲學片段是否更接近音位的序列，在此將續用上一章節的分析方法，比對並檢驗引入次詞單位是否有機會得到更好的語音表徵，進而有機會用於無文字（Textless）架構 [45, 5, 42] 中。

4.2 相關研究

在無文字架構被提出後的約兩年後，藉助次詞單位組合離散單元的研究逐步出現。任氏（Ren）等人 [11] 最先提出聲學片段的概念，該論文比對離散單元序列及對應的文字轉寫，從中觀察到許多相似的規律重複出現，而且不限於單一語者。受此啟發，本論文首先將離散單元，透過文字處理中常用以獲得次詞單位的「句片段（SentencePiece）」[58] 套件獲得新的符記——「聲學片段」，並用於語音辨識的預訓練上。

不久，由吳氏（Wu）提出的 Wav2seq [10] 論文中，考量文字與語音的序列長

¹指資料序列中的離散基本單位。

度差異，並基於離散單元和音位的關聯性，將離散單元視為字符（Character），嘗試將這些字符透過次詞單位組成「虛擬語言（Pseudo-language²）」，來幫助語音到文字的模型。在實際應用中，因為解碼器生成的目標文字序列亦是由次詞單位組成，因此該篇研究旨在讓模型在預訓練後可以快速適應下游任務。與前一篇呼應，聲學片段對語音預訓練的效果在 [59] 中被探討，此後聲學片段更被應用於縮短資料序列長度 [49]、語音生成 [60]，或學習更穩健（Robust）的語音表徵 [61]。

近期，張氏（Chang）等人 [12] 將以分詞方法處理離散單元的流程（Pipeline）納入 ESPNet 套件 [62] 中，並在語音辨識、語音翻譯等任務中獲得了超越以往的表現，進一步證明了這個方法的效果。

4.3 文字處理中的分詞演算法

在以文字為主體的自然語言處理中，文字文本除了以單詞（Word）或字元（Character）為處理單位，更常見的作法是透過分詞演算法（Tokenization）將文本分段，以「次詞單位」構成詞彙表來重新編碼文本，用於文字模型的訓練與推論。

使用次詞單位的優點包含：

1. 固定詞彙表大小，避免未登錄詞（Out-of-vocabulary，OOV）。
2. 縮短資料序列的長度，提升訓練和推論的效率。
3. 分解單詞，捕捉更細緻的語意關係，模擬如英語中的字首（Prefix）、字尾（Suffix）等等具有特定意義的文字組合。

²偽語言對應之離散單元被視為「虛擬文字（Pseudo-text）」

4.3.1 常見演算法

以下介紹幾種常見的分詞方法：

位元組對編碼（Byte Pair Encoding，BPE）

位元組對編碼 [63, 64] 是一種常用的分詞方法，最初來自資料壓縮技術 [63]，後來被引入到自然語言處理領域，用以處理機器翻譯問題 [64]。該演算法從字元開始，根據詞彙表中各個次詞單位的頻率，反覆合併常見的字元成為新的次詞單位，直到達到預定的詞彙表大小。

單詞片段（WordPiece）

WordPiece [65] 演算法由 Google 用以訓練機器翻譯系統，並在 BERT [28] 模型中被使用而廣為人知。與位元組對編碼相似，同樣是透過反覆合併的策略，但合併的依據改以機率模型取代出現頻率。

單一詞語言模型（Unigram Language Model）

單一詞語言模型 [66] 是基於語言模型的分詞方法，以機率分佈選擇次詞單位，並以最大化輸入文本的機率來為文本分段。

4.3.2 「句片段（SentencePiece）」套件

「句片段（SentencePiece）[58]」是由 Google 開發的分詞套件，實作了前述的位元組對編碼和單一詞演算法。其優勢在於可應用於不同語言，尤其用於處理中文、日文等不使用空格分隔單詞的語言文本時，此套件大大的簡化了前處理的流程。考慮到語音訊號本身不如英語等文字，在書寫時就已經具備空格分隔單詞，因此本章節的所有次詞單位皆以句片段套件中的單一詞演算法取得。

4.4 分析方式

本章節沿用上一章節 LibriSpeech 資料集的 train-clean-100 訓練子集，以單一詞演算法取得次詞單位，並嘗試 500、1000、8000、10000、20000 五種符記種數，對每一種語音表徵和 K-平均模型的分群數，各自取得五種聲學片段文本。

比照第三章的分析方式，本章除了整體的純度與相互資訊數據外，亦同樣從聲學片段與音位的角度分別探討，藉由調整次詞單位的種類數量，探討引入次詞單位並改變符記數量，將如何影響這些符記序列與音位標註間的相關性。然而，為了避免結果呈現過於複雜，細部分析時將著重比對 500 和 1000 種次詞單位的結果變化。

由於本章節探討重點為次詞單位種數變化的影響，延續第三章的發現，後續分析將以表現最好的 HuBERT 離散表徵為主。在需要比較離散單元分群數影響時，我們將比對分群數為 50 與 100 時的差異，否則為避免數據過於複雜，離散單元的分群數預設為 50 進行細部探討。

4.5 分析結果

承繼上一個章節的分析方法，我們先將純度等數據與條件機率熱圖 $p_{y|z}(i|j)$ ³ 兩者互相對照，並以語音學排序呈現，觀察聲學片段與音位之間兩者的分佈關係。

³ 由於共同機率分佈熱圖 p_{yz} 的數值對於觀察符記對應到音位的關係較不明顯，因此仿照 SpeechTokenizer [44]、DinoSR [50] 等論文使用 $p_{y|z}(i|j)$ 呈現。

4.5.1 由聲學片段角度探討

聲學片段數量的影響

次詞單位種數	音位純度	分群純度	音位熵	離散單元熵	PNMI
離散單元	0.5256	0.3382	3.3152	3.8681	0.4993
500	0.5574	0.0829	3.3152	6.0282	0.5357
1000	0.5744	0.0556	3.3152	6.6594	0.5466
8000	0.5957	0.0257	3.3152	8.5192	0.5729
10000	0.5955	0.0238	3.3152	8.7207	0.5750
20000	0.5921	0.0182	3.3152	9.3527	0.5820

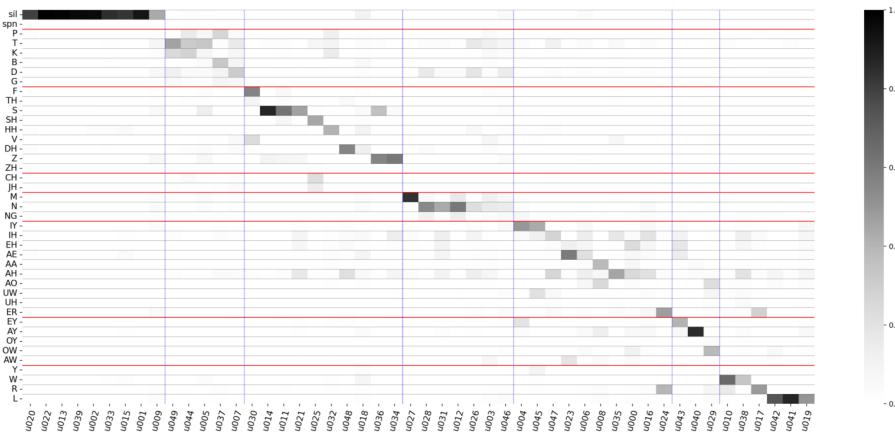
(a) 分群數 = 50

次詞單位種數	音位純度	分群純度	音位熵	離散單元熵	PNMI
離散單元	0.6097	0.2553	3.3152	4.5704	0.5786
500	0.6260	0.0972	3.3152	6.0655	0.5990
1000	0.6372	0.0631	3.3152	6.7181	0.6089
8000	0.6536	0.0237	3.3152	8.5954	0.6308
10000	0.6527	0.0219	3.3152	8.7938	0.6324
20000	0.6490	0.0173	3.3152	9.4123	0.6378

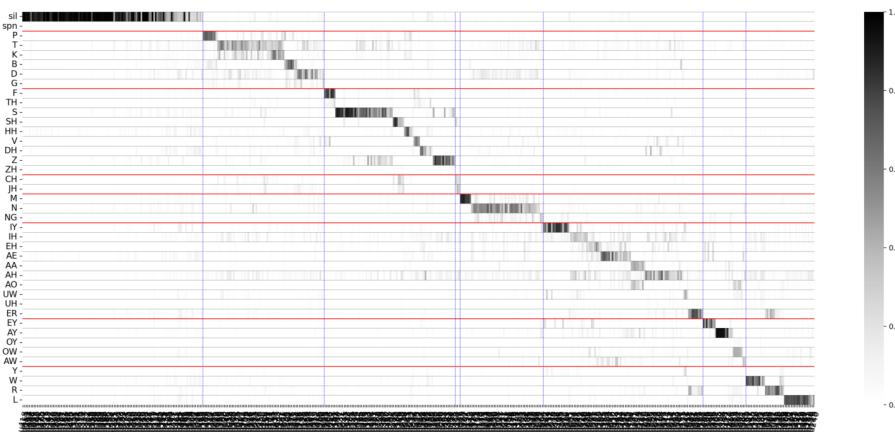
(b) 分群數 = 100

表 4.1: HuBERT 模型在不同次詞單位種類數量時的純度分析數據

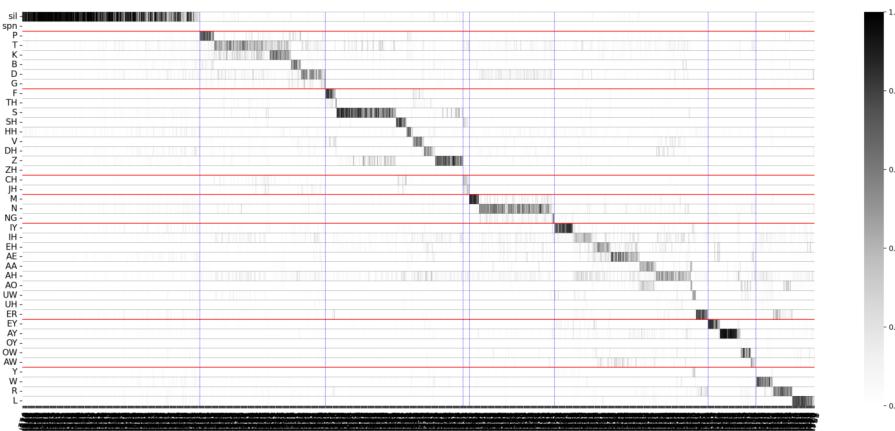
表 4.1 是 HuBERT 模型透過離散單元與不同次詞單位數量之聲學片段的純度與相互資訊數據。首先，為了觀察聲學片段數量對於機率熱圖與純度數據的影響，圖 4.1 與圖 4.2 分別以 HuBERT 表徵、分群數為 50 和 100 的離散單元為基礎，比較原始離散單元、500 和 1000 種次詞單位三種設定下，不同聲學片段數量的條件機率熱圖。從中我們可以看出，當聲學片段數量上升時，熱圖可以觀察出許多更深的色塊，也就是有更多的聲學片段可以更集中的對應到特定音位。由此可見，



(a) 紛散單元



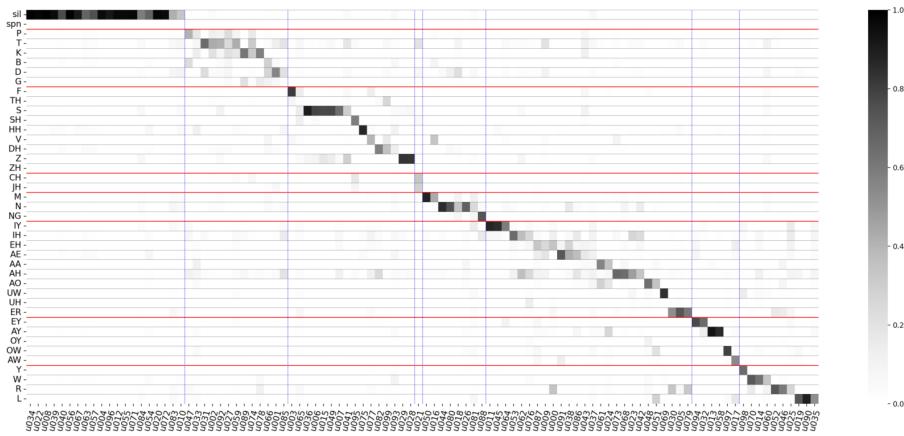
(b) 500 種次詞單位



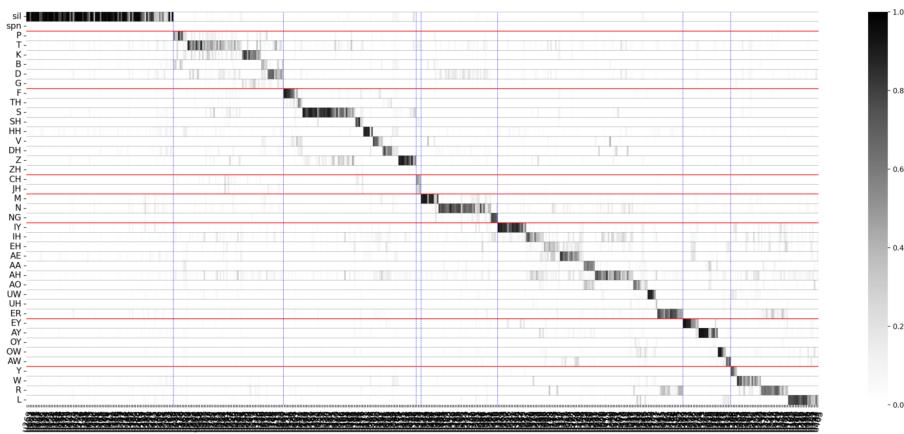
(c) 1000 種次詞單位

圖 4.1: HuBERT 表徵在 K-平均演算法使用分群數 50 後，

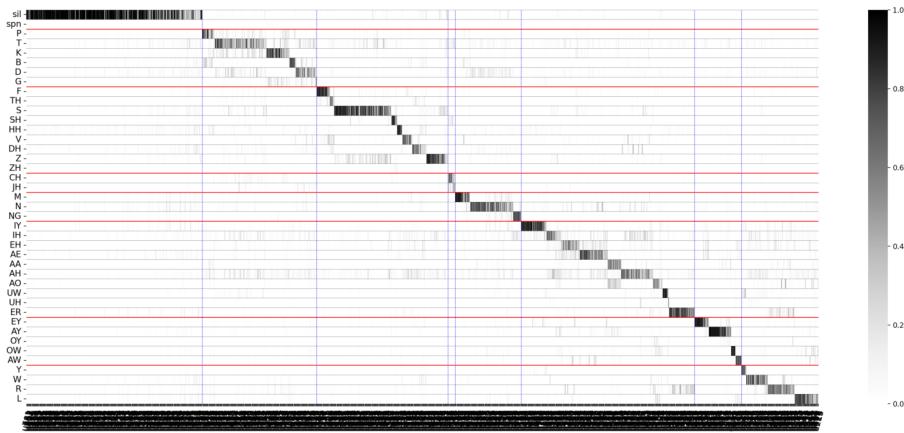
比較不同次詞單位數量的條件機率分佈 $p_{y|z}(i|j)$ 热圖



(a) 離散單元



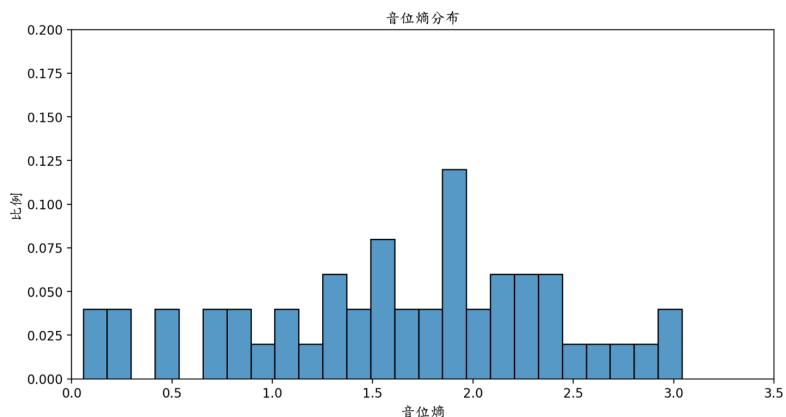
(b) 500 種次詞單位



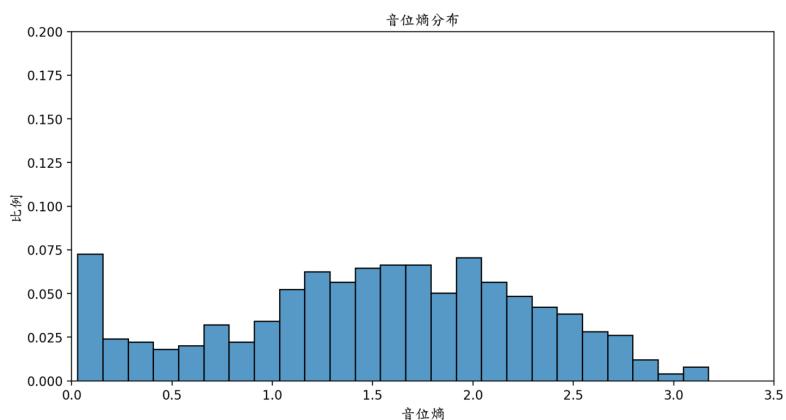
(c) 1000 種次詞單位

圖 4.2: HuBERT 表徵在 K-平均演算法使用分群數 100 後，

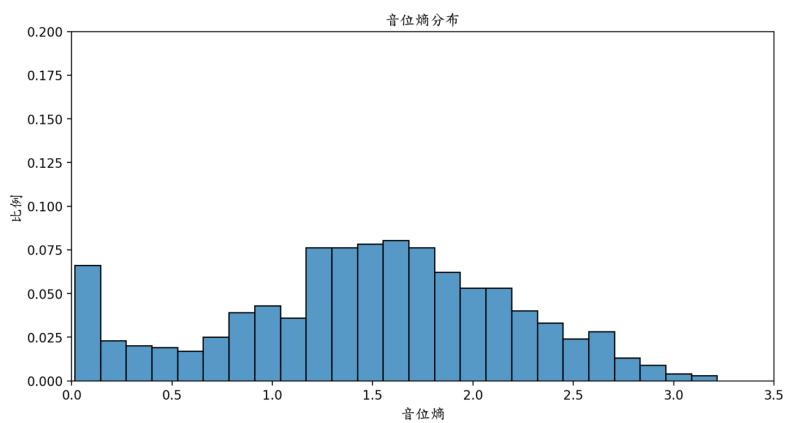
比較不同次詞單位數量的條件機率分佈 $p_{y|z}(i|j)$ 熱圖



(a) 紛散單元

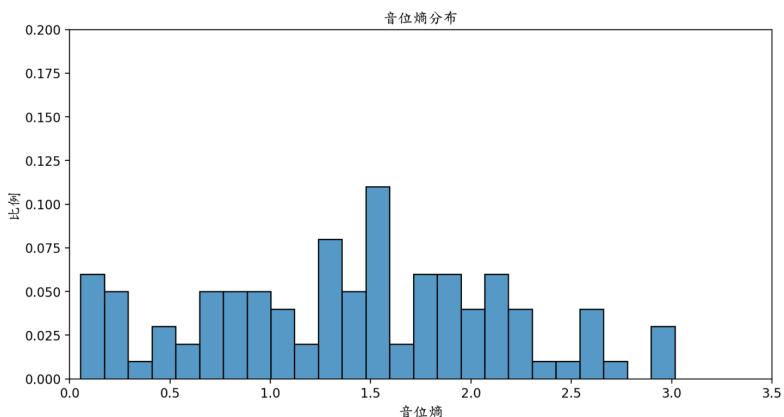


(b) 500 種次詞單位

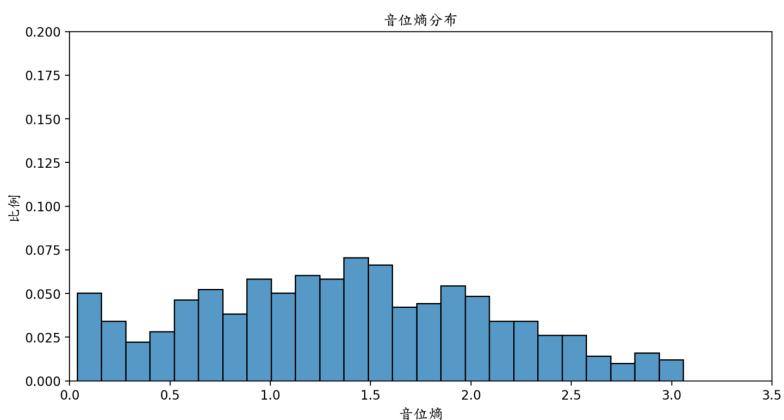


(c) 1000 種次詞單位

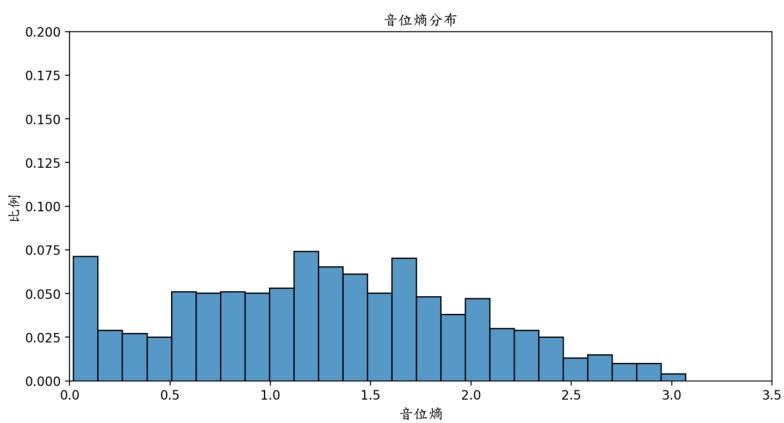
圖 4.3: HuBERT 表徵在 K-平均演算法使用分群數 50 後，
比較不同次詞單位數量的音位條件熵 $H(y|z)$ 直方圖



(a) 紛散單元



(b) 500 種次詞單位



(c) 1000 種次詞單位

圖 4.4: HuBERT 表徵在 K-平均演算法使用分群數 100 後，
比較不同次詞單位數量的音位條件熵 $H(y|z)$ 直方圖

有了更多樣的符記可以區別出更細節的發音差異，使整體的純度數值有所提升；然而，機率熱圖整體也變得更加破碎，因此歸類同樣音位的效果也相對變得較不明顯。

為了確認各自聲學片段對應音位之集中狀況，我們可以考慮這些機率熱圖的條件音位熵 $H(y|z)$ ，以直方圖呈現來確認變化。透過觀察圖 4.3 與圖 4.4 的結果，可以確認相比第三章的離散單元，引入次詞單位確實能降低整體的條件音位熵，亦即新的符記各自能夠有更明確對應的音位，與我們從機率熱圖上所觀察到的趨勢符合。

雖然改用聲學片段會使熱圖更加破碎而複雜，但除純度與相互資訊的數值變化外，觀察每個聲學片段對應之最高機率音位 $i^*(j)$ 以及它們的音位分類比例變化，也可以驗證「更多符記可以區別發音細節差異」這點。再次觀察 HuBERT 在分群數 50 時的機率熱圖（圖 4.1），圖 4.1 中三章熱圖的藍色鉛直線是每個符記在找出對應音位 $i^*(j)$ 後，按音位分類分區排序的結果。因此，比較藍色鉛直線在橫軸上各區的比例變化，可以知道有多少比例的符記能表示特定類型的發音。第三章結尾時提及過，在離散單元分群數為 50 時，由於符記數量較少，並沒有任何單元最能直接對應塞擦音音位。然而，當將這些離散單元以次詞單位進行重組後，不管在新符記種數為 500 或 1000 的機率熱圖上，都可以發現至少出現一個以上的符記得以對應到塞擦音。由此，我們驗證了引入次詞單位，對捕捉更細微的發音差異的確有所幫助。

離散單元分群數對聲學片段表現的影響

然而，儘管引入次詞單位一定程度上能幫助區別語音訊號中的細微發音差異，在語音表徵進行離散化時，K-平均演算法的分群數仍是決定這些符記捕捉語

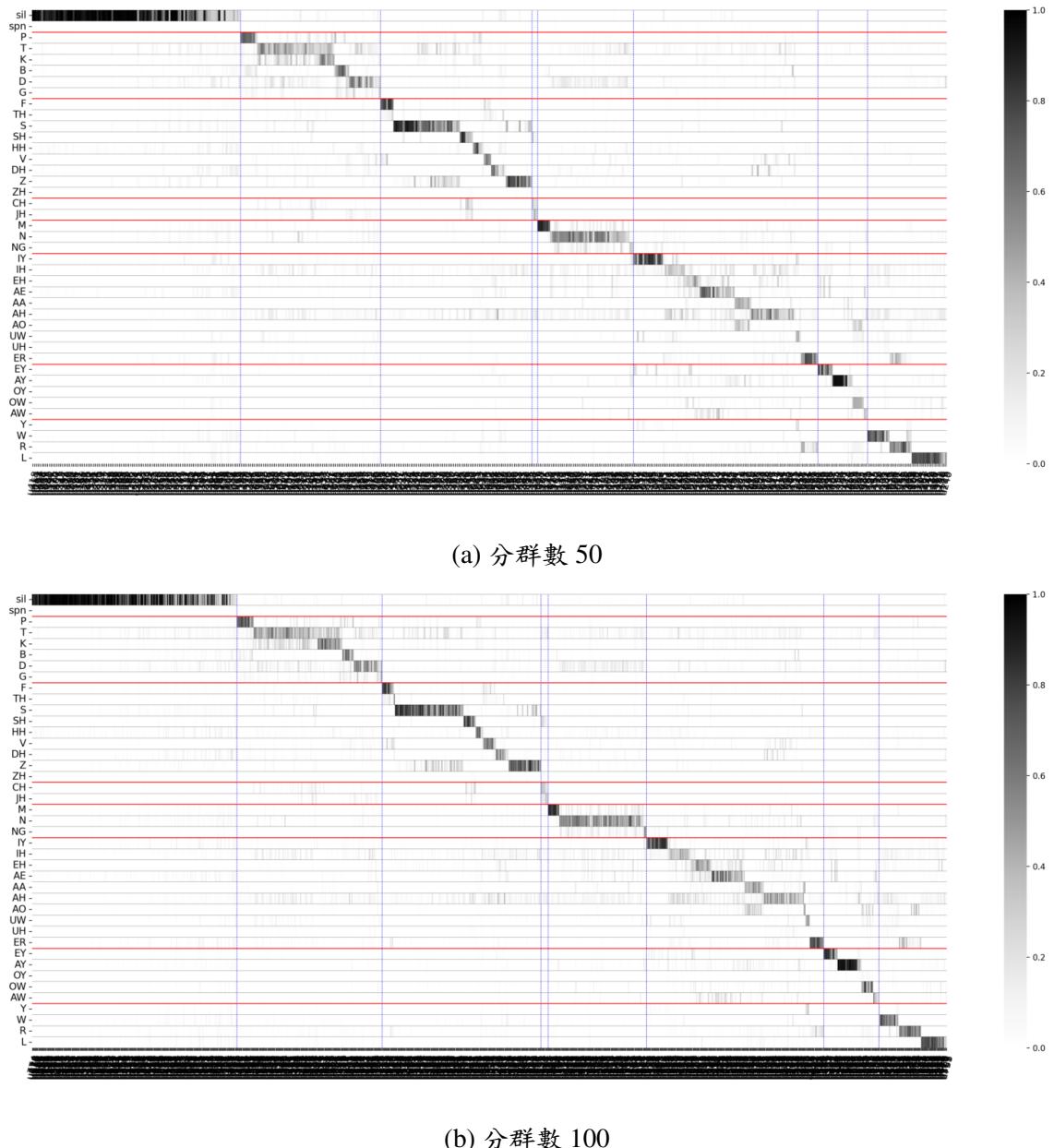


圖 4.5: 比較同樣 500 種次詞單位的聲學片段模型，著重比較 HuBERT 表徵

在 K-平均演算法使用分群數 50 與 100 的條件機率熱圖 $p_{y|z}(i|j)$ 差異

音資訊更關鍵的決定因素。圖 4.5 比較了同樣是 500 種次詞單位，K-平均演算法的離散表徵分群數選擇 50 和 100 的機率熱圖差異，不難發現分群數為 100 的機率熱圖能更加平均的對應到不同音位。然而，即便與音位的對應效果最大取決於 K-平均的分群數，但分群演算法本身相當消耗計算資源。因此當遇到運算資源限制，致使 K-平均演算法的分群數難以設置得很大時，次詞單位的引入仍舊能提升整體表現。

聲學片段對應最高機率之音位間的比較

接下來，我們比較各個聲學片段與音位之間的對應關係，亦即每個符記所對應最可能的前幾個音位之間，是否依然如離散單元那樣存在特定特徵。觀察以 HuBERT 模型、分群數 50 為基礎，分別以「離散單元」與「500 種次詞單位的聲學片段」為符記的虛擬文字文本，將對應到塞音、擦音和單元音部分的次詞單位取出觀察，將每個符記對應前五高機率的音位排名呈現在圖 4.6 中（並附上最高機率音位 $i^*(j)$ 的條件機率值 $p_{y|z}(i^*(j)|j)$ ），圖中上半部是離散單元，下半部則是聲學片段的結果。相互比較後可以發現，由於聲學片段的符記數量比離散單元更多，因此在維持對應音位之間相關性的同時，卻能呈現出不同音位間更細節的相關性。例如在圖 4.6a 中，上半部顯示原先以離散單元為符記時，因為只有 50 種符記，因此只能看出 T、B 與 D 比較容易和哪些其他音位比較相關，但聲學片段卻可以呈現出 P、T、B、D 等更多細節的音位關係。特別值得注意的是，圖 4.7 是對應到塞擦音的幾個聲學片段，這些對應到 CH 和 JH 兩種塞擦音的聲學片段也確實給予了同樣是塞擦音的其他音位較高的機率。

仿照第三章，藉由以音位分類作為新的標註計算純度，我們可以確認聲學片段給予同類音位較高機率的效果。然而從表 4.2 可以發現，隨著符記種數的提升，

	u049	u044	u005	u037	u007
rank 1	T	T	T	B	D
	45.88 %	31.37 %	33.04 %	32.67 %	30.73 %
rank 2	K	K	S	P	T
rank 3	D	P	K	D	B
rank 4	G	D	P	G	G
rank 5	AH	CH	Z	AH	K

	rank 1	rank 2	rank 3	rank 4	rank 5
u181	P 60.58 %	T K D B			
u308	P 71.21 %	T D K B			
u478	P 71.28 %	B T AH D			
u423	P 64.55 %	B G AH D			
u314	P 64.92 %	B T K G			
u346	P 67.61 %	B G AH D			
u401	P 56.57 %	G D B T			
u219	P 61.56 %	T B K G			
u060	P 25.55 %	B D AH DH			
u069	T 38.81 %	K CH P D			
u159	T 43.81 %	K HH EH AH			
u302	T 63.11 %	IH AH UW K			
u200	T 57.41 %	IH UW AH D			
u128	T 36.80 %	K D CH G			
u044	T 26.70 %	Z D S K			
u252	T 48.80 %	AH IH K UW			
u059	T 32.70 %	D AH IH K			
u028	T 23.10 %	D N AH IH			
u400	T 57.30 %	K S D Z			
u373	T 42.52 %	K AH IH UW			
u335	T 47.48 %	K EH HH EY			
u070	T 35.97 %	K CH D P			

	rank 1	rank 2	rank 3	rank 4	rank 5
u035	B 36.05 %	P G AH IH			
u204	B 46.17 %	G D AH V			
u394	B 63.72 %	D G AH V			
u196	B 61.52 %	G D AH P			
u371	B 75.12 %	AH G IH P			
u215	B 57.65 %	P IH G R			
u304	B 24.82 %	D G JH V			
u316	B 33.74 %	D G V AH			
u036	D 49.74 %	T JH G DH			
u132	D 63.13 %	JH G T V			
u295	D 56.94 %	G AH IH T			
u123	D 28.65 %	T G P JH			
u179	D 39.80 %	T IH AH HH			
u369	D 67.06 %	T G HH AH			
u157	D 45.86 %	T HH AH IH			
u009	D 14.40 %	T N HH Z			
u167	D 32.12 %	T AH IH V			
u459	D 55.26 %	G JH AH B			
u012	D 19.69 %	T V DH N			
u312	D 47.99 %	T JH AE IH			
u375	D 42.09 %	IH G T AH			
u050	D 15.14 %	N T sil M			

(a) 塞音

圖 4.6: HuBERT 表徵、K-平均演算法分群數 50，比較單一離散單元與使用 500 種次詞單位，依據不同音位分類比較符記各自對應的前五高音位上半部為離散單元，下半部為聲學片段。

圖中的百分比為最高機率音位的條件機率 $p_{y|z}(i^*(j)|j)$

	u030	u014	u011	u021	u025	u032	u048	u018	u036	u034
rank 1	F	S	S	S	SH	HH	DH	DH	Z	Z
	56.54 %	87.39 %	62.24 %	46.31 %	44.75 %	40.85 %	56.04 %	12.11 %	56.40 %	60.86 %
rank 2	V	Z	SH	AH	CH	K	AH	sil	S	IH
rank 3	TH	T	Z	IH	JH	P	TH	HH	T	AH
rank 4	AH	spn	F	Z	T	T	EH	AE	AH	S
rank 5	R	HH	TH	EH	S	AE	IH	W	HH	EH

	rank 1		rank 2	rank 3	rank 4	rank 5
u093	F	67.74 %	V	TH	R	AH
u168	F	77.92 %	V	TH	R	AH
u306	F	82.32 %	TH	V	R	sil
u053	F	46.58 %	V	TH	AH	R
u344	F	84.94 %	TH	AH	R	V
u251	F	81.48 %	TH	AH	R	spn
u407	F	64.31 %	TH	AH	HH	spn
u180	S	84.29 %	AH	IH	EH	Z
u277	S	90.94 %	Z	AH	T	spn
u331	S	86.23 %	AH	IH	EH	Z
u241	S	81.50 %	AH	IH	EH	Z
u235	S	83.94 %	AH	EH	IH	Z
u047	S	54.86 %	SH	TH	F	Z
u223	S	87.27 %	Z	T	AH	spn
u476	S	93.37 %	Z	T	spn	D
u225	S	86.81 %	Z	T	D	TH
u391	S	85.69 %	Z	EH	AH	IH
u328	S	86.29 %	IH	AH	Z	EH
u147	S	60.49 %	Z	T	HH	spn
u329	S	89.34 %	Z	T	AH	spn
u434	S	61.40 %	T	P	K	Z
u015	S	41.05 %	SH	Z	TH	F
u399	S	62.53 %	T	Z	K	P

	rank 1		rank 2	rank 3	rank 4	rank 5
u023	HH	46.20 %	P	K	T	AE
u085	HH	63.62 %	P	K	T	AE
u243	HH	78.42 %	P	K	AE	EH
u032	HH	33.33 %	P	K	T	AA
u037	HH	26.92 %	P	K	R	T
u224	HH	16.69 %	sil	DH	W	AE
u030	V	44.92 %	F	TH	AH	HH
u332	V	62.37 %	AH	TH	IH	F
u422	V	53.33 %	F	AH	IH	TH
u054	V	46.19 %	TH	F	DH	AH
u043	DH	59.91 %	TH	EH	IH	AH
u106	DH	64.70 %	EH	AH	IH	EY
u208	DH	39.38 %	AH	EH	AE	TH
u311	DH	42.62 %	AH	ER	IH	IY
u097	DH	14.43 %	HH	sil	W	N
u150	DH	18.01 %	sil	HH	W	AE
u088	DH	31.73 %	TH	AH	IH	sil
u102	DH	11.75 %	sil	HH	AE	W
u100	Z	52.66 %	S	T	HH	spn
u263	Z	85.20 %	AH	IH	S	HH
u256	Z	74.83 %	IH	AH	EH	ER
u390	Z	75.49 %	IH	AH	EH	S
u299	Z	75.65 %	AH	IH	S	EH

(b) 擦音

	u004	u045	u047	u023	u006	u008	u035	u000	u016	u024
rank 1	IY	IY	IH	AE	AE	AA	AH	AH	AH	ER
	49.75 %	43.68 %	27.04 %	59.57 %	21.24 %	38.63 %	45.05 %	25.15 %	21.04 %	47.05 %
rank 2	EY	UW	AH	AW	IH	AO	IH	EH	IH	R
rank 3	NG	IH	T	EH	AH	AY	V	OW	EH	AH
rank 4	IH	Y	UW	AH	EH	AH	T	AA	ER	IH
rank 5	AY	HH	ER	AA	AO	AW	ER	AE	IY	HH

	rank 1		rank 2	rank 3	rank 4	rank 5
u087	IY	55.06 %	EY	IH	NG	N
u164	IY	65.67 %	EY	IH	NG	N
u049	IY	36.42 %	UW	Y	IH	HH
u151	IY	81.23 %	IH	HH	Y	N
u236	IY	83.38 %	IH	Y	N	HH
u038	IY	38.67 %	UW	IH	Y	UH
u212	IY	79.52 %	IH	Y	NG	HH
u326	IY	70.32 %	EY	IH	NG	N
u444	IY	87.01 %	IH	HH	N	Y
u305	IY	82.55 %	IH	HH	Y	D
u429	IY	84.49 %	IH	N	Y	HH
u182	IY	73.72 %	IH	NG	Y	N
u075	IY	33.53 %	EY	NG	AY	IH
u216	IY	76.32 %	IH	Y	NG	HH
u334	IY	79.76 %	IH	HH	Y	UW
u385	IY	58.00 %	NG	IH	Y	EY
u231	IY	15.58 %	ER	sil	D	L
u004	IH	32.68 %	AH	T	ER	EH
u074	IH	22.11 %	AH	EH	AY	N
u045	IH	35.75 %	AH	T	ER	HH
u046	IH	36.34 %	AH	ER	HH	N
u041	IH	30.11 %	IY	UW	Y	AH
u162	IH	26.01 %	AH	EH	AY	N

	rank 1		rank 2	rank 3	rank 4	rank 5
u121	AA	39.28 %	AO	AY	R	L
u171	AA	41.48 %	AO	AY	L	R
u283	AA	45.77 %	AO	AY	AH	L
u249	AA	39.58 %	AO	N	AH	T
u345	AA	40.42 %	AO	N	AH	T
u117	AA	31.29 %	AY	AO	AH	R
u260	AA	33.15 %	AH	N	AO	T
u442	AA	43.54 %	AO	AY	L	R
u120	AA	24.49 %	AH	AY	R	AO
u057	AH	58.87 %	DH	IH	IY	T
u010	AH	41.37 %	IH	V	T	ER
u005	AH	38.66 %	IH	ER	T	V
u193	AH	48.91 %	DH	IH	T	ER
u003	AH	22.09 %	EH	AY	AE	OW
u279	AH	51.51 %	EH	V	DH	M
u042	AH	22.60 %	IH	ER	N	EH
u079	AH	26.20 %	V	IH	T	B
u218	AH	45.60 %	M	EH	V	N
u194	AH	51.09 %	DH	ER	IH	IY
u297	AH	70.86 %	DH	IH	IY	M
u029	AH	20.74 %	IH	ER	IY	EH
u144	AH	42.31 %	IH	ER	V	T
u211	AH	48.92 %	V	IH	T	TH

(c) 單元音

	rank 1		rank 2	rank 3	rank 4	rank 5
u414	CH	33.07 %	SH	JH	AH	ZH
u327	JH	41.40 %	CH	AH	SH	T
u439	JH	26.49 %	CH	AH	IH	SH

圖 4.7: 對 HuBERT 分群數 50 離散單元取得 500 種次詞單位後，
對應到塞擦音的聲學片段之音位條件機率排名

符記種數	音位分類純度	以音位分類標註之分群純度
離散單元	0.7006	0.1509
500	0.7116	0.0340
1000	0.7186	0.0226
8000	0.7080	0.0119
10000	0.7048	0.0113
20000	0.6929	0.0089

(a) 群數 = 50

符記種數	音位分類純度	以音位分類標註之分群純度
離散單元	0.7584	0.0882
500	0.7578	0.0326
1000	0.7576	0.0223
8000	0.7382	0.0097
10000	0.7346	0.0090
20000	0.7235	0.0074

(b) 群數 = 100

表 4.2: HuBERT 模型在不同詞表大小時的語音學類別分析數據

僅有分群數 50 時在較少符記種類時，音位分類的純度有微幅提升，多數時候符記種數的提高，伴隨的反而是音位分類純度些微的降低。由此可以推斷，次詞單位的引入雖能帶來更多樣的符記，對應音位間的關係卻在 K-平均演算法得到的離散單元已經大致抵定，次詞單位帶來的效果幾乎已經沒什麼改善的空間。其理由很可能是源自次詞單位演算法的特性，聲學片段的計算過程可以把原先代表不同種類音位的離散單元合在一起。因此，即便整體新的符記對應音位的純度有所提升，對音位分類的相關性卻低上不少。

4.5.2 由音位角度探討

考慮完聲學片段，接著我們一樣以音位的角度切入，觀察各自音位的符記分佈集中程度。圖 4.8 是對 HuBERT 模型所得的離散單元（比較分群數 50 和 100），以不同次詞單位種數取得聲學片段後，對應符記熵 $H(z|y)$ 最高與最低的排名。比對最左側直行顯示的離散單元排名，亦即第三章不引入次詞單位的結果，可以發現整體排名趨勢雖有些微變動，但對應符記最分散的音位仍以 AH、IH、T、D 為主，而最集中的亦仍然是 ZH、SH、F、EY，與上一章的觀察接近。由此可以推論，音位本身的較容易或較難以歸類的特性，在對語音表徵進行分群時就已經大致呈現；然而，即使聲學片段的演算法允許將代表不同類別音位的離散單元重新組合成新的符記，卻仍舊維持了音位本身分散程度的趨勢。因此，音位本身對應符記，不論是使用 K-平均演算法離散化獲得，或是以次詞單位重新歸類，這個分散程度的趨勢都是差不多的，音位本身的發音特徵確實是超出單一音框、影響範圍更廣的特性。

最後，我們可以將音位分類分別考慮，統計其各自的純度與相互資訊數據，與上一章節比對。對 HuBERT 分群數 50 離散單元文本以不同次詞單位種數處理

符記熵最高 (HuBERT · 分群數 = 50)							符記熵最低 (HuBERT · 分群數 = 50)						
次詞 單位數	單元	500	1k	8k	10k	20k	次詞 單位數	單元	500	1k	8k	10k	20k
# 1	spn	spn	spn	spn	spn	AH	# 1	ZH	ZH	ZH	ZH	ZH	ZH
# 2	AH	AH	T	AH	AH	spn	# 2	SH	SH	SH	SH	SH	SH
# 3	IH	T	AH	IH	IH	IH	# 3	E	E	E	EY	EY	OY
# 4	T	IH	IH	T	T	T	# 4	EY	EY	EY	W	OY	EY
# 5	D	D	D	sil	sil	sil	# 5	AA	AA	L	OY	W	W
# 6	EH	sil	sil	D	D	D	# 6	W	W	W	Y	Y	Y
# 7	TH	EH	K	EH	EH	EH	# 7	S	OW	Y	F	F	JH
# 8	HH	N	N	N	N	N	# 8	CH	L	CH	UW	UW	CH
# 9	UH	UH	EH	AE	AE	AE	# 9	IY	Y	OY	JH	JH	E
#10	G	G	G	K	DH	DH	#10	Y	CH	OW	CH	CH	UW
#11	sil	K	UH	G	K	HH	#11	OW	R	UW	L	L	OW
#12	N	TH	AE	DH	G	K	#12	Z	AO	M	OW	OW	L
#13	K	AE	TH	P	HH	G	#13	AO	M	R	IY	IY	AW
#14	AE	HH	NG	S	S	S	#14	L	UW	JH	M	M	IY
#15	P	NG	HH	TH	P	P	#15	ER	AY	AA	AO	AO	M

(a) 分群數 = 50

符記熵最高 (HuBERT · 分群數 = 100)							符記熵最低 (HuBERT · 分群數 = 100)						
次詞 單位數	單元	500	1k	8k	10k	20k	次詞 單位數	單元	500	1k	8k	10k	20k
# 1	spn	spn	spn	spn	spn	spn	# 1	SH	SH	SH	ZH	ZH	ZH
# 2	AH	AH	AH	AH	AH	AH	# 2	Y	ZH	ZH	OY	OY	OY
# 3	IH	T	T	T	T	T	# 3	ZH	Y	Y	Y	Y	Y
# 4	T	IH	sil	sil	sil	sil	# 4	E	NG	JH	SH	SH	SH
# 5	D	D	IH	IH	IH	IH	# 5	NG	E	NG	NG	NG	NG
# 6	sil	N	D	D	D	D	# 6	EY	EY	UW	UW	EY	EY
# 7	EH	sil	EH	EH	EH	EH	# 7	UW	AW	OY	F	UW	UW
# 8	HH	EH	N	S	S	S	# 8	W	UW	F	EY	F	F
# 9	UH	HH	AE	N	N	N	# 9	AW	JH	CH	JH	JH	JH
#10	AE	R	S	AE	AE	AE	#10	AY	AY	AW	CH	CH	CH
#11	K	AE	ER	K	K	Z	#11	M	OY	EY	AW	AW	AW
#12	N	TH	K	Z	Z	K	#12	AA	CH	OW	W	OW	W
#13	R	ER	HH	R	R	R	#13	OW	W	L	L	W	OW
#14	G	UH	R	HH	HH	DH	#14	CH	OW	AA	M	M	AY
#15	P	K	Z	DH	DH	HH	#15	L	AA	M	OW	L	M

(b) 分群數 = 100

圖 4.8: HuBERT 表徵、K-平均演算法分群數 50 和 100，

比較不同次詞單位種數時，符記熵最高與最低的音位排名

次詞單位數	XXX 音位純度	塞音 音位純度	擦音 音位純度	塞擦音 音位純度	鼻音 音位純度	單元音 音位純度	雙元音 音位純度	近音 音位純度
-	0.9924	0.4744	0.7033	0.6616	0.7580	0.5222	0.7813	0.8658
500	0.9931	0.5732	0.7390	0.7358	0.7745	0.5491	0.8093	0.8849
1000	0.9933	0.6002	0.7550	0.7402	0.7821	0.5658	0.8221	0.8922
8000	0.9944	0.6462	0.7949	0.8156	0.8152	0.6210	0.8607	0.9154
10000	0.9946	0.6505	0.7983	0.8209	0.8190	0.6271	0.8653	0.9181
20000	0.9951	0.6677	0.8067	0.8372	0.8313	0.6443	0.8807	0.9230

次詞單位數	XXX 分群純度	塞音 分群純度	擦音 分群純度	塞擦音 分群純度	鼻音 分群純度	單元音 分群純度	雙元音 分群純度	近音 分群純度
-	0.1733	0.2621	0.4688	0.4760	0.3633	0.3252	0.4996	0.4130
500	0.0493	0.0568	0.1063	0.0982	0.0771	0.0786	0.1308	0.1262
1000	0.0217	0.0345	0.0728	0.0876	0.0656	0.0539	0.0794	0.0977
8000	0.0149	0.0122	0.0343	0.0277	0.0321	0.0219	0.0334	0.0522
10000	0.0147	0.0117	0.0313	0.0248	0.0308	0.0198	0.0316	0.0469
20000	0.0131	0.0090	0.0252	0.0187	0.0233	0.0148	0.0219	0.0336

次詞單位數	XXX PNMI	塞音 PNMI	擦音 PNMI	塞擦音 PNMI	鼻音 PNMI	單元音 PNMI	雙元音 PNMI	近音 PNMI
-	0.8295	0.2082	0.5596	0.1065	0.3512	0.3961	0.5683	0.7110
500	0.8441	0.3531	0.6299	0.2460	0.4142	0.4389	0.6350	0.7618
1000	0.8480	0.3842	0.6477	0.2620	0.4363	0.4573	0.6682	0.7781
8000	0.8701	0.4596	0.7009	0.4226	0.5160	0.5207	0.7472	0.8185
10000	0.8746	0.4668	0.7057	0.4391	0.5253	0.5282	0.7559	0.8229
20000	0.8876	0.4947	0.7200	0.4852	0.5548	0.5505	0.7860	0.8333

圖 4.9: HuBERT 分群數 50 的離散單元，以不同符記種數取得聲學片段後，按照音位分類分開各自計算的純度與相互資訊

次詞單位數	XXX 音位純度	塞音 音位純度	擦音 音位純度	塞擦音 音位純度	鼻音 音位純度	單元音 音位純度	雙元音 音位純度	近音 音位純度
-	0.9943	0.5480	0.7535	0.6917	0.8447	0.6306	0.8273	0.8952
500	0.9948	0.6062	0.7719	0.7061	0.8490	0.6450	0.8443	0.8973
1000	0.9949	0.6415	0.7804	0.7474	0.8627	0.6554	0.8490	0.9028
8000	0.9955	0.7399	0.8193	0.8372	0.8915	0.6858	0.8797	0.9257
10000	0.9956	0.7455	0.8224	0.8446	0.8946	0.6895	0.8834	0.9275
20000	0.9960	0.7636	0.8312	0.8645	0.9030	0.7001	0.8941	0.9327

次詞單位數	XXX 分群純度	塞音 分群純度	擦音 分群純度	塞擦音 分群純度	鼻音 分群純度	單元音 分群純度	雙元音 分群純度	近音 分群純度
-	0.0855	0.1936	0.3591	0.3197	0.3325	0.2507	0.3978	0.3147
500	0.0380	0.0688	0.1288	0.1713	0.0946	0.0999	0.1791	0.1197
1000	0.0185	0.0438	0.0846	0.1380	0.0777	0.0630	0.1077	0.0830
8000	0.0093	0.0162	0.0293	0.0293	0.0294	0.0246	0.0370	0.0342
10000	0.0093	0.0143	0.0266	0.0274	0.0287	0.0221	0.0354	0.0320
20000	0.0092	0.0109	0.0206	0.0195	0.0251	0.0163	0.0270	0.0248

次詞單位數	XXX PNMI	塞音 PNMI	擦音 PNMI	塞擦音 PNMI	鼻音 PNMI	單元音 PNMI	雙元音 PNMI	近音 PNMI
-	0.8672	0.3351	0.6318	0.1998	0.5609	0.5168	0.6667	0.7734
500	0.8711	0.4236	0.6641	0.2414	0.6053	0.5406	0.7132	0.7856
1000	0.8716	0.4700	0.6773	0.3044	0.6292	0.5541	0.7260	0.8007
8000	0.8854	0.5947	0.7282	0.5106	0.7038	0.5958	0.7853	0.8391
10000	0.8880	0.6026	0.7333	0.5279	0.7113	0.6010	0.7919	0.8426
20000	0.8995	0.6308	0.7490	0.5797	0.7338	0.6170	0.8134	0.8534

圖 4.10: HuBERT 分群數 100 的離散單元，以不同符記種數取得聲學片段後，按照音位分類分開各自計算的純度與相互資訊

後，不同音位分類各自的純度與相互資訊數據以圖 4.9 呈現。由結果可以發現，隨著次詞單位種數的增加，除了原本音位純度較低的塞音在音位純度與相互資訊的提升較為明顯外，其他音位分類的音位純度與相互資訊就已經較高，因而雖然增加得不是很明顯，但整體大致仍然有所改善。比較圖 4.10，可以確認此一變化在分群數改為 100 時依然可見。

4.5.3 分析結論

藉由改以次詞單位重新組合離散單元得到聲學片段，透過符記種類數量的提升，聲學片段可以區別出語音訊號中更細節的語音差異，進而得以提升音位純度與相互資訊等數據，提高符記與音位之間的相關性。同時，次詞單位的特性雖然允許對應到不同音位的離散單元重新組合，然而如此產生的新符記，每個音位對應符記的集中或分散程度卻差異不大。因此，透過將次詞單位應用在離散單元的嘗試，結合多個離散單元重新編碼語音訊號，聲學片段在捕捉語音訊號規律上，儘管效果不如直接對語音表徵進行分群的離散單元好，卻能作為需要探索更細微語音資訊差異時，除了 K-平均演算法之外對語音訊號離散化的另一個選擇。

4.6 本章總結

本章節首先介紹了文字處理中常用的次詞單位，並嘗試對離散單元序列重新組合成聲學片段。接著，仿照第三章的分析方式，將離散單元與聲學片段互相比較，對比兩種不同符記與音位間對應關係的變化。結果顯示，無論是使用離散單元或次詞單位，儘管兩種方式在語音資訊捕捉效果上有所不同，但隨著符記種類的增加，都能獲得更加細節的語音資訊，並提升與音位之間的相關性。期望這一發現可以在未來建立語音語言模型時，除了 K-平均演算法的離散單元外，考慮與

次詞單位的演算法結合使用，以更全面和細緻的從語音訊號中提取與音位相關的語音學資訊。

第五章 結論與展望

5.1 研究貢獻與討論

本論文的主旨，在於分析語音基石模型的離散表徵，與語音標註之間的純度和相互資訊等數據的相關性，並且透過分詞方法的引入，嘗試將多個離散單元進行結合後，觀察學習到的新符記是否和音位等標註更加一致。

首先，論文第三章介紹了與無文字架構以及語音表徵相關的分析研究，隨後簡介語音學知識中，對於不同音位之間如何按照發音特性分門別類。有了音位與語音學分類兩種語音標註後，借鑑 HuBERT 提及的純度和相互資訊的分析方式，對離散表徵與語音標註之間，兩者的相關性進行分析與觀察，比對無文字架構中不同語音離散表徵的統計特性。結果可發現，HuBERT 作為目前無文字架構最常用的語音離散表徵模型的理由，很可能來自於它們的音位純度與相互資訊都相對較高，因而更能捕捉到語音中與內容相關的重要資訊，且同樣的趨勢在語音學分類的標註也可以被觀察到。

其後在論文第四章，考慮到音位與離散單元往往是一對多的關係，藉著嘗試引入自然語言處理常用的分詞方法，重新對離散單元的序列進行分組，並且比較不同詞表大小對這些分析數據的影響。考量語音不如英語的文字系統具備明確的空格提示，本研究採取單一詞作為分詞方法進行實驗，並比較不同模型與不同詞表大小對第三章的分析數據是否造成影響。

結果顯示，藉助加入分詞方法並提供足夠大的詞表，確實能夠讓不同音位的純度以及相互資訊有所提升，讓多個離散單元之間有機會相互結合、重新分組，更能捕捉語音訊號中的內容資訊。且在四種語音表徵之間，HuBERT 依然是所有模型中音位純度和相互資訊最高者，還達成了一定的序列長度壓縮比率，相較之

下 CPC 模型雖然壓縮比率更低，卻犧牲掉過多語音資訊導致相關數據反而較差。這也解釋了為什麼目前在離散單元相關的研究中，無論是使用單一離散單元，或是使用分詞方法進行長度壓縮等，HuBERT 都仍是最有利於後續語音任務的訓練與應用的模型。

5.2 未來展望

希望這些對離散單元與分詞方法應用的嘗試，能幫助我們在訓練任務之前，決定哪種語音基石模型更適合作為離散編碼語音訊號的基礎。接下來，我們期望能針對常見的語音任務，特別是語音辨識和語音翻譯等內容處理相關的任務，比對離散單元促成的實際成效和分析數據之間的關係，並對這些任務中的錯誤案例進行統計和個案探討。

另外，對於如何結合語音離散單元，除了將其視為文字進行分詞演算法外，我們還可以使用其他方式對離散單元序列進行分組，以達成壓縮序列長度並使其與音位等語音內容更加一致的目標。例如，將此目標形塑為語音分段（Speech Segmentation）任務等，也是未來可以嘗試的離散單元分組方式。

最後，利用語音學分組的切入點，或許可以在未來分析離散單元或連續語音表徵時，不再僅限於參考音位或文字，還可以從語音學知識提供的相似性資訊出發，為錯誤發音修正等任務提供衡量的依據。

參 考 文 獻

- [1] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 27th ed. Dallas, Texas: SIL International, 2024. [Online]. Available: <https://www.ethnologue.com>
- [2] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, “Unsupervised speech representation learning using wavenet autoencoders,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019, publisher: IEEE.
- [3] L.-W. Chen, S. Watanabe, and A. Rudnicky, “A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12 644–12 652, Jun. 2023, number: 11. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/26488>
- [4] X. Zhao, Q. Zhu, J. Zhang, Y. Zhou, and P. Liu, “Speech Enhancement with Multi-granularity Vector Quantization,” in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Oct. 2023, pp. 1937–1942, iSSN: 2640-0103. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10317485>
- [5] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, “On Generative Spoken Language Modeling from Raw Audio,” *Transactions of the Association for*

Computational Linguistics, vol. 9, pp. 1336–1354, 2021, place: Cambridge, MA

Publisher: MIT Press. [Online]. Available: <https://aclanthology.org/2021.tacl-1.79>

- [6] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” Jan. 2019, arXiv:1807.03748 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [7] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, “Hubert: How Much Can a Bad Teacher Benefit ASR Pre-Training?” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6533–6537. [Online]. Available: <https://ieeexplore.ieee.org/document/9414460/>
- [8] 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*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>
- [9] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong, H. Inaguma, S. Popuri, C. Wang, J. Pino, W.-N. Hsu, and A. Lee, “Speech-to-Speech Translation for a Real-world Unwritten Language,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4969–4983. [Online]. Available: <https://aclanthology.org/2023.findings-acl.307>

- [10] F. Wu, K. Kim, S. Watanabe, K. J. Han, R. McDonald, K. Q. Weinberger, and Y. Artzi, “Wav2Seq: Pre-Training Speech-to-Text Encoder-Decoder Models Using Pseudo Languages,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10096988/>
- [11] S. Ren, S. Liu, Y. Wu, L. Zhou, and F. Wei, “Speech Pre-training with Acoustic Piece,” in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2648–2652. [Online]. Available: https://www.isca-archive.org/interspeech_2022/ren22_interspeech.html
- [12] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe, Y. Fujita, T. Maekaku, P. Guo, Y.-F. Cheng, P. Denisov, K. Saijo, and H.-H. Wang, “Exploring Speech Recognition, Translation, and Understanding with Discrete Speech Units: A Comparative Study,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11481–11485, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/10447929>
- [13] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, publisher: Springer. [Online]. Available: <https://doi.org/10.1007/BF02478259>
- [14] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Review*, vol. 65, no. 6, pp. 386–

- 408, 1958, publisher: American Psychological Association. [Online]. Available: <https://doi.apa.org/doi/10.1037/h0042519>
- [15] K.-I. Funahashi, “On the approximate realization of continuous mappings by neural networks,” *Neural Networks*, vol. 2, no. 3, pp. 183–192, Jan. 1989. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0893608089900038>
- [16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/323533a0>
- [17] D. E. Rumelhart and J. L. McClelland, “Learning Internal Representations by Error Propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, 1987, pp. 318–362. [Online]. Available: <https://ieeexplore.ieee.org/document/6302929>
- [18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Jan. 1998, conference Name: Proceedings of the IEEE. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/726791>
- [19] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/>
- [20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

- [21] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: <https://aclanthology.org/W14-4012>
- [22] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” 2014.
- [23] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [24] 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*, vol. 30, 2017.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.
- [26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Sep. 2013, arXiv:1301.3781 [cs]. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [27] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 59–67. New Orleans, Louisiana: Association for Computational Linguistics, 2018. [Online]. Available: <https://www.aclweb.org/anthology/N18-1008.pdf>

- Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202>
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [29] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, “Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders,” Oct. 2019. [Online]. Available: <https://arxiv.org/abs/1910.12638v2>
- [30] L. T, LiShang-Wen, and LeeHung-yi, “TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, publisher: IEEE. [Online]. Available: <https://dl.acm.org/doi/10.1109/TASLP.2021.3095662>
- [31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
- [32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss,

- G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- [33] Y.-A. Chung and J. Glass, “Generative Pre-Training for Speech with Autoregressive Predictive Coding,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3497–3501, iSSN: 2379-190X.
- [34] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, “Multi-Task Self-Supervised Learning for Robust Speech Recognition,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6989–6993, iSSN: 2379-190X.
- [35] 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*, vol. 29, pp. 3451–3460, 2021, publisher: IEEE.
- [36] T. Maekaku, X. Chang, Y. Fujita, L.-W. Chen, S. Watanabe, and A. Rudnicky, “Speech representation learning combining conformer cpc with deep cluster for the zerospeech challenge 2021,” 2022.

- [37] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised Pre-Training for Speech Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.
- [38] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, “Unsupervised pretraining transfers well across languages,” 2020.
- [39] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [40] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [41] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” *arXiv preprint arXiv:1910.05453*, 2019.
- [42] “Textless NLP: Generating expressive speech from raw audio,” Sep. 2021. [Online]. Available: <https://ai.meta.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/>
- [43] G.-T. Lin, Y.-S. Chuang, H.-L. Chung, S.-w. Yang, H.-J. Chen, S. Dong, S.-W. Li, A. Mohamed, H.-y. Lee, and L.-s. Lee, “Dual: Discrete spoken unit adaptive learning for textless spoken question answering,” *arXiv preprint arXiv:2203.04911*, 2022.
- [44] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Speechtokenizer: Unified speech tokenizer for speech large language models,” 2024.
- [45] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, “Generative Spoken Language

- Modeling from Raw Audio,” Sep. 2021, arXiv:2102.01192 [cs]. [Online]. Available: <http://arxiv.org/abs/2102.01192>
- [46] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *International Conference on Learning Representations*, 2021.
- [47] A. Sicherman and Y. Adi, “Analysing discrete self supervised speech representation for spoken language modeling,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.
- [48] B. M. Abdullah, M. M. Shaik, B. Möbius, and D. Klakow, “An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech,” in *Proc. INTERSPEECH 2023*, 2023, pp. 2883–2887.
- [49] X. Chang, B. Yan, Y. Fujita, T. Maekaku, and S. Watanabe, “Exploration of Efficient End-to-End ASR using Discretized Input from Self-Supervised Learning,” in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1399–1403. [Online]. Available: https://www.isca-archive.org/interspeech_2023/chang23b_interspeech.html
- [50] A. H. Liu, H.-J. Chang, M. Auli, W.-N. Hsu, and J. Glass, “Dinosr: Self-distillation and online clustering for self-supervised speech representation learning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [51] Z. Huang, C. Meng, and T. Ko, “Repcodec: A speech representation codec for speech tokenization,” *arXiv preprint arXiv:2309.00169*, 2023.

- [52] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux, and G. Wisniewski, “Probing phoneme, language and speaker information in unsupervised speech representations,” in *Proc. Interspeech 2022*, 2022, pp. 1402–1406.
- [53] D. Wells, H. Tang, and K. Richmond, “Phonetic Analysis of Self-supervised Representations of English Speech,” 2022, pp. 3583–3587. [Online]. Available: https://www.isca-archive.org/interspeech_2022/wells22_interspeech.html
- [54] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/7178964>
- [55] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.
- [56] “The CMU Pronouncing Dictionary.” [Online]. Available: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict?stress=-s&in=CITE>
- [57] A. Klautau, “Arpabet and the timit alphabet,” *an archived file. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf* (Accessed Mar. 12, 2020), 2001.
- [58] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds. Brussels, Belgium:

- Association for Computational Linguistics, Jan. 2018, pp. 66–71. [Online]. Available: <https://aclanthology.org/D18-2012>
- [59] A. Elkahky, W.-N. Hsu, P. Tomasello, T.-A. Nguyen, R. Algayres, Y. Adi, J. Copet, E. Dupoux, and A. Mohamed, “Do coarser units benefit cluster prediction-based speech pre-training?” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.
- [60] F. Shen, Y. Guo, C. Du, X. Chen, and K. Yu, “Acoustic bpe for speech generation with discrete tokens,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 746–11 750.
- [61] H.-J. Chang and J. Glass, “R-spin: Efficient speaker and noise-invariant representation learning with acoustic pieces,” *arXiv preprint arXiv:2311.09117*, 2023.
- [62] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1456>
- [63] P. Gage, “A new algorithm for data compression,” *C Users J.*, vol. 12, no. 2, p. 23–38, feb 1994.
- [64] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith,

Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: <https://aclanthology.org/P16-1162>

- [65] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.
- [66] T. Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” 2018.