目錄

一、導論 4

1.1 研究動機 4

1.2 研究方向 6

1.3 主要貢獻 7

1.4 章節安排 8

二、背景知識 9

2.1 類神經網路 9

2.1.1 簡介 9

2.1.2 卷積式類神經網路 11

2.1.3 遞迴式類神經網路 11

2.1.4 序列至序列（Sequence-to-sequence，Seq2seq）模型 12

2.1.5 專注機制（Attention mechanism） 12

2.1.6 轉換器（Transformer）類神經網路 13

2.2 表徵（Representation）學習與自監督式學習（SSL） 14

2.2.1 聲學特徵 14

2.2.2 表徵學習 14

2.2.3 自監督學習 15

2.2.4 向量量化（Vector quantization） 16

2.2.5 離散單元與無文字（Textless）架構 16

2.3 本章節總結 17

三、單一語音離散表徵與語音標記的對應模式 18

3.1 動機簡介 18

3.2 相關研究 18
3.2.1 語音表徵的語音學分析 18
3.2.2 無文字（textless）語言模型 19
3.3 衡量方式（metric） 19
3.3.1 音位（phoneme）長條圖（bar chart） 19
3.3.2 純度（purity） 19
3.3.3 熵（entropy）和相互資訊（mutual information，MI） 20
3.3.4 對齊（alignment） 20
3.4 語音學分類（phone type） 21
3.4.1 簡介 21
3.4.2 解釋意義 21
3.5 分析結果 22
3.5.1 基於各自音位的分析 22
3.5.2 基於語音學分類的分析 22
3.6 本章總結 22
四、多個語音離散表徵組合與語音標記間的關係 23
4.1 動機 23
4.2 相關研究 23
4.2.1 對語音離散表徵的分詞（tokenization）研究 23
4.3 分詞方法 23
4.4 衡量方式 23
4.4.1 字符（token）與音位之間的關係 23
4.4.2 壓縮比率 23

4.5 分析結果 23
4.5.1 基於各自音位的分析 23
4.5.2 基於語音學分類的分析 23
4.6 應用在語音任務的實驗 24
4.6.1 語音辨識 24
4.7 本章總結 24
五、結論與展望 25
5.1 研究貢獻與討論 25
5.2 未來展望 25
參考文獻 26

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，各種書寫系統其背後承載的語言必定有語音的形式作為對應。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自然不言而喻。

然而，相對於穩定、易於處理和保存的文字文本，語音訊號的變化萬千，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、研究相比之下複雜度與難度劇增。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，更好的讓電腦、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注過程，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注、相較容易取得的語音或文字的語料，找出其中的語音、語法及語義等等結構，形成帶有對人類語言有前所未見表現的基石模型（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開始廣泛被採用。然而，相比於文字符號的穩定與單純，語音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，[5] 基於模仿人類學習語言的過程，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，企圖以單純大量的語音資料中訓練出一個不依賴文字的語言模型。此種學習架構的優勢在於在能保有利用大量未標注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）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（Language Model）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（Discrete Unit）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，然而其作為「完全文字的替代」仍然有相當的距離。借鑑過往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 中從消息理論（Information Theory）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（Phone 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（Linguistics）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）」）或單字（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的「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。

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 BPE 等等常見的次詞單位（Subword）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規律（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，從「將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。

1.3 主要貢獻

本論文達成的主要成果是以更細緻的方式，對現在愈來愈廣為使用的離散單元以音位和語音類別等語音知識的視角給出一個基礎相關性的分析方法，並將單一離散單元本身與將多個單元透過分詞演算法（Tokenization）重新編碼前後進行比較，初步試探離散單元與音位之間的關係，並期望作為「離散單元可否一定程度上的『被視為文字』或『有機會從中發掘出文字單位』」的判斷基礎，為往後研究往語音語言模型（Spoken Language Model）中「對語音編碼」這個重要的程

序，提供一個在實際上開始耗費資源的模型訓練之前，可比較的判斷標準。

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（Deep Learning）、表徵學習與自監督學習相關的基礎背景知識。
- 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：講述為何單一離散單元本身或許不全然足夠發掘出類似音位進而對應到文字的單位，以及近年人們嘗試以離散單元為基礎，透過分詞演算法（Tokenization Algorithm）發展之聲學片段（Acoustic Piece）的進展，接著我們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，以論證對離散單元進行分詞是否可以找出更接近音位的單位，驗證「離散單元可被文字化」或「離散單元學到的是否為更精細的語音訊號規律或結構（Structure）」等論述。
- 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　類神經網路

(Like this [13, 8] and fade out.)

2.1.1　簡介

((深層類神經網路是 McColloh 在 1950 年提出的計算模型，其概念取法自連結主意學派，旨在用計算模型模擬生物神經連結的現象。類神經網路最基本的單元為一個神經元（Neuron），來自 1960 年 Rosenblatt 提出的感知器（Perceptron）模型，旨在根據輸入訊號給出的線性分類器，其數學是可以表達成：))

類神經網路是 McCulloch 和 Pitts [14] 在 1943 年提出的計算模型，旨在模仿生物神經的連結，並能透過演算法的最佳化，擬和我們想要的函數以實現特定的應用功能。由於其彈性，目前已成為人工智慧發展的主流。

類神經網路最基本的單元是「神經元」，其本質為一個線性分類器，會接受一串數字作為輸入並計算出一個數字作為輸出，可用下列數學運算式描述：

$$y = \sigma(w^\top x + b)$$

其中 x 是 N 個輸入的數字，可描述為一個 N 維的向量；w 為該神經元對每個輸入值給予的權重，再對加權平均後的結果加上偏差值 b 後，經過激發函數（Activation Function）σ 的非線性轉換後最後得到輸出。常用的激發函數包含 ReLU、sigmoid 和 tanh 等等。

藉由結合好幾個神經元的運算，基於 Universal Approx. Thm. ，在數學理論上我

們可以近乎模擬一切函數，這樣的機器學習模型即為 Perceptron [15]。然而單純增多神經元數目仍無法解決如 XOR 等分類問題，於是爾後 MLP 的概念被 _____ 提出，即結合多層感知器，在輸入與輸出之間加入隱藏層（Hidden layer）以對運算數據進行表徵空間上的轉換，可以更好的拓展類神經網路的適用範圍，解決更加複雜的現實問題。此種透過加深類神經網路隱藏層形成的計算模型便被稱為深層類神經網路（Deep neural network）。

然而單純擁有一個可以表達複雜函數的模型是不夠解決工程應用問題的，為了增加函數擬和（Fit）的效率，_____ 在 _____ 年提出了 backpropagation 的演算法，旨在藉由計算輸出層與目標函數之間的誤差，透過最佳化演算法計算出梯度後，經由隱藏層反向往輸入層對於整個類神經網路進行修正，便能配合 GPU 大量平行運算的能力，很好的從資料中找尋出我們想要的函數。這樣透過深層類神經網路，從資料輸入與輸出之間尋找函數的機器學習演算法，就稱之為深層學習。由於深層學習的 scalibility 與泛用性，不論在圖像、語音、文字等多個模態，深層類神經網路都已經獲得了廣泛應用。

然而根據資料特性的不同，並不是所有的資料都單純適用這樣輸入與輸出向量直接對應的模式，因此類神經網路又發展出不同的架構以適應資料本身的特性。前述的類神經網路由於運算過程單純是從輸入層經由多層感知器直接進行矩陣運算完成函數的模擬，因此被稱之為「前饋式類神經網路」。與此直接計算不同，在輸入與輸出之間調整連接關係，可以得到 CNN、RNN 與 Transformer 等架構。由於這些架構在語音與文本處理上已是主流選擇，接下來分別介紹：

2.1.2 卷積式類神經網路

卷積式類神經網路（Convolutional neural network）為 1998 年由楊氏（LeCun）提出 [13]，旨在利用訊號處理上卷積（Convolution）的運算模擬人類視覺皮質感知 [16] 的特性，利用其移動不變性（Shift-invariance）來捕捉二維影像中的局部（Local）特徵，以便於後續的類神經網路可以對輸入的資料進行更整體而全面的判斷。

有別於圖像中經常是以 pixel 的 RGB 亮度進行卷積運算，在語音中 CNN 處理的對象除了直接是空氣壓力波形的物理訊號以外，為了更方便機器模型判斷語音訊號的內容，透過聲學知識得到的聲學特徵或深層學習得出的語音表徵也經常是語音處理中卷積層運算的對象。然而不論是何種輸入，有別於影像的二維資料，語音訊號的資訊是被呈現在時間軸的維度上，因此通常使用一維的卷積式類神經網路，以模仿人耳聽覺對時變訊號的窗框（Window）處理過程，讓模型可以觀察到輸入語音在不同解析度（Resolution）上的資訊，例如本研究特別著重的音位（Phoneme）等。

// text embedding 先不寫

// speech acoustic features 拖到後面去寫

2.1.3 遞迴式類神經網路

有別於運算過程由輸入往輸出單向的 FFN 和 CNN，為了處理有記憶和狀態的資料，特別是會隨時間變化的序列資訊，在語音和文字的機器學習中，會將輸出訊號重新接回輸入層的 RNN 是一個相當符合語言特性的選擇。RNN 以每個時間點（Timestep）為考慮對象，在每一步會對輸入層的向量進行運算後，不但將此結果算出一個輸出向量，還會得到另外一些數據保留作內部狀態，表示此前經歷過所

有序列資料的記憶。常用的 RNN 的類型有 LSTM 和 GRU，這兩種 RNN 的示意圖與運算式如下：

// 放運算式和示意圖

此類類神經網路通常會以下列介紹的序列至序列的形式被用在如語音辨識、語音合成或機器翻譯等和語言密切相關的任務中。

2.1.4 序列至序列（Sequence-to-sequence，Seq2seq）模型

由於許多以語言為主的資料經常以兩個序列互相配對的形式呈現，因此專門用以處理此類資料的模型被特別稱為序列至序列模型。此類模型一般的架構是由一個編碼器（Encoder）和一個解碼器（Decoder）構成，旨在模擬輸入與輸出序列之間的變化與相依關係（Dependency）。

此類模型一般有兩種模式：

其一是每個時間點都取得一個輸出的向量，用在輸入與輸出等長的任務之中，此模式又被稱為 token classification。

但更常見的狀況是，輸入與輸出兩者序列長度並不總是相同，此時典型的作法是，讓編碼器將輸入序列在每個時間點一一與模型進行運算，藉由內部表徵（Latent representation）的調整對整個輸入序列進行編碼，完成後將最後一個時間點的表徵作為整個序列的代表，此表徵向量會被稱為「語境向量（Context vector）」，接著被傳遞給解碼器依時序生成輸出訊號的序列。

2.1.5 專注機制（Attention mechanism）

然而由於 RNN 本身需要編碼和解碼的資訊量是整個序列，對時間點距離比較遠的輸入容易被遺忘，也就是難以處理長期相依性（Long-term dependency）的問

題。為解決這種困境，Luong 等人提出了「專注機制（Attention mechanism）」，讓解碼器除了依據語境向量的資訊以外，還可以對輸入序列的不同時間點分配權重，在生成輸出序列時重新從輸入序列中得到所需的訊息。

專注機制一般涉及三個向量之間的運算：query、key 和 value，其運算式如下：

(KQV 運算)

具有專注機制的序列至序列模型又被稱為 AED，透過專注機制的引入，大大改善了如語音辨識、機器翻譯等任務的效能。

2.1.6 轉換器（Transformer）類神經網路

儘管 RNN 本身善於處理時序資料，然而它難以平行化的架構限制卻大大束縛了其在訓練和推理（Inference）時的效率。由 attention 獲取靈感，2017 年瓦氏（Vaswani）等人在 ___cite___ 提出了一種完全由專注機制構成，不需依賴遞迴運算的序列至序列模型，用以解決最經典的機器翻譯任務。

轉換器架構

轉換器一樣沿用了 Attention 的 KQV 三組向量的邏輯，以 positional encoding 對序列中每個位置的時間點進行編碼，取代原先在 RNN 模型需要一一運算的過程，在實行平行計算的同時也能考慮到資料在不同時間點出現的效應。其整體架構如下：

(tfm 的圖)

(講多頭專注、KQV、FFN 那些)

由於轉換器不需對每個時間點一一運算，使其得以實現高度平行化的優勢，類神經網路得以透過專注機制的幫助同時進行序列資料的大量訓練，這種 scalibility

因而在自然語言和語音處理都獲得了巨大的進展，近乎取代了原先 RNN 的應用場景，近年甚至被電腦視覺的研究者推廣應用在圖像類的資料上 (___vit___) ，足以展現此種模型架構的彈性與泛用性，是目前最前沿人工智慧的主流架構。

除了模型架構，機器學習中不可或缺的另一大 component 即是對資料的編碼過程。如何更有效率的讓機器可以理解、處理和輸出，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，語音和文字處理如何對資料去蕪存菁，表徵學習更是重中之重。

2.2 表徵 (Representation) 學習與自監督式學習 (SSL)

2.2.1 聲學特徵

為了讓機器可以理解輸入的資料，表徵學習是機器學習中不可或缺的一部分。

在語音處理中，在過往機器運算能力還沒有那麼強大的時候，人們基於聲學原理，使用 MFCC 為處理的對象。在文字中則通常使用 TF-IDF

(這邊寫 mfcc 的介紹)

2.2.2 表徵學習

後來 Mikolov 提出了 word2vec 的做法，使用 distributed representation 對文字的單詞（Word）進行編碼，透過大量的文本單詞之間的 co-occurence 去找出每個單詞最適合的語義表徵。其後 ELMo 提出了 contextualized embedding 的想法，更細緻的在單詞本身之外也嘗試對句子脈絡的語義進行詞嵌入（Word embedding）編碼。

2.2.3 自監督學習

爾後在 Transformer 模型被提出後，BERT (cite BERT) 被提出，從大量的文本與自專注機制之中，工程師們便可不借助人為的標記，透過預先設定的 pretext tasks 的引導，使得模型可以自己從大量文本中更細緻、更 contextualized 的自行找出語義關係，並在許多 NLP 的任務上獲得了 SoTA 的成績。自此「self-supervised learning (SSL)」的概念大行其道，這種以 pretext tasks 代替標註資料本身，從大量的未標註資料中利用資料本身結構進行學習的模式成為主流。由於其學習對象是發掘自大量的資料本身，可以更好的利用 NN 的泛化（Generalization）能力去找出什麼樣的資訊對於人們日常應用任務中是重要的並予以保留。

有鑑於文字處理方面的成功有許多的語音處理學者便嘗試將類似的模式套用在語音訊號之上自此提出了很多的語音基石模型而這些藉由語音基石模型得出來的語音表徵，也在很多任務上被證明可以超越傳統上使用的 fbank、MFCC 等工程特徵，因為大量的語音資料庫本身可以幫助模型去萃取出更適用於各種語音任務上的向量表徵。

依照這些語音自監督模型的學習模式，大致可以分為重建式、預測式與對比式模型，以下分別按照這三類模式介紹這些語音基石模型：

(這邊接下來直接看以前碩論怎麼分。宏毅老師的 review paper 再說）}

重建式學習

此類模型在文字處理以 BERT 為代表，BERT 本身屬於遮罩語言模型（Masked language model，MLM）。在語音中以 Mockingjay、TERA 為主要採取此模式的基石模型。

預測式學習

此類模型在文字處理以 GPT 為典型，不同於 BERT 是任意對資料進行擾動作為預訓練任務，這類模型的目標即是單純的自迴歸（Auto-regressive），可以用以下式子來表達其訓練 objective：

（寫那個$y = p(x|x < t)$ 什麼的式子）

在語音中以 APC 為代表。

（是不是有一個寫法把前兩個都當成預測，只是一個是重建一個是自迴歸？）

對比式學習

這類模型以 CPC 為主。在電腦視覺有 BYOL 等等。

2.2.4 向量量化（Vector quantization）

基於向量本身容易受噪聲擾動而導致訓練不穩定，因此為了穩定訓練，向量量化（Vector quantization）的技巧變常常為機器學習，尤其是語音這邊所使用。例如 vq-wav2vec 與其後的 wav2vec 2.0 就使用了這樣的技巧，HuBERT 本身對語音表徵進行 KMeans clustering 也是一個向量量化的手段。（是不是應該寫一下 KMeans 是什麼？）

（補說一下什麼是 discrete unit）

2.2.5 離散單元與無文字（Textless）架構

由於 HuBERT 本身的成功（好像要寫理由？），其後 Meta 提出了完全基於 HuBERT unit 的抽取方式，完全只依賴語音而不依靠文字標註的「無文字

(Textless)」架構被提出，其代表作為 GSLM。(好像也要提一下 speech-to-speech 翻譯嗎?)

無文字目前在 QA (cite 實驗室的 DUAL) 跟語音到語音翻譯 (Cite 並描述臺語翻英語翻譯) 獲得了很好的成功。自此這類「離散單元 (Discrete unit)」被視為一項類似文字卻不需要真的依賴人類文字標記的語音表徵，其最大優勢為儲存的 bit rate 低與可以套用 NLP 文字的「語言模型」之訓練模式。

然而，雖然在系統與應用 task 上獲得了很大的成功，但 unit 本身是否已經真的很好的可以替代文字，或能夠多少的幫助 spoken language model 的訓練與建立，仍然是目前本領域探討的焦點議題。有鑑於此，本論文基於語言知識，從最接近文字但又跟語音訊號最密切相關的 phoneme 開始探討，期望對 unit 本身究竟能夠帶給我們什麼特徵、如何幫助後續應用進行進一步的研究。

2.3 本章節總結

本章節先是對作為 building block 的類神經網路進行了基本原理的介紹，其後對本論文研究的核心 — 「representation」與「discrete unit」的發展演進與歷史進行了簡單的梳理。此後兩章節就會緊扣著這些基石模型得到的離散特徵，將其與尤其是 phoneme 這類語音學標記之間的統計關係進行更進一步的分析。

(Ch 3 好像要講一下為什麼不做連續特徵 & 為什麼要以 phoneme 為客體了 T_T)

第 三 章　單一語音離散表徵與語音標記的對應模式

（Ch 3 好像要講一下為什麼不做連續特徵 & 為什麼要以 phoneme 為客體了 T_T）

3.1　動機簡介

由於 HuBERT 之後，unit 的使用很廣泛，因此為了研究 unit 本身為什麼會被如此適當的可以讓模型視為文字對語音資料進行訓練，我們先從離散表徵本身的特徵分析起。

3.2　相關研究

近期已經有多項相關的研究，嘗試在 SSL 這麼厲害的表現之後找原因，因此有針對 unit 背後 repr 的特性進行分析的 work，例如 CITEUSPLEASE。

3.2.1　語音表徵的語音學分析

在 HuBERT 出來之後，有一些研究像是 cite 等等，試圖探討對於語音表徵這樣語音模型的基礎進行各種從統計和語音學領域知識角度的分析，以期望能夠解釋為什麼模型可以擁有如此的表現。

此後，cite 等等作品則是從原先連續的表徵出發，開始往離散的量化向量，甚至是離散單元進行分析比對。雖然分析的切入角度可以相當多樣，例如 ABX、tsne 降維分群等等，但本次研究主要著重比對兩者之間在同一段語音序列上給予標籤的相關性，也就是以「偽標籤（pseudo-label）」的角度進行衡量。

3.2.2 無文字（textless）語言模型

這系列 textless 以 GSLM 為最主要代表作，旨在探討 unit 作為一種替代文字的方案。

本論文以 GSLM textless 採用的模型 units 為主要分析對象，企圖銜接兩者的脈絡，來佐證這些 unit 作為一種「類似或可替代文字的語音紀錄方式」在能夠發揮 LM 的特長背後，是否是基於符合語音學特徵帶來的，抑或有什麼其他特徵。

3.3 衡量方式（metric）

為了測量這些 unit 跟 phn 這類語言學 labels 之間的 correlation，我們需要先介紹本論文會探討的指標

3.3.1 音位（phoneme）長條圖（bar chart）

要先對 unit 做頻率上的統計

3.3.2 純度（purity）

phoneme purity: 每個 cluster unit 內的 phoneme purity，代表此 unit 是否有 phoneme 代表性

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

cluster purity：每個 phoneme 對應到的 cluster 統計，若 cluster purity 低代表 less liguistical meaning? 單一 phoneme 本身對應的 unit 的一致性。如果

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

3.3.3 熵（entropy）和相互資訊（mutual information，MI)

除了「最大」的對應關係，根據 Hubert 原先的 paper CITEME (hubert) 中的分析方式，我們也可以從 info theory 的角度，去探討「觀察到一個 unit 對於 label 不確定性的降低」來考慮 unit 本身提供了多少背後 phn 的資訊

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

$$\frac{I(y;z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i,j) \log \frac{p_{yz}(i,j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \tag{3.1}$$

$$= \frac{H(y) - H(y|z)}{H(y)} \tag{3.2}$$

$$= 1 - \frac{H(y|z)}{H(y)} \tag{3.3}$$

3.3.4 對齊（alignment）

cluster 是否保留 segment 資訊，不將不同 phoneme 合併 segmentation → 怕說只是每個 frame 本身 unit 或 piece 可以跟 phoeneme 特徵相關，但放在連續的語音中被切得很碎/前後文相關的東西不知道有沒有抓到

要計算 phoneme alignment ，我們參考 CITMEALIGNM 本身需要考慮 P 值、R 值等，以下說明如下：

(算 alignment 的一些介紹)

3.4 語音學分類（phone type）

3.4.1 簡介

除了單一 phn 本身的特性以外，由於 phn 本身彼此不是完全獨立的，而是彼此之間就存在相似的特徵，可以分成幾個組別。因此，依照 CITEME (tanghao 等三篇) 的分組方式，對英語的 phn 進行分類並合併比對數據，看看這些 unit 本身是否有 capture 到相似的發聲特徵，而不單純只是把 phn 分成約五十類完全獨立的標籤。(基於語音表徵本身就是 acoustic sisgnals 來的，應該 by nature 要可以對語音特徵分組吧?)

以下為各分組進行簡單介紹：

consonants

子音可以分成五類

vowels

母音在這邊為了簡單起見，會被分在一起？

3.4.2 解釋意義

- 純度（purity）：換成 type 之後有何變化（關聯性更強?)
- 熵（entropy）(放直方圖解釋）–> phone type 更明顯？
- 對齊（alignment）：是否減少 segment 資訊的保留（連續子音母音被合併？

3.5 分析結果

3.5.1 基於各自音位的分析

(放數據)

3.5.2 基於語音學分類的分析

(放數據)

3.6 本章總結

第 四 章　多個語音離散表徵組合與語音標記間的關係

4.1　動機

4.2　相關研究

4.2.1　對語音離散表徵的分詞（tokenization）研究

4.3　分詞方法

4.4　衡量方式

4.4.1　字符（token）與音位之間的關係

4.4.2　壓縮比率

4.5　分析結果

4.5.1　基於各自音位的分析

4.5.2　基於語音學分類的分析

4.6 應用在語音任務的實驗

4.6.1 語音辨識

實驗設定與資料集

實驗結果與其和分析數據間的關係

4.7 本章總結

第 五 章　結論與展望

5.1　研究貢獻與討論

5.2　未來展望

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.

[3] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 644–12 652.

[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)*. IEEE, 2023, pp. 1937–1942.

[5] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[6] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[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)*. IEEE, 2021, pp. 6533–6537.

[8] 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.

[9] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong *et al.*, "Speech-to-speech translation for a real-world unwritten language," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 4969–4983.

[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)*. IEEE, 2023, pp. 1–5.

[11] S. Ren, S. Liu, Y. Wu, L. Zhou, and F. Wei, "Speech pre-training with acoustic piece," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2648–2652. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-981

[12] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe *et al.*, "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)*. IEEE, 2024, pp. 11 481–11 485.

[13] 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, 1998.

[14] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[15] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[16] 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, p. 574, 1959.