目錄

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] 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

由於本論文用到的 units 來自自監督學習的語音表徵，因此在介紹主要研究內容之前，我們會先介紹自監督學習

2.1　深層類神經網路（deep neural network）

深層類神經網路（Deep neural network）是來自於神經心理學家麥氏（McCulloch）等人在 1943 年提出 [12]，取法自生物神經連結的計算模型。以發展此類模型為主軸的心理學流派，在計算認知神經科學中被稱為「連結派（connectionism）」，旨在模擬生物神經系統的連結，以模仿生物的各項功能。爾後在工程界進而透過機器學習的最佳化演算法，使得整個模型能夠藉由資料去貼合（fit）理想的函數，以達成應用或工程上所需要的各種任務。因為該類網路的彈性與計算上易於平行化的特徵，能夠很恰當的利用諸如圖形處理器（graphics processing unit，GPU）等硬體裝置的優勢，以求更好的描述資料分佈、達到前所未有的效能，因此近年在電腦科學的機器學習領域中獲得重大進展，現已成為人工智慧發展的主流。

2.1.1　簡介

2.1.2　訓練方式

2.1.3　前饋式類神經網路

基於深層類神經網路的神經架構有 CNN、RNN、Transformer 等等，由於這些架構在語音與文字處理上都已經被廣泛使用，因此在下面分別介紹：

2.1.4 卷積式（convolutional）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 [14] 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。

在語音處理的領域中，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。

2.1.5 遞迴式（recurrent）類神經網路

2.1.6 序列至序列（sequence-to-sequence）模型

由於許多語言相關的資料都是兩個序列之間互相配對的關係，包含語音和文字等時序訊號等等，都是以時間軸為主要資料呈現的維度。因此這類資料通常會使用序列至序列的模式進行訓練，旨在模擬輸入與輸出序列之間的變化與相依關係（dependency）。

此類模型一般的架構是由一個編碼器（encoder）和一個解碼器（decoder）構成，其中編碼器是將輸入訊號藉由內部表徵（latent representation）進行編碼，依據每個時間點輸入訊號的變化來調整其內部表徵狀態，接著將最後一個時間點的表徵作為整個序列的代表，傳遞給解碼器生成輸出訊號的序列。該過程可由以下

數學式表示：

$$\mathcal{L}_{ST}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{t=1}^{T}\sum_{i=1}^{V} P_{\hat{y}_t}(\hat{y}_t = v_i) \log P_{y_t}(y_t = v_i)$$

2.1.7 專注（attention）機制

原本的序列自序列模型本身。需要讓解碼器單純透過最後一個時間點。的表征資訊來完全儲存輸入序列的一切資訊以工解碼器判斷。并生成輸出序列。然而，由于單就最後一個向量進行判斷對於解碼器而言過於不易。因此。盧氏提出。在編碼器中對輸入序列的不同時間點進行注意力機制亦即讓解碼器可以根據當下所需要輸出的內容判斷應該要重新對輸入序列的哪些部份進行更多的加權。

2.1.8 轉換器（Transformer）

其後，由瓦氏（Vaswani）cite 提出的論文中提出了一個完全由注意機制。所構成的序列自序列模型。原先該模型適用於解決機器翻譯。的問題。由於其能夠高度平行化的特性，日後在自然源處理和語音處理，甚至到電腦視覺領域等近乎整個深層學習的領域都被廣泛的應用。

2.2 表徵（representation）學習與自監督式學習（SSL）

2.2.1 特徵

原本在文字和語音。文字會用 TF-IDF 等等，語音則會用 mel 和 MFCC

2.2.2 表徵學習

後來基於 Mikolov 的 Word2Vec word2vec 是 mikolov (?) 最早提出跟對於文字進行語義表徵的 work。在其後開始嘗試從大量資料去學習出表徵

結合 contextulaized embedding 有了 ELMo

2.2.3 SSL (這邊接下來直接看以前碩論怎麼分。宏毅的再說)

從 contextulaized 的精神，結合 SAttn，BERT 被提出來，並有了 SSL 的概念

SSL 的好處是可以更好的利用 NN 的學習與泛化（generalization）能力，從大量的未標註資料中，就由 pretext tasks 的引導，在 unsupervised 的情形下利用資料本身結構進行學習。

（提到「提出了很多語音基石模型」）

MLM/recon

BERT

CLM/predictive

GPT, APC

contrastive

CPC // 所以這個在後面會不會出事？

2.2.4 離散單元

2.3 本章總結

第 三 章　單一語音離散表徵與語音標記的對應模式

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

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

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

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

3.5.2 基於語音學分類的分析

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] 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.

[12] 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.

[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] "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.