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，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] 是

在近年，已經有不少相關的研究開始嘗試往將離散單元（discrete unit）作為除了連續表徵（continuous representation）以外，可以編碼（encode）語音訊號的另外一種方式。離散表徵（discrete representation）跟連續表徵相比，具有資訊更濃縮（位元率（bit rate）更低）因此更好儲存、處理與傳輸，以及形式上更

像文字的特性。

1.2 研究方向

本論文旨在探討自監督學習模型的離散語音表徵與音位（phoneme）之間的關係。

儘管離散表徵在語音社群（community）中常被當成一種類似文字的存在，另外有一些文獻則是將其當成連續表徵的精簡表示法。

然而

因此，我們藉由分析各種離散單元和人類理解語音最直接的處理層次：音位（phoneme）之間的關係，並將兩者進行各種統計上的序列比較，可以作為訓練大型語音基石模型（foundation model）的分詞（tokenization）基礎，選定最適合的表徵最為系統的輸入符記（token）。

例如 [6] 等作品便是首先嘗試將離散單元進行 tokenization 後做進一步處理的，其後的

1.3 主要貢獻

結果我們發現，藉由觀察這些 unit 並嘗試藉由分詞演算法找出更 high-level 的單位之後，我們觀察到這些機器學習 figure out 的「偽標記」一定程度上的符合了人類音素的特性，因此可以當成某種類似拼音文字的存在。當然這跟人類真正使用的拼音文字仍有距離，因此雖然無法直接當成 exact 的文字使用，一些人類的標注還是需要的，不過透過 ML 我們已經可以盡量更有效的利用珍貴的標注資料，幫助那些尚不容易取得文字的語言發展語音語言科技，以協助保存他們的語言。

1.4 章節安排

由於本論文是以剖析既有的語音離散表徵為主軸，因此就相關研究方面需要從各角度入手，單獨成一章節。接著我們會從單一的離散單元，以及將單元視為像文字的字符（character）並進行分詞演算法兩種對語音離散單元處理的層次分別成章進行分析，最後將這些表徵嘗試做在語音的任務上，以驗證其具有一定的語音表徵能力，且能保留語音學的特徵。

本論文之章節安排如下：

- 第二章：介紹本論文相關背景知識。
- 第三章：介紹 A。
- 第四章：介紹 B。
- 第五章：本論文之結論與未來研究方向。

第 二 章　背景知識

2.1　深層類神經網路（deep neural network）

深層類神經網路（Deep neural network）是麥氏（McCulloch）在 1943 年提出[7]，取法自生物神經連結的計算模型，旨在模擬生物神經系統的連結，以模仿生物的各項功能，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理想的函數，以達成應用或工程上所需要的各種任務。。

以發展此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）」，其後因為該網路的彈性與平行化的能力，和諸如圖形處理器（graph processing unit，GPU）等硬體裝置能夠最好的利用。并能夠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，並因此現已成為人工智慧發展的主流。

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：

2.1.1　卷積式（convolutional）類神經網路

卷積式類神經網路一開始是在 cite 中提出，主要是鑑於影像中的局部性（locality），讓 NN 可以在。在語音中，因為語音訊號的資訊是被呈現在時間維度上，因此通常使用一維的卷積式類神經網路，以捕捉時間維度上的局部性特徵，例如本研究特別探討的 phoneme、morpheme 等等。

2.1.2 遞迴式（recurrent）類神經網路

2.1.3 序列至序列（sequence-to-sequence）模型

由于許多實際上的資料都是 2 個序之間互相配對的關系此類的資料包含語音文字。信號等等，都是以時間軸為主要演變方向的資料。因此有一類模型。會被以序列制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和解碼器構成其中編碼器是將輸入訊號借由內部表征進行編碼。依據每個時間點輸入訊號的順序來改變其內部表征的狀態接著將最後一個時間點的表徵作為整個序列的特征傳遞給解碼器進行輸出訊號生成。

2.1.4 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。

2.1.5 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

2.2.1 文字的語意表徵

2.2.2 語音特徵與表徵

2.3 語音基石模型與自監督式學習

2.3.1 自監督式學習

2.3.2 語音基石模型

2.3.3 離散單元

2.4 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

3.2　相關研究

在 HuBERT 出來之後，

3.2.1 語音表徵的語音學分析

3.3 衡量方式

3.3.1 音位（phoneme）長條圖（bar chart）

3.3.2 純度（purity）、熵（entropy）

3.3.3 對齊（alignment）

3.4 語音學分類

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] 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.

[6] 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.

[7] 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.