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

2.0.1　簡介

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 [**?**]，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　訓練方式

2.1.2　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：

2.1.3 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 [**?**]，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 [**?**] 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。

2.1.4 遞迴式（recurrent）類神經網路

2.1.5 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：

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

2.1.6 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。

2.1.7 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 自監督學習（self-supervised learning，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」）

重建式學習

BERT 以重建被遮罩語言模型

預測式學習

GPT APC

對比式學習

CPC // 所以這個在後面會不會出事？

2.2.4 向量量化（vector quantization）

2.2.5 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。

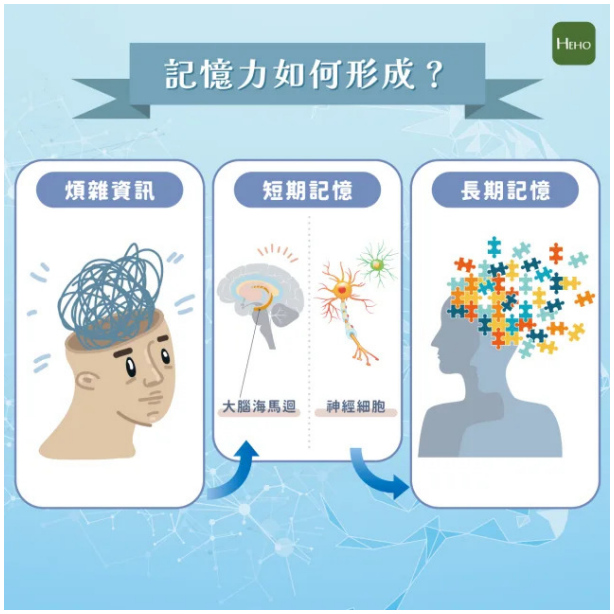

圖 3.1: Enter Caption2

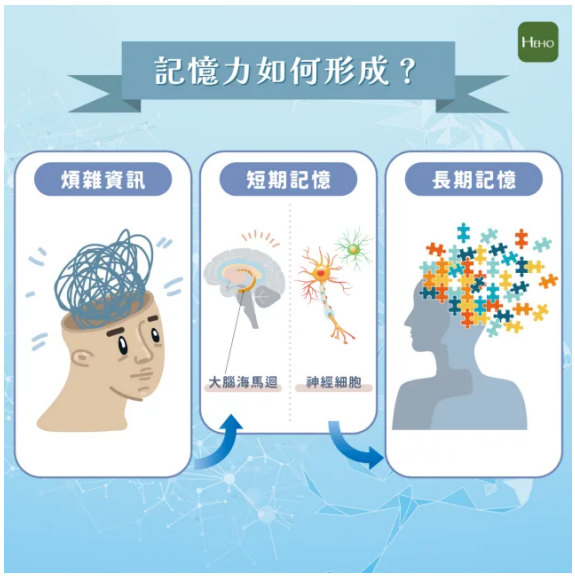

圖 3.2: Enter Captddion

圖 3.3: Enter Caption

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

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.