第一章　導論

1.1　研究動機

語言是人與人之間最主要的溝通方式，是我們對外界世界互動最主要的媒介。隨著時代的發展，在這個資訊爆炸的社會，人們接觸到語言、使用與互動的情形變得非常頻繁。然而，語言的多樣性在此卻成為一項阻礙不同人們交流的障礙，因此，發展語言科技為人們的互動提供協助，因而成為了一項不可避免的趨勢。藉助科技與知識的力量，人們可以透過不同語言的視角去獲取更多元的資訊，激發思考、開拓視野，因此諸如語音辨識、機器翻譯等等，一直都是人們熱衷於研究的議題。

過往，為了達成此一遠大目標，機器學習工程師們與語言學家之間互相合作，期望藉由對領域知識的了解去建立模型，打造出創新的技術以應對語言的各種變化，來滿足人們溝通的需求。近年來，由於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-supervision）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礎模型（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。

……

……

……

然而相較於穩定、易於處理的文字文本，語音訊號的變化複雜萬千，蘊藏了大量如內容、韻律、語者等等不同層次的訊息，使得處理的難度劇增。更何況，目前世界上大約七千多種語言中，絕大多數仍然沒有成熟且普及的文字系統。

（先從語音表徵開始）

因此，除了發展文字的模型以外，語音處理的技術更是必不可少的。而在語音處理這邊，近期藉由自監督學習，提出了許多語音這邊的基礎模型。另外，由於語音的獨特性質的關係

因此在語音處理界，「無文字（textless）」的發展是相當吸引人的。

（加上語言公平性問題？）

近期由於 HuBERT 等等起來的 GSLM 等架構，已經在一定程度上做到完全不依賴文字轉寫、單純在大量的語音語料之上建立一定程度上媲美於「大型語言模型」的成果，甚至達成了閩南語和英語的互相對譯。此一里程碑大力的推動了語言科技的進展，有望推動藉助科技達成的降低語言障礙。

……

不過，即便語音的技術已經相當成熟，在追求模型表現的同時，語音與語言的技術開始與過往的人類對於語言學、語音的理解逐漸產生脫節。似乎在為了讓機器可以擁有良好表現的同時，理解模型如何運作似乎是被犧牲的必然。但人們在追求更好的模型表現的同時，有一群人開始注意到語音處理模型是否有可能抓取到人類語音中特有的、區別於一般音訊的特徵，並嘗試使用過往用來研究、歸類人類語音的方式，結合機器學習與統計學去解釋為什麼，並期求可以比較甚至

改善機器模型在進行語音處理時的表現，不僅僅只是使用資料集本身的分數，而有更多更多元、更穩健的衡量標準。由於離散表徵在當今語音模型與語音處理技術已經愈來愈具備重要性，因此探討與分析為什麼語音離散表徵可以幫助下游任務的背後成因是相當重要的研究方向，其中一個驗證離散表徵能夠幫助模型處理語音訊號的方式，便是驗證其與音位（phoneme）之間的對應關係。

我們想要知道的是，在離散單元推動語音處理發展的同時，它究竟與人類書寫和使用的文字還有多遠的差異，以及在使用上是否能夠達成如同文字的效果，仍舊是領域中尚待探討的議題。所以我們先看看，他是不是起碼符合語音學上，作為「捕捉語音內容」的基本特徵。

（加上跟 LLM 的關係嗎？）

也就是 Phonology 去討論這件事，看它跟聲學特徵跟語音到底有多像，從這個方向我們可以進而去改善說，也許我們可以去用不同的方式使用 Unit ，來進一步推進讓我們的 Unwritten Language 這些技術，能夠更好的跟，就是等於說進而發揮類似文字的效果，然後同時也能促進我們去知道，機器是怎麼樣去理解這些語音訊號，它可能看的是哪些部分。

……

……

……

1.2 研究方向

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。

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 [1] 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。

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 [2] 仿生數學模型，旨在模擬生物神經系統的連結。

深層類神經網路是一個取法自生物神經連結的數學模型，其在計算認知神經科學中以連結派（connectionism）為主要代表，後在電腦科學與機器學習中有不同結構的進展。在此之後，基於其彈性與平行化的能力，能在 GPU 上面很有效率的進行運算並達到前所未有的效能，因此現在已經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。

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，

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

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] 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.

[2] 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.