# 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，因此語言科技發展是一種必然。近年來，由於硬體平行運算技術的進步，深層學習快速崛起成為人工智慧的主流，進而使得大量利用語音和文字的資料發展語言科技成為可能。在這波發展歷程中，藉由大量訓練資料發展而成的基礎模型是這個領域的一大里程碑，不論是文字或是語音都取得了長足的進展，自然語言處理領域甚至出現了幾乎與人類難辨真偽的生成式模型，改變了人們生活的方方面面。

語言是人與人之間最主要的溝通方式，因此語言科技發展是一種必然。……近年來，由於硬體平行運算技術的進步，深層學習快速崛起成為人工智慧的主流，進而使得大量利用語音和文字的資料發展語言科技成為可能。在這波發展歷程中，藉由大量訓練資料發展而成的基礎模型是這個領域的一大里程碑，不論是文字或是語音都取得了長足的進展，自然語言處理領域甚至出現了幾乎與人類難辨真偽的生成式模型，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，加深了處理的難度。更何況在世界上大約七千多種語言中，絕大多數其實仍然沒有成熟且普及的文字系統。因此在語音處理界，「無文字（textless）」的發展是相當吸引人的。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。

## 1.2 研究方向

## 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 [1] 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

我要用這樣的 citation：[1]

# 參 考 文 獻

[1] 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.