目錄

第 三 章　單一語音離散表徵與音位的關係

由於 HuBERT 和 wav2vec 2.0 等語音基石模型的成功，不但在語音任務上達到了前所未有的表現，還促使從語音表徵離散化的想法得以發展。以此產生的「無文字（Textless）」架構，讓人們在處理語音訊號時，有了連續表徵以外的新選擇。離散形式的表徵，可以直接應用文字領域發展的技術，如機器翻譯、生成式模型等，為語音技術帶來新的突破。另一方面，基於離散「符記（token）」的共同形式，離散語音表徵可以更好的整合文字資料，促成多模態領域的發展。跨模態離散表徵的成功，甚至驅使影像領域也開始嘗試發展離散表徵，如探討唇語的 AV-HuBERT [?] 等等，展現了離散表徵在資料處理上的優勢。

此外，除了技術的角度切入，為了探討離散語音表徵成功背後的可能因素，以及它們和語言學對人類語音理解之間的差異，甚至是進而得以利用這些技術協助他們更細緻的探討人類的語音現象。因此，原先在連續語音表徵上的語音學分析，也開始關注離散表徵背後有多能描述對語音現象，將其列入考量，成為除了連續語音特徵和時頻譜之外的另一個選擇。

3.1　相關研究

3.1.1　無文字與離散語音表徵

自從 HuBERT 帶起來的研究之後，愈來愈多離散表徵相關的研究，例如 [這邊要 cite 一些東西] 等等。它們在提出自己的離散表徵時，也會採用原先 HuBERT 提供的那些衡量方式，來驗證這些離散單元確實是與語音中的內容與人類對語音的詮釋具有一定程度的相關性，並從消息理論的角度，證明這些偽標記

找出來的符號之間，確實可以做到區分語音中不同形式的資訊。

3.1.2 語音學分析

另一個層面，由於語音處理本身所針對與探討的終究是人類的語音。因此，有一群研究者通過對人類語音本身的理解，將這些知識應用在分析模型如何對語音訊號建構表徵之上。例如 [這邊要 cite 那些如唐顥等分析語音本身的] 等。

基於這些作品都對語音的離散表徵感到興趣而做出的探討，本論文也先透過過往幾個常用來分析語音表徵的方式，特別是 HuBERT 原始 paper 中提出的標準進行初步的分析。以下介紹此次分析語音表徵的衡量方式：

3.2 衡量方式

首先是純度（Purity）、熵（Entropy）和相互資訊（Mutual Information，MI），這類標準是在原始 HuBERT 論文中採用的指標 [?, ?]，在比對機器學習過程得到的偽標記與人類知識的標註之間，兩者的相關性（Correlation）。以下對各標準進行詳細解釋：

不論是什麼語音基石模型，語音表徵的基本單位是音框（Frame）。因此一段語句（Utterance）的語音的離散單元被表示為 $[y_1, \cdots\cdots, y_T]$。其中 T 是該段語句的音框總數。對於該段語句，若給予一段在音框上對齊的語音學標註（Phonetic Label）$[z_1, \cdots\cdots, z_T]$，此時我們可以將離散單元與標註之間配對的出現次數，寫為一個雙變數的共同分佈（Joint Distribution）

$$p_{yz} = \frac{\sum_{t=1}^{T}[y_t = i \wedge z_t = j]}{T} \tag{3.1}$$

其中 i 是第 i 個音位類別，而 j 指編號為 j 的離散單元。兩個變數的邊際機

率（Marginal Probability）分別為

$$p_z(j) = \sum_i p_{yz}(i,j)p_y(i) = \sum_j p_{yz}(i,j) \tag{3.2}$$

因此，對於每一個音位 i 而言，這個音位對應最可能的離散單元為

$$z^*(i) = \arg\max_j p_{yz}(i,j) \tag{3.3}$$

與之相對應的，對於每一個離散單元的類別 j 則可以找到機率最高的音位

$$y^*(j) = \arg\max_i p_{yz}(i,j) \tag{3.4}$$

於是我們可以計算出以下指標：

3.2.1 純度

本指標考慮音位和離散單元兩個序列之間對應的最高機率，因此從音位與離散單元的角度出發，可以得到以下兩項數據：

音位純度（Phoneme Purity） 考慮每個離散單元對應的音位中，最高機率音位的機率，表示為

$$\mathbb{E}_{p_z(j)}\left[p_{y|z}(y^*(j)|j)\right] \tag{3.5}$$

此指標表示該單元是否對與它對應的音位有足夠的代表性。

分群純度（Cluster Purity） 與音位純度相對，改以每個音位的角度，考慮對應單元類別的機率

$$\mathbb{E}_{p_y(i)}\left[p_{z|y}(z^*(i)|i)\right] \tag{3.6}$$

由於離散表徵進行分群演算法時的類別數是一項超參數（Hyperparameter），且通常離散單元的分群數量會比音位多，因此該統計數據本身不直接具有語音學的解釋意義，而且在分群數量很多時會顯著下降。

然而該指標在考量音位純度時必須一併考慮，因為當分群數非常多時，分群純度過低可能使得音位純度相較失去意義。一個極端的情形是每一個音框都給予不同的離散單元編號，如此音位純度可以達到 100%。但如此一來，離散單元做不到歸納音位類別的目的，音位純度也就失去了意義。

3.2.2 熵和相互資訊

除了純度提供「最高機率」的對應關係，根據 HuBERT 論文 [?] 中的分析方式，我們也可以從資訊理論（Information Theory）的角度，觀察兩個序列的熵和相互資訊。

熵　　熵的定義按照資訊理論，衡量兩個序列中標籤類別出現機率的不確定性（Uncertainty），公式寫作：

$$H(y) = \sum_i p_y(i) \log p_y(i) H(z) = \sum_j p_z(j) \log p_z(j) \tag{3.7}$$

式子中 $H(y)$ 和 $H(z)$ 分別為音位和離散單元的熵。

以音位標準化之相互資訊 (Phone-normalized Mutual Information，PNMI)

本數據以「觀察到某一個離散單元，能降低多少音位標註的不確定性」，定

義該離散單元的出現背後提供了多少音位的資訊。公式寫為：

$$\frac{I(y;z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i,j) \log \dfrac{p_{yz}(i,j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \tag{3.8}$$

$$= \frac{H(y) - H(y|z)}{H(y)} \tag{3.9}$$

$$= 1 - \frac{H(y|z)}{H(y)} \tag{3.10}$$

該項數據愈高，表示離散單元的分群愈是足以提供語音音位的資訊，是一個品質更好的分群結果。由於離散單元能多好的對應到音位才是人們所關心的問題，因此與純度不同，只以音位的角度出發，而不考慮以離散單元分群的角度。

3.2.3 以語音分段指標衡量對齊（Alignment）程度

為了分析離散單元跟音位在語句序列之間的對齊程度，本研究根據 [?] 的方法採用語音分段（Speech Segmentation）的標準去衡量。具體動機為將被分到同一個離散單元編號的音框當成語音分段的同一類別的音位，以此期望可以觀察出在每一段語句中，離散單元出現的順序與範圍，與音位標註指示的範圍一致的程度。

3.3 語音學分類（Phone Type）

3.3.1 簡介

除了單一音位本身的特性以外，由於音位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 [tanghao 等三篇] 的分組方式，對英語的音位進行分類並合併比對數據，看看這些離散單元本身是否有擷取到相似的發聲特徵，而不單純只是把音位分成約 50 類完全獨立的標籤。

英語中的音位分為元音（Vowel）與輔音（Consonant）兩大類別，其中又可依照發音的共同特性一共分成七個類別。

元音　根據發音的位置是否發生改變，英語的元音可分為：

- 單元音（Monophthong）
- 雙元音（Diphthong）

兩大類別。

輔音　而輔音按照發音的方式，可分為以下五類：

- 爆破音（Plosive）
- 擦音（Fricative）
- 塞擦音（Affircate）
- 鼻音（Nasal）
- 近音（Approximant）

3.3.2　解釋意義

- 純度（Purity）：換成以語音學的類別作為新的語音標籤後，有何變化（關聯性更強？）

- 熵（Entropy）（放直方圖解釋）→ *Alignment*

3.4 實驗集與分析模型

本研究的分析對象參考無文字架構 [?, ?, ?] 研究，採用當中提及的 CPC、Wav2vec 2.0 和 HuBERT 三個語音基石模型，並與作為比對的聲學特徵梅爾時頻譜（Mel-Spectrogram），一共四種語音表徵模型。

((寫一下 resolution 跟層數)) (3.11)

此後亦跟隨該研究選用特定模型層數，和其釋出的 K-平均量化模型。這些模型層數與量化為該研究中被證明與語音學特徵最相關，且被使用於無文字架構後續研究之語音離散表徵的抽取方法。無文字研究中已透過 (某某) 語料對四種語音表徵進行 K-平均分群演算法，分別得到群數為 50、100 和 200 的三個量化模型。

本論文以公開的 LibriSpeech 資料集為分析對象，採取其 train-clean-100 為分析的語音語料庫。本研究將語音語料庫的語音資料經過四個模型獲取連續表徵後，再經過量化模型得到完全由離散單元組成的「偽文字」語料。

針對語音學的音位標註，吾人透過 Montreal 強迫對齊器（Forced-Aligner）[[CITE montreal]] 的英語預訓練模型，從語料庫的文字轉寫取得語音資料的音位標註與對應的時間範圍。最後透過語音表徵各自的時間解析度生成以音框為單位的音位標註語料。最後將兩者對語音資料集進行音位標註相關性的分析。

3.5 分析結果

3.5.1 基於各自音位的分析

由表 ?? 中可以看出，分群的群數愈多時，音位的純度確實有所上升，但這可能是犧牲分群純度得來的。因此再看 PNMI 的指標可以發現，整體離散單元和

	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.5256	0.3382	3.3152	3.8681	0.4993
wav2vec 2.0	0.4006	0.2676	3.3152	3.8215	0.3706
CPC	0.5188	0.3812	3.3146	3.7918	0.4992
LogMel	0.3253	0.1473	3.3158	3.8630	0.2647

(a) 群數 = 50

	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.6097	0.2553	3.3152	4.5704	0.5786
wav2vec 2.0	0.4877	0.2118	3.3152	4.5284	0.4596
CPC	0.5895	0.2674	3.3146	4.5034	0.5557
LogMel	0.3348	0.0931	3.3158	4.5591	0.2789

(b) 群數 = 100

	音位純度	分群純度	音位熵	離散單元熵	PNMI
HuBERT	0.6474	0.1644	3.3152	5.2681	0.6289
wav2vec 2.0	0.5427	0.1467	3.3152	5.2173	0.5188
CPC	0.6098	0.1789	3.3146	5.1885	0.5882
LogMel	0.3474	0.0569	3.3158	5.2322	0.2955

(c) 群數 = 200

表 3.1: 不同群數在四種基石模型的分析數據

音位標註的相關性還是有所提升的。

此外，就不同模型來觀察，HuBERT 的表現是四種語音表徵之中最好的，一定程度上可以佐證 HuBERT 在找出語音中有意義單位上的效能，以及為什麼無文字架構通常以 HuBERT 當成抽取語音離散表徵的模型。

3.5.2 基於語音學分類的分析

(放數據)

3.6 本章總結

本章節探討以音框為單位取出的語音離散表徵與對應的音位標註之間的關係，從分析結果中可以得到，HuBERT 模型的離散表徵確實與人類理解的語音單位「音位」之間，具有最明顯的相似性，也進一步證明為何 HuBERT 目前是抽取語音離散表徵時最常使用的模型。