目錄

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，例如從音位（Phoneme）、構詞（Morphology）、語法（Syntax）等等用以刻劃人類語音和語言特性的概念，將之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] 基於模仿人類學習語言的過程，藉助諸如 CPC（[6] ）、HuBERT （[7] ）、wav2vec 2.0 （[8] ）等自監督學習模型的幫助，引入向量量化的技術，提出了「無文字（Textless）」的學習架構，轉而以語音表徵量化後的「離散單元（Discrete Unit）」作為操作對

象，企圖以單純大量的語音資料中訓練出一個不依賴文字的語言模型。此種學習架構的優勢在於在能保有利用大量未標注文字轉寫語音資料的同時，與連續表徵相比資訊的位元率（Bit Rate）利用更有效率、容易儲存、處理與傳輸，以及形式上更像文字的特性，因而可以將其視為一種「機器自己學習出來的文字」，接下來借用長久以來只能在自然語言處理（Natural Language Processing，NLP）領域中各種語言模型（Language Model）的相關技術和任務的解決方法，套用在語音處理的領域中，期望可以像文字那樣從大量的語音資料中，找尋出「語音訊號版本的文字」。自此之後有一系列如應用於英語和閩南語之間的語音到語音翻譯 [9] 等等使用離散單元（Discrete Unit）進行任務訓練的研究，一定程度的印證了這些離散單元捕捉語音內容的效果。

儘管離散單元在編碼語音之上固然有不錯的效果，並有相關研究展現了離散單元具有一定程度上與文字的相似性，然而其作為「完全文字的替代」仍然有相當的距離。借鑑過往在自監督學習的語音表徵出來之後，便嘗試重新從語言學（Linguistics）的概念汲取靈感，對其進行語音學（Phonetics）層面的分析。本論文期望初步結合原先 HuBERT 中從消息理論（Information Theory）的統計數據，結合語音學分析的視角，對於離散表徵（Discrete Representation）本身與音位（Phoneme）和語音類別（Phone Type）之間的關係進行相關性的統計與分析，期望可以對 HuBERT 等自監督學習表徵進行量化（Quantization）後所得的離散單元所編碼、擷取到的資訊是什麼有較為深入程度的了解。

1.2 研究方向

本研究論文為了探究離散單元本身是否具有潛力可以單純透過大量語音資料的自監督學習與統計過程，從文本中找尋出語音中更精細的結構，乃至於類似

文字或是從語言學（Linguistics）等人類知識領域定義出的「離散單位」— 如音素（Phone）、音位（Phoneme）、字符（Character）、「詞綴與字根」（即「詞素（Morpheme）」）或單字（Word）等等。因此，本研究取法自 HuBERT 本身為了證明其離散單元具有一定的「聲學單元（Acoustic Unit）」特性的「純度（Purity）」和「相互資訊（Mutual Information，MI）」的分析數據作為分析離散語音表徵和「音位」— 作為人類知識理解語音中最基礎的單位 — 之間相關性（Correlation）的參考。

此外，基於訊號速率（如序列的長度）的考量，結合在文字處理中如 BPE 等等常見的次詞單位（Subword）分詞（Tokenization）演算法，基於形式上的相似性，因而也可以套用在像是 HuBERT 離散單元這種離散的符號上，將離散單元序列中相似的規律（Pattern）發掘出來。近期如 Wav2Seq [10]、[11]、[12] 等作品也先進行了類似的嘗試。本論文則是在除了經驗上（Empirically）將其用於大量資料訓練的視角以外，從「將其視為另一種離散單位」的觀點進行統計數據的量化分析（Quantitative Analysis），作為在計算資源有限的前提下決策數據編碼的一個判斷標準。

1.3 主要貢獻

本論文達成的主要成果是以更細緻的方式，對現在愈來愈廣為使用的離散單元以音位和語音類別等語音知識的視角給出一個基礎相關性的分析方法，並將單一離散單元本身與將多個單元透過分詞演算法（Tokenization）重新編碼前後進行比較，初步試探離散單元與音位之間的關係，並期望作為「離散單元可否一定程度上的『被視為文字』或『有機會從中發掘出文字單位』」的判斷基礎，為往後研究往語音語言模型（Spoken Language Model）中「對語音編碼」這個重要的程

序，提供一個在實際上開始耗費資源的模型訓練之前，可比較的判斷標準。

1.4 章節安排

本論文將以如下的方式進行章節安排：

- 第二章：介紹後面章節所需要的與深層學習（Deep Learning）、表徵學習與自監督學習相關的基礎背景知識。
- 第三章：從介紹離散單元本身提出後，「無文字」的相關前作文獻開始，帶出對從無文字系列作品用到的各種自監督學習模型抽取之離散單元本身的純度（Purity）和相互資訊（Mutual Information，MI）等統計數據，進行比較與分析。
- 第四章：講述為何單一離散單元本身或許不全然足夠發掘出類似音位進而對應到文字的單位，以及近年人們嘗試以離散單元為基礎，透過分詞演算法（Tokenization Algorithm）發展之聲學片段（Acoustic Piece）的進展，接著我們將單元進行分詞法重新編碼處理前後，觀察數據上與第三章結果間的差異，以論證對離散單元進行分詞是否可以找出更接近音位的單位，驗證「離散單元可被文字化」或「離散單元學到的是否為更精細的語音訊號規律或結構（Structure）」等論述。
- 第五章：總結前面的觀察結果，並進一步探討本研究還可以如何延伸，並怎麼幫助語音語言模型的發展。

第 二 章　背景知識

2.1　深層類神經網路

2.1.1　簡介

深層類神經網路（Deep Neural Network，DNN）又稱「人工類神經網路（Artificial Neural Network，ANN）」，是由神經科學中連結主義（Connectionism）學派的麥氏（McCulloch）與皮氏（Pitts）等人，在 1943 年 [13] 提出的計算模型。該學派主張藉由模仿生物神經連結的方式建立模型，以建立系統模擬各種心智現象，而其中深層類神經網路由於它在貼合目標函數上的彈性和運算上易於平行化的特徵，得以恰當利用圖形處理器（Graphics Processing Unit，GPU）等硬體裝置的優勢，能夠很好的結合最佳化演算法對貼合（Fit）資料分佈、找出最好的函數（Function），在各類任務與應用上取得前所未有的效能，因而近年為機器學習與電腦科學領域獲得重大進展，現已成為人工智慧發展的主流。

深層類神經網路最基本的單位是「神經元（Neuron）」，其本質為一個線性分類器，為了模擬生物神經細胞接收訊號、處理到傳出的過程。每個神經元會接收一串數字 $(x_1, x_2, \cdots\cdots, x_N)$ 作為輸入後，得出一個數字 y 作為輸出。其關係可用下列數學運算式描述：

$$y = \sigma(w^\top x + b)$$

其中輸入 $x = (x_1, x_2, \cdots\cdots, x_N)$ 被描述為一個 N 維向量，該神經元對每個輸入取值分別給予權重（Weight）$w = (w_1, w_2, \cdots\cdots, w_N)$ 相乘後，再對加權平均的結果加上偏差值 b 當成線性輸出值。最後，為了描述分類器函數的非線性（Non-linear）特性，類似神經細胞觸發與否的過程，此輸出值通常會經過激發函

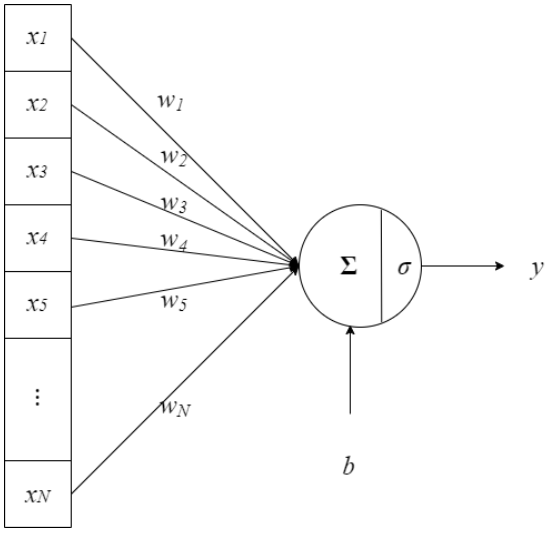

圖 2.1: 神經元示意圖

數（Activation Function）σ 的轉換後才得到最終輸出值 y。常見的激發函數包含線性整流單元（Rectified Linear Unit，ReLU）、S 函數（Sigmoid Function）或雙曲正切函數（Hyperbolic Tangent Function，tanh）等等。

其後羅氏（Rosenblatt）在 1958 年 [14] 提出感知器（Perceptron），本質上是結合數個神經元的運算，來實現更為複雜的函數功能。基於數學理論中的通用近似定理（Universal Approximation Theorem）[15]，感知器理想上可以近乎模擬一切函數。然而後續研究發現單層感知器具有線性不可分[1]的限制，於是出現了在輸入與輸出函數之間增加隱藏層（Hidden Layer）的多層感知器（Multi-layered Perceptron，MLP），由於在輸入與輸出之間的多層感知器可以藉助隱藏層的幫助實現函數的多次轉換，因而大大拓展了模型的適用範圍。此種計算模型由於是「加深隱藏層」得來的，因而被稱為深層類神經網路（Deep Neural Network）。

然而，單純擁有能夠表達複雜函數的模型是不夠解決現實複雜的工程問題的。為了增加函數貼合的效率，魯氏（Rumelhart）與辛氏（Hinton）等人 [16, 17] 提出了反向傳播（Backpropagation）演算法，期望藉由計算輸出層與目標

[1] 例如無法貼合異或（Exclusive OR，XOR）運算等函數

函數之間的誤差（Error），透過最佳化演算法計算出梯度（Gradient）後，經由隱藏層反向往輸入層，對整個類神經網路進行修正。於是配合圖形處理器平行運算的能力，資料中貼合函數的過程變得更有效率。如此透過深層類神經網路，找出資料輸入與輸出之間的函數關係的機器學習演算法，就稱之為深層學習（Deep Learning）。由於深層學習的可擴展性（Scalibility）與泛用性（Generalizability），不論在圖像、語音、文字等多個模態，深層類神經網路都已經獲得了廣泛應用，解決更加複雜的現實問題。

實際上，根據資料特性的不同，並不是所有的資料都能單純的適用於這類輸入與輸出向量直接對應的模式，因此類神經網路又發展出不同的架構以適應資料本身的特性。前述的類神經網路由於運算過程單純是從輸入層經由多層感知器直接進行矩陣運算完成函數的模擬，因此被稱之為「前饋式類神經網路（Feed Forward Network，FFN）」。為了適應各種資料型態的特徵，藉由調整各神經元之間的連接關係，後續發展出了如卷積式（Convolutional）、遞迴式（Recurrent）與轉換器（Transformer）類神經網路等架構，以應對不同任務的需求。由於這些架構在語音與文字處理上相當常用，接下來將一一分別介紹：

2.1.2 卷積式類神經網路

卷積式類神經網路（Convolutional Neural Network，CNN）為 1998 年由楊氏（LeCun）[18] 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 [19] 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。

有別於圖像中經常是以像素（Pixel）的三原色亮度進行卷積運算，在語

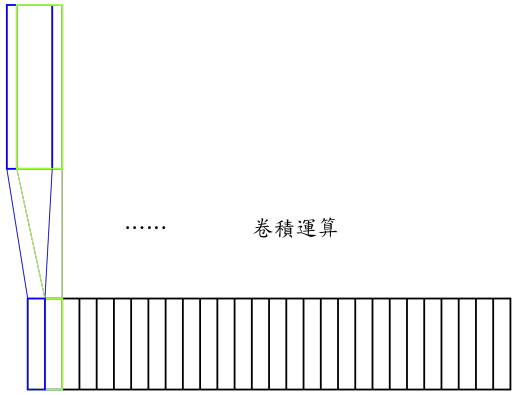

圖 2.2: 卷積神經網路作用於語音訊號示意圖

音中卷積式類神經網路的處理對象，除了直接是空氣壓力波形的物理訊號以外，為了更方便機器模型判斷語音訊號的內容，透過聲學知識得到的聲學特徵或深層學習得出的語音表徵，也經常是語音處理中卷積層運算的對象。然而不論是何種輸入，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。

2.1.3 遞迴式類神經網路與序列至序列（Sequence-to-sequence，Seq2seq）模型

遞迴式類神經網路

不同於運算過程由輸入往輸出單向的前饋式和卷積式類神經網路，為了處理有記憶和狀態的資料，特別是會隨時間變化的序列資訊，在語音和文字的機器學習中，會將輸出訊號重新接回輸入層的遞迴式類神經網路（Recurrent Neural Network，RNN）是一個相當符合語言特性的選擇。此種網路以每個時間點（Timestep）為考慮對象，在每一步會對輸入層的向量進行運算後，不但

將此結果算出一個輸出向量，還會得到另外一些數據保留作內部狀態，表示此前經歷過所有序列資料的記憶。常用的遞迴式類神經網路類型有長短期記憶（Long Short-term Memory，LSTM）和閘門循環單元（Gated Recurrent Unit，GRU）等。

此類類神經網路通常會以下列介紹的序列至序列的形式被用在如語音辨識、語音合成或機器翻譯等和語言密切相關的任務中。

序列至序列（Sequence-to-sequence，Seq2seq）模型

由於許多以語言為主的資料經常以兩個序列互相配對的形式呈現，因此專門用以處理此類資料的模型被特別稱為序列至序列模型。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，旨在模擬輸入與輸出序列之間的變化與相依關係（Dependency）。

此類模型一般有兩種模式：

其一是每個時間點都取得一個輸出的向量，用在輸入與輸出等長的任務之中，此模式又被稱為符記分類（Token Classification）。但更常見的狀況是，輸入與輸出兩者序列長度並不總是相同，此時典型的作法是，讓編碼器將輸入序列在每個時間點一一與模型進行運算，藉由內部表徵（Latent Representation）的調整對整個輸入序列進行編碼，完成後將最後一個時間點的表徵作為整個序列的代表，此表徵向量會被稱為「語境向量（Context vector）」，接著被傳遞給解碼器依時序生成輸出訊號的序列。

2.1.4 專注機制(Attention Mechanism)與轉換器(Transformer)類神經網路

專注機制

由於 RNN 本身需要編碼和解碼的資訊量是整個序列，對時間點距離比較遠的輸入容易被遺忘，也就是難以處理長期相依性（Long-term dependency）的問題。為解決這種困境，Luong 等人提出了「專注機制（Attention mechanism）」，讓解碼器除了依據語境向量的資訊以外，還可以對輸入序列的不同時間點分配權重，在生成輸出序列時重新從輸入序列中得到所需的訊息。

專注機制一般涉及三個向量之間的運算：詢向量（Query）、鑰向量（Key）和值向量（Value），其運算式如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

序列至序列模型透過專注機制的引入，能夠更好的分配輸入序列的運算，因而大大改善了如語音辨識、機器翻譯等任務的效能。

轉換器類神經網路

儘管遞迴式類神經網路本身善於處理時序資料，然而它難以平行化的架構限制卻大大束縛了其在訓練和推理（Inference）時的效率。由專注機制獲取靈感，2017 年瓦氏（Vaswani）等人在 ___cite___ 提出了一種完全由專注機制構成，不需依賴遞迴運算的序列至序列模型，用以解決最經典的機器翻譯任務。

轉換器一樣沿用了專注機制三組向量的運算邏輯，以位置編碼（Positional Encoding）的方式對序列中每個位置的時間點進行編碼，取代原先在遞迴式類神經網路需要一步一步運算的過程，在實行平行計算的同時也能考慮到資料在不同

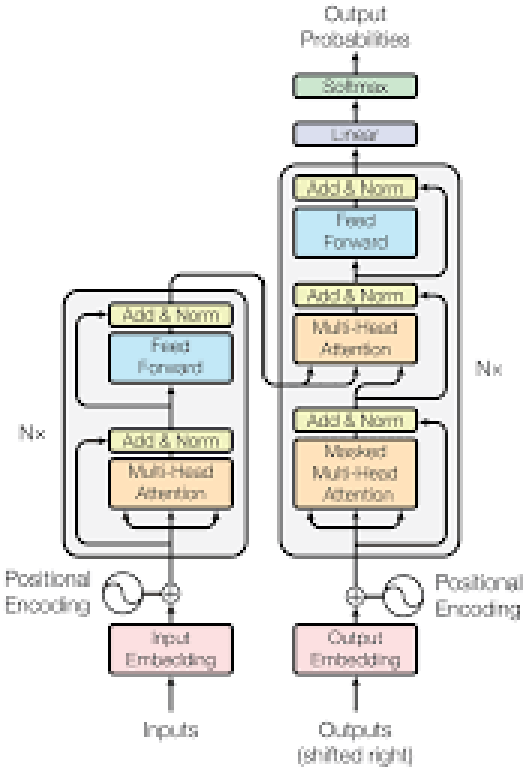

圖 2.3: 轉換器架構圖

時間點出現的效應。其整體架構如圖 2.3

由於轉換器不需對每個時間點一一運算，使其得以實現高度平行化的優勢，類神經網路得以透過專注機制的幫助同時進行序列資料的大量訓練，這種可擴展性（Scalability）因而在自然語言和語音處理都獲得了巨大的進展，近乎取代了原先遞迴式類神經網路的應用場景，近年甚至被電腦視覺的研究者推廣應用在圖像類的資料上（___vit___），足以展現此種模型架構的彈性與泛用性，是目前最前沿人工智慧的主流架構。

除了模型架構，機器學習中不可或缺的另一大部分即是對資料的編碼過程。如何更有效率的讓機器可以理解、處理和輸出，是機器學習乃至深層學習的一大課題。面對捉摸不定、抽象且變化萬千的人類語言，語音和文字處理如何對資料去蕪存菁，表徵學習更是重中之重。

2.2 表徵與自監督式學習

2.2.1 特徵抽取與表徵學習

不論採用何種模型，為了讓機器可以處理並捕捉輸入資料中的訊號與模式（Pattern），如何對資料編碼和運算的步驟，在機器學習稱之為特徵抽取（Feature Extraction）或表徵學習（Representation Learning），是模型建構時不可或缺的重要步驟。

對於抽象的語言概念，早期工程領域根據對語音和文字的理解，分別進行了不同的處理。對離散因而可以一個一個計數的文字，人們使用詞頻統計衍生出如 n 連詞（n-gram）、TF-IDF（Term-Frequency Inverse Document Frequncy）等特徵當成模型學習的前處理步驟；而對於連續又複雜的語音，工程師則是透過聲學原理與訊號處理的知識，使用如濾波器組（Filter Bank）、梅爾倒頻譜係數（Mel-Frequency Cepstrum Coefficient，MFCC）等特徵，當成人耳捕捉語音訊號過程之類比。

在深層學習逐漸發展的過程中，自然語言處理領域的一大里程碑，是米氏（Mikolov）提出的「word2vec」模型 [20]，以連續的向量表徵（Vector Representation）取代稀疏（Sparse）的統計數據，對離散的文字單詞進行「詞嵌入（Word Embedding）」編碼，並透過大量的文本運算，將各單詞之間的共現（Collocation）以跳躍詞（Skip-gram）、連續詞袋（Continuous Bag-of-Word，CBOW）等演算法轉換成高維向量空間中的點，找出每個單詞最適合的語義表徵。爾後，為了更細緻的捕捉同一單詞在不同句子中可能的脈絡變化，ELMo（Embeddings from Language Model）[21] 提出了「脈絡化詞嵌入（Contextualized Embedding）」的概念，使得各個單詞在運算出表徵的過程可以根據上下文進行些

微調整。

2.2.2 自監督學習

隨著轉換器模型的提出後，BERT（來自轉換器的雙向編碼器表徵，Bidirectional Encoder Representations from Transformers）[22] 被提出，從大量的文本與自專注機制之中，工程師們便可不藉由人工標記，透過預先設定任務（Pretext Task）的引導，使得模型可以自己從大量文本中自行找出更細緻且考量脈絡（Contextualized）的語義關係，並在許多文字的任務上獲得了超越以往的成績。

自此，楊氏（Yann LeCun）將此種以特定的任務作為引導，藉助資料本身的結構替代標註，以從大量的未標註資料中進行學習資料模式（Pattren）訓練方式，被稱之為「自監督學習（Self-supervised Learning，SSL）」。BERT 的成功使得自監督學習自此大行其道，並出現了許多由巨量資料進行預訓練（Pre-train）而成的基石模型（Foundation Model），很好的解決在語言處理領域中資料飢餓（Data Hungry）的問題。此後人們在解決語言相關任務時，便不需要從頭蒐集資料與進行耗時耗能的訓練過程，而是可以透過基石模型優良的泛化（Generalization）能力，找出對應各種應用任務的資訊予以解決。相比於預訓練的任務，這些更貼近日常現實的任務被稱為「下游任務（Downstream Task）」，而可以應對廣泛的下游任務種類，則是這些基石模型最大的優勢。

有鑑於文字處理方面的成功，語音領域的研究者便嘗試將相似的模式套用於語音之上，眾多語音基石模型也隨之出現。因為大量的語音資料庫本身，可以幫助模型去萃取出有助於下游任務的語音表徵（Speech Representation），並在各式任務上獲得了超越傳統聲學特徵的表現。由於語音表徵具備的無窮潛力，已

經逐漸成為聲學特徵以外，用來處理語音訊號時的新選擇。

依照這些語音自監督模型在預訓練的學習模式，大致可以分為重建式、預測式與對比式模型，以下分別按照這三類模式介紹這些語音基石模型：

重建式學習（Reconstruction Learning）

此類模型藉由對輸入訊號進行擾動（Perturb）後，期望模型將被更動的輸入重新預測回原始的資料，通常減損函數表示為：

$$\mathcal{L}_{recon} = \mathbb{E}_x[|f_\theta(\tilde{x}) - x|]$$

其中 $\tilde{x}$ 為 x 擾動後的資料，$f_\theta(\cdot)$ 為模型代表的函數。擾動的方式可能以遮蔽為主要方式，在文字處理以 BERT 為代表，因此又被稱為「遮蔽語言模型（Masked Language Model，MLM）」。在語音中採用此方式學習的有 Mockingjay [23]、TERA [24] 等模型。

預測式學習（Predictive Learning）

這類模型透過預訂一些學習目標的函數，製造類似輸入與輸出的配對資料，讓模型去預測該函數的結果，來學習資料中的特定結構。其訓練減損函數可表示為：

$$\mathcal{L}_{pred} = \mathbb{E}_x[\text{eval}(f_\theta(x), \hat{f}(x))]$$

其中 $\hat{f}$ 是期望模型學習到的目標函數，$f_\theta(\cdot)$ 為模型代表的函數，eval 則是用以評估此預測好壞的標準。

目標函數最典型的代表是單純的自迴歸（Auto-regressive），也就是期望模型可以預測未來時間點的輸入表徵。文字方面以生成式預訓練轉換器（Generative

Pretrained Transformer，GPT[25, 26]）系列為代表，而語音上的 APC [27] 也是採用此種模式。

此外，語音基石模型還可以使用其他的訓練目標，如 PASE+ [28] 要求模型得以預測他種模型的表徵，而本文著重探究的隱藏單元 BERT（Hidden-unit BERT，HuBERT）[7, 29], 則是以預測輸入表徵進行分群（Cluster）後的結果。HuBERT 預測的目標又被視為偽標註（Pseudo-label），在後面將特別細究相關內容。

對比式學習（Contrastive Learning）

這種學習方式的訓練目標是要求模型區分正樣本（Positive Sample）與負樣本（Negative Sample）的差異，減損函數通常定義為：

$$\mathcal{L}_{contr} = -\mathbb{E}_x \left[\log \left(\frac{\sum_{\tilde{x} \in x_{pos}} \exp(\text{sim}(x, \tilde{x}))}{\sum_{\tilde{x} \in \mathcal{X}} \exp(\text{sim}(x, \tilde{x}))} \right) \right]$$

其中 x 為輸入，x_{pos} 為正樣本，$\mathcal{X}$ 為整個包含正負樣本的資料集；$\text{sim}(\cdot, \cdot)$ 是評估兩個樣本之間相似程度的函數，最常使用的相似度函數為內積運算得出的餘弦相似度（Cosine Similarity）。語音上最早使用對比式學習的模型為對比預測編碼（Contrastive Predictive Coding，CPC）[30]，之後如 Wav2vec [31]、Modified CPC [32]、Wav2vec 2.0 [33] 等等模型亦是以對比正負樣本的模式訓練，只是訓練時正負樣本的定義有差異，如 Wav2vec 僅以時間維度上相同的向量為正樣本，其餘則以固定一段時間內的向量皆為正樣本。

對比式學習藉由正負樣本的定義，將預訓練任務形塑成分類問題，因此減損函數本質上為交叉熵（Cross-Entropy），使模型可以將訓練資料中的結構差異判斷出來。

2.2.3 向量量化與離散單元

語音訊號雖然本身記錄的是語言的資訊，然而卻和影像資料一樣都是連續數值的資料，不像離散的文字相較之下易與處，因而發展出了許多應用廣泛的模型。為了使語音模型的訓練可以套用自然語言處理領域的演算法，從連續語音中找出離散的表徵因而逐漸發展起來，而這類研究又被稱為「聲學單元發掘（Acoustic Unit Discovery，AUD）」。

由於語言概念本質上是有離散符號的，因此向量量化的技術常被用在學習需要涉及語言標註的情境之下，如電腦視覺經典的量化向量變分自編碼器（Vector-Quantized Variational Autoencoder，VQ-VAE）[34]，便是利用影像標註是離散語言單詞的特性，使得模型學習出來的表徵向量被約束在編碼簿（Codebook）的幾個向量之中。

在語音的領域，基於 Wav2vec 之上，Vq-wav2vec [35] 和 Wav2vec 2.0 將連續的語音特徵量化加入訓練的目標中，在語音辨識等任務上獲得了不少的進步。

HuBERT 則是應用事先對連續的 MFCC 特徵進行 K-平均（K-Means）演算法分群後，以所得的群心（Centroid）編號作為訓練目標，實施類似文字 BERT 的遮蔽語言模型訓練，並將第一次訓練得到的語音表徵再次分群後替代 MFCC 特徵再次訓練。這些經過兩輪訓練後，從模型表徵分群得到的群心即被視為「隱藏單元（Hidden Unit）」，編碼了語音訊號中較為代表性的若干個聲學特徵。而透過找出隱藏單元的過程，HuBERT 可以在低資源的情形下達到與 Wav2vec 2.0 接近的語音辨識成績。

2.2.4 離散單元與無文字（Textless）架構

奠基於 HuBERT 等語音基石模型的成功，Meta 利用隱藏單元的想法，將大量的語音資料表徵進行 K-平均演算法，作為這些語音訊號的偽標籤。如此得到的大量離散隱藏單元便形成了形同「偽文字（Pseudo-text）」的語料庫，於是他們基於這些離散單元作為文字訓練語言模型，稱為「生成式口語語言模型（Generative Spoken Language Model，GSLM）」。配合反向以語音合成的方式訓練一個基於離散單元的語音生成模型，整體架構便能完全不依賴文字標註，就訓練出一個純語音的語言模型，因而被稱為「無文字（Textless）架構」[36]。

無文字的模式目前在語音問答（Spoken Question Answering）[37] 跟語音到語音翻譯 [9] 獲得了前所未有的進展。自此這類「離散單元（Discrete Unit）」被視為一項類似文字卻不需要真的依賴人類文字標記的語音表徵，以儲存所需的位元率（Bit Rate）較低與可以套用文字的「語言模型」之訓練模式作為最大優勢為語音社群廣泛借鑑，後續也出現了許多如 [38] 等等嘗試將語音以離散表徵編碼的研究。

然而，雖然在系統與應用任務上獲得了很大的成功，但這些離散單元本身究竟與文字存在多少差異，或能夠多少的幫助語音語言模型的訓練與建立，仍然是目前本領域探討的焦點議題。有鑑於此，本論文基於語言知識，從最接近文字但又跟語音訊號最密切相關的「音位（Phoneme）」開始探討，期望對離散單元本身究竟能夠帶給我們什麼特徵、如何幫助後續應用進行進一步的研究。

2.3 本章節總結

本章節先是對深層學習模型的核心部件 — 類神經網路進行了基本原理的介紹，其後對本論文研究的核心 — 「語音表徵」與「離散單元」的發展演進與歷史進行

了簡單的梳理。此後兩章節就會緊扣著這些基石模型得到的離散特徵，將其與尤其是「音位」這類語音學標記之間的統計關係進行更進一步的分析。

第 三 章　單一語音離散表徵與語音標記的對應模式

（Ch 3 好像要講一下為什麼不做連續特徵 & 為什麼要以 phoneme 為客體了 T_T）

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。

由于 HuBERT 模型在使用离散单元进行语音处理方面取得了显著成功，为了研究这些单元为什么能够如此有效地训练模型，我们从离散表征本身的特征分析起。探讨这些单元是否能够替代文字作为语音数据的表征，并分析其与语言学标签之间的关系。

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。

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。

相关研究

近期有多项相关研究尝试在 SSL 表现优异之后探讨其原因，例如 Hsu 等人的研究。他们的研究表明，通过使用 k-means 聚类生成的隐藏单元作为预训练的目标标签，HuBERT 模型能够在低资源条件下取得卓越的语音识别性能。这一发现表明，pseudo-label 技术在自监督学习中起到了关键作用。

语音表征的语言学分析 在 HuBERT 出现之后，一些研究试图从统计和语言学领域知识角度分析语音表征，例如 Wells 等人的研究。这些研究通过分析离散表征与语言学标记的对应关系，探讨模型为何能够如此有效地捕捉语音数据的结构和模式。常见的方法包括 ABX 测试、t-SNE 降维分群等。

无文字（Textless）语言模型 本论文以 GSLM textless 采用的模型 units 为主要分析对象，企图探讨这些 unit 是否符合语言学特征，或者有其他特征。

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

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

cluster purity：每個 phoneme 對應到的 cluster 統計，若 cluster purity 低代表 less liguistical meaning? 單一 phoneme 本身對應的 unit 的一致性。如果

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

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

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

$$\frac{I(y;z)}{H(y)} = \frac{\sum_i \sum_j p_{yz}(i,j) \log \frac{p_{yz}(i,j)}{p_y(i)p_z(j)}}{\sum_i p_y(i) \log p_y(i)} \tag{3.1}$$

$$= \frac{H(y) - H(y|z)}{H(y)} \tag{3.2}$$

$$= 1 - \frac{H(y|z)}{H(y)} \tag{3.3}$$

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

要計算 phoneme alignment ，我們參考 CITMEALIGNM 本身需要考慮 P 值、R 值等，以下說明如下：

(算 alignment 的一些介紹)

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

(放數據)

3.5.2 基於語音學分類的分析

(放數據)

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, publisher: IEEE.

[3] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12 644–12 652, Jun. 2023, number: 11. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26488

[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)*, Oct. 2023, pp. 1937–1942, iSSN: 2640-0103. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10317485

[5] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On Generative Spoken Language Modeling from Raw Audio," *Transactions*

of the Association for Computational Linguistics, vol. 9, pp. 1336–1354, 2021, place: Cambridge, MA Publisher: MIT Press. [Online]. Available: https://aclanthology.org/2021.tacl-1.79

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 2019, arXiv:1807.03748 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1807.03748

[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)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6533–6537. [Online]. Available: https://ieeexplore.ieee.org/document/9414460/

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[9] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong, H. Inaguma, S. Popuri, C. Wang, J. Pino, W.-N. Hsu, and A. Lee, "Speech-to-Speech Translation for a Real-world Unwritten Language," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds.

Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4969–4983. [Online]. Available: https://aclanthology.org/2023.findings-acl.307

[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)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096988/

[11] S. Ren, S. Liu, Y. Wu, L. Zhou, and F. Wei, "Speech Pre-training with Acoustic Piece," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2648–2652. [Online]. Available: https://www.isca-archive.org/interspeech_2022/ren22_interspeech.html

[12] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe, Y. Fujita, T. Maekaku, P. Guo, Y.-F. Cheng, P. Denisov, K. Saijo, and H.-H. Wang, "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)*, Apr. 2024, pp. 11 481–11 485, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/10447929

[13] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5,

no. 4, pp. 115–133, Dec. 1943, publisher: Springer. [Online]. Available: https://doi.org/10.1007/BF02478259

[14] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, publisher: American Psychological Association. [Online]. Available: https://doi.apa.org/doi/10.1037/h0042519

[15] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, Jan. 1989. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0893608089900038

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/323533a0

[17] D. E. Rumelhart and J. L. McClelland, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, 1987, pp. 318–362. [Online]. Available: https://ieeexplore.ieee.org/document/6302929

[18] 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, Jan. 1998, conference Name: Proceedings of the IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/726791

[19] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013, arXiv:1301.3781 [cs]. [Online]. Available: http://arxiv.org/abs/1301.3781

[21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[23] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay:

Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders," Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.12638v2

[24] L. T, LiShang-Wen, and LeeHung-yi, "TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, publisher: IEEE. [Online]. Available: https://dl.acm.org/doi/10.1109/TASLP.2021.3095662

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, and others, "Language models are unsupervised multitask learners."

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[27] Y.-A. Chung and J. Glass, "Generative Pre-Training for Speech with Autoregressive Predictive Coding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3497–3501, iSSN: 2379-190X.

[28] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-Task Self-Supervised Learning for Robust Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6989–6993, iSSN: 2379-190X.

[29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021, publisher: IEEE.

[30] T. Maekaku, X. Chang, Y. Fujita, L.-W. Chen, S. Watanabe, and A. Rudnicky, "Speech representation learning combining conformer cpc with deep cluster for the zerospeech challenge 2021," 2022.

[31] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.

[32] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," 2020.

[33] 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.

[34] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[35] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[36] "Textless NLP: Generating expressive speech from raw audio," Sep. 2021. [Online]. Available: https://ai.meta.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/

[37] G.-T. Lin, Y.-S. Chuang, H.-L. Chung, S.-w. Yang, H.-J. Chen, S. Dong, S.-W. Li, A. Mohamed, H.-y. Lee, and L.-s. Lee, "Dual: Discrete spoken unit adaptive learning for textless spoken question answering," *arXiv preprint arXiv:2203.04911*, 2022.

[38] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," 2024.