國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

基於自監督式語音表徵之無參考客觀語音品質評估

No-reference Objective Speech Quality Assessment Based on Self-supervised Speech Representations

曾韋誠

Wei-Cheng Tseng

指導教授: 李琳山 博士

Advisor: Lin-shan Lee Ph.D.

中華民國 111 年 6 月

June, 2022

doi:10.6342/NTU202300029

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

基於自監督式語音表徵之無參考客觀語音品質評估

No-reference Objective Speech Quality Assessment Based on Self-supervised Speech Representations

本論文係曾韋誠君（R09942094）在國立臺灣大學電信工程學研究所完成之碩士學位論文，於民國 111 年 6 月 28 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Communication Engineering on 28 June, 2022 have examined a Master's thesis entitled above presented by Wei-Cheng Tseng (R09942094) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李琳山	王小川	陳信宏
（指導教授 Advisor）		
鄭秋豫 鄭秋豫 (2022年6月28日 13:28 GMT+8)	簡仁宗	李宏毅 HungYi Lee (2022年6月28日 10:13 EDT)

系主任/所長 Director: 劉錦添

致謝

從加入語音實驗室以來，也已經過了兩年半。這兩年半的碩士生活裡，有限的時間被各種死線切割成一塊塊碎片，這些碎片與數不清的想法混搭，彼此間不停地映射、交錯：第一次參加 group meeting、第一次上台報告 paper、第一次試著做研究、第一次寫 paper、投到國際研討會、幸運地被接受、報告成果；第二次投稿、被拒絕；第三次投稿、被接受；寫碩士論文、整個大改、補實驗、校稿、修訂、口試、Interspeech。至此驀然回望，已然拼湊成一部蒙太奇電影。在這趟旅途上，我常覺得自己幸運得過分，到底何德何能可以得到這麼多人的幫助。這本碩士論文的完成，不只給了當初一無所知仍毅然投入語音處理研究的自己一個交代，也承載了對於所受到的這些幫助難以言盡的感謝。

感謝琳山老師當初願意收我進實驗室，讓我能擁有如此優良的環境來做研究，而在老師一次又一次的談話中，我逐漸了解何謂「做學問的能力」，也讓我深諳「見樹又見林」的重要性。除了經師，琳山老師更是位人師，在我對於前途感到迷惘時，幫我照亮前方的路，仔細地分析各種選擇的利與弊，並在日常生活的言行舉止中，樹立了讓我能夠一生遵循的典範。

感謝宏毅老師總是在研究上為我指點迷津。有好幾次看見老師行程表早已塞爆，但仍排除萬難、擠出三十分鐘與我討論研究進度，先全面地分析目前研究上的不足，再給出最精闢的建議，讓我在研究上少走了許多冤枉路。而老師樂天開朗的性格，也深深感染了我與周遭的同學，每一句「這聽起來很 promising 欸」，都讓我們有良好的心態來面對更多未知且艱難的研究問題。

doi:10.6342/NTU202300029

感謝健祐、伯翰、義聖、子賢、政豪、書文，很慶幸當初剛進入實驗室時坐在你們附近，在模仿烘焙王以及聽健祐跟記良吵架之間，讓我這個外校生逐漸融入語音實驗室這個大家庭，並總是不厭其煩的回答我在基礎知識與研究實作上的疑問，讓我能迎頭趕上其他同學們的進度，真正踏上做研究的旅途。

感謝瑋聰、凱為，早就已經數不清楚有幾個夜晚一起跑實驗、寫 paper，只為了能趕上幾天後的 conference deadline，可能是 Interspeech、也可能是 ICASSP，雖然當時總覺得熬夜讓人頭痛，下次不想再這樣了。但如今回想起來，何其有幸能有目標一致的夥伴不停激盪彼此的想法，一起努力，然後一起受苦。與你們做研究的時光著實是相當快樂的（喝酒也相當快樂）。

感謝實驗室的其他夥伴們，大家平時一起討論研究、聊實驗室八卦、走去 118 吃飯，一幕幕場景在我腦海中仍是鮮明的記憶。

感謝室友法鈞、宇呈、波哥，跟你們一同住在景平路的時光相當充實，這兩年下來打了好多場 LOL，法鈞的秀操作跟我哥的秀下限，讓平日間的寂寥得以受到排遣。而無數個清晨的 677、Q、到每個週末的 889，不僅沈澱了碩士班的壓力，也讓我的體重得到了相當多的積累。

感謝女友利真，從大學以來便陪伴在我身邊，在不同的人生階段中互相砥礪、一同成長，在每個難捱的時刻也都支持著對方。與你相處的每個週末，無論晴天或雨天、在台北或新竹，都因那些歡笑與感動成了記憶中特別的日子。與你去過的大小旅行，澎湖、宜蘭、花東、基隆、九份，都為我的碩士生活描繪了明亮的色彩。即使還不清楚未來會朝何方走去，但接下來的日子，還請你多多指教！

最後感謝我的家人們：阿公、阿嬤、爸爸、媽媽，由於求學過程間你們無條件的支持與鼓勵，我才能無所顧忌的在北部唸書，盡情地探索自己的興趣，並得以將我的熱忱充分發揮在喜愛的事物上。是你們的支持與鼓勵，才造就了今天的我，親情恩重如山，再多感謝都難以回報。

doi:10.6342/NTU202300029

摘要

語音品質評估（Speech Quality Assessment）多年來，一直是語音處理（Speech Processing）領域的重要課題。傳統上，經許多人聆聽後所獲得的平均主觀意見分數（Mean Opinion Score）一直是語音品質評估的金科玉律，但由於需舉辦聆聽測驗來獲取許多受測者對於待測語音訊號的主觀評分，因而必須耗費大量的人力與時間。另一方面，確有多項基於模擬人類聽覺系統所發展而來的全參考客觀語音品質評估方法（Full-reference Objective Speech Quality Assessment）被普遍使用，並證實與平均主觀意見分數成高度相關。然而，由於這些方法中需要乾淨真實的參考訊號作為待測訊號的比較對象，使得它們在無法取得參考訊號的情況下無法使用。因此，開發一套無參考客觀語音品質評估方法（No-reference Objective Speech Quality Assessment），也就是不須參考語音訊號，且與平均主觀意見分數的評量結果呈現高度相關的語音品質評量技術，乃成為本研究的主題。

另一方面，近年自監督式學習（Self-supervised Learning）的預訓練（Pre-trained）技術在語音處理領域上已經相當成熟，可以由大規模無標記語料庫中，提取出隱含豐富資訊的特徵向量（Feature Vector）。這些特徵向量被證實能增進多項語音處理任務的表現，如語音辨識、語者辨識、語音翻譯等；只是在無參考客觀語音品質評估上的潛力還未被充分發掘。在本論文中，我們首先分析了自監督式語音表徵用於無參考客觀語音品質評估上的可行性，在實驗中發現，自監督式語音表徵中含有豐富的聲學（Acoustic）資訊及語言（Linguistic）內容的資訊，且能區隔不同品質的語音訊號，說明其可能相當適合用於無參考語音品質評估。接

doi:10.6342/NTU202300029

著，我們基於上述結果，提出了一套全新的、基於 HuBERT 表徵的深層（Deep）無參考語音品質評估技術。實驗結果顯示，這套技術全面超越過去使用傳統語音表徵的所有方法，並在不同語言上有更好的泛化能力。最後，我們也使用探測分析（Probing Analysis）更深入理解影響模型表現的因素。

關鍵字：深度學習、自監督式語音表徵、語音處理、語音品質評估、無參考客觀語音品質評估

doi:10.6342/NTU202300029

Abstract

Speech quality assessment is to evaluate the quality of audio, and it has been an essential part of speech processing to measure the performance of a system for decades. Conventionally, the mean opinion score (MOS) has been considered the ”golden standard” for speech quality assessment, but such measurement involves a large number of human listeners, making it costly and time-consuming. Full-reference objective speech quality assessment approaches have thus been developed to simulate the auditory system of human beings and have been shown to have a high correlation with MOS. However, these approaches require a clean reference signal for comparison with the test signal, limiting their utility when such a signal is unavailable. Accordingly, there is a need to develop a no-reference objective speech quality assessment method that correlates well with human perception and does not require a reference signal, which is the main focus of this thesis.

On the other hand, self-supervised pre-trained models that enhance the utility of large-scale unlabeled speech datasets have emerged in the research field of speech processing. The self-supervised models can extract high-level, informative, and compact

doi:10.6342/NTU202300029

representation vectors from the raw audio inputs. The extracted representations have been demonstrated beneficial for downstream tasks like speech recognition, speaker verification, speech translation, and spoken language understanding. Nonetheless, the capability of these self-supervised speech representations for speech quality assessment has yet to be well addressed. In this thesis, we first conduct a preliminary analysis to investigate the feasibility of adopting self-supervised speech representations for speech quality assessment. The analysis results demonstrated that these representations contain rich acoustic and linguistic information and can distinguish audio signals of different qualities, suggesting their potential for evaluating speech quality. Accordingly, we proposed a novel, deep no-reference objective speech quality assessment model based on the HuBERT feature. The experiment results showed that our model significantly outperforms the previous state-of-the-art approaches and has better generalization ability across different languages. Moreover, we also conducted several probing analyses to further understand the factors that affect the model performance.

Keywords: Deep Learning, Speech Processing, Self-supervised Speech Representation, Speech Quality Assessment, No-reference Objective Speech Quality Assessment

doi:10.6342/NTU202300029

目錄

doi:10.6342/NTU202300029

doi:10.6342/NTU202300029

doi:10.6342/NTU202300029

doi:10.6342/NTU202300029

圖目錄

doi:10.6342/NTU202300029

doi:10.6342/NTU202300029

表目錄

doi:10.6342/NTU202300029

第一章　導論

1.1　研究動機

「語音」一直是人類個體之間互相溝通的重要方式。日常生活中，不論是買東西、學習、抑或是進行科學研究，都免不了與人交談。此外，相比於文字，語音傳達的資訊更為豐富，短短幾秒的聲音內便包含了說話的人想表達的內容、語氣以及情緒。而語音處理（Speech Processing）領域的研究，則試圖開發與語音相關的應用，以提升人們生活的便利性，例如蘋果公司所推出的 Siri、Google 的語音搜尋、會議記錄使用的語音逐字稿軟體、大眾運輸中以文句翻語音（Text-to-speech）系統合成出的廣播內容、甚至到電話中聲音的傳遞，都屬於語音處理領域的研究成果，足見其影響層面之廣。

在眾多的語音處理研究領域中，為了讓人們能聽到品質更好的語音，語音品質評估（Speech Quality Assessment）——顧名思義指的是衡量語音品質好壞的方法——一直是重要的研究主題之一。其中，平均主觀意見分數（Mean Opinion Score）為最具代表性的語音品質評估方法。平均主觀意見分數主要透過聆聽測驗（Listening Test）來評估語音訊號的品質，一群受試者在聆聽待測的語音訊號後，根據語音品質好壞給出特定範圍內的評分，而所有受試者的分數平均即為該語音訊號的平均主觀意見分數。由於平均主觀意見分數幾乎適用於各種場景中，且評估結果也於人們的普遍喜好直接相關，傳統上經常將它做為語音品質評估的金科玉律。然而，這個方法在實際應用層面上有相當大的缺點：收集多筆待測語音的

doi:10.6342/NTU202300029

評分結果的過程不可避免地需要為了舉辦聆聽測驗（Listening Test）耗費大量的時間與精力，這大大降低了平均主觀意見分數的便利性，並且也嚴重地減緩實驗流程推進的速度。

另一方面，確有多項基於模擬人類聽覺系統所發展而來的全參考客觀語音品質評估方法（Full-reference Objective Speech Quality Assessment）被普遍使用，這些方法根據不同使用情境設計，透過特定演算法分析乾淨真實的參考訊號（Reference Signal）以及失真的待測訊號之間的差異，並將這些差異映射至某種指標上，不僅能有效率地衡量待測訊號的品質，其評估結果也被證實與平均主觀意見分數呈現高度相關。常見的全參考客觀語音品質評估方法包含：語音品質感知評估（Perceptual Evaluation of Speech Quality，記為 PESQ）[1]、短時客觀可理解性（Short-Time Objective Intelligibility，記為 STOI）[2]、以及梅爾倒頻譜距離（Mel-cepstral Distance）[3] 等。

然而，很多時候，事實上無法取得參考訊號—例如在文句翻語音與語音轉換（Voice Conversion）任務時，想生成一段目標語者沒有說過的內容；或在通訊時，接收端只收到失真的訊號—因而無法使用上述的全參考客觀語音品質評估方法。為了在這些情況下也能夠準確且有效率的衡量語音品質的好壞，近年來有多個提出基於深層學習（Deep Learning）的無參考客觀語音品質評估方法（No-reference Objective Speech Quality Assessment）被提出 [4–8]，試圖讓模型從「資料」中學習人類為語音品質進行評分時隱含的行為準則。這些方法僅需將待測訊號的梅爾時頻譜（Mel-Spectrum）、深層時頻譜（Deep Spectrum）等語音表徵（Speech Representation）輸入至深層類神經網路模型，即可一定程度地模擬待測訊號的主觀評估結果（即其平均主觀意見分數）。不僅讓人們在語音品質評估方法上看見了新的可能性，也使得深層無參考客觀語音品質評估模型成為一項重要的研究主題。

另一方面，近年來，以自監督式學習（Self-supervised Learning）預訓練

doi:10.6342/NTU202300029

（Pre-trained）的模型在語音處理（Speech Processing）領域也已經取得了大量的關注 [9–13]。這些模型可以從大量無標記語料中，習得語音資料內隱含的、與內容、語者、情緒相關的複雜信息。而利用預訓練好的自監督式語音模型，我們可以從原始的語音訊號中，抽取含有豐富信息的表徵向量，這些表徵被證實能增進多個語音處理任務的表現，並超越單純使用監督式學習（Supervised Learning）的方法，這些語音處理任務包含：語音辨識（Speech Recognition）、語者辨識（Speaker Recognition）、口語理解（Spoken Language Understanding）及語音翻譯（Speech Translation）等。

雖然自監督式學習之語音表徵在多個語音處理任務上被證實有良好的表現，但其在無參考客觀語音品質評估上的潛力尚未被完整探討。我們猜測，在其他任務上的好表現應該能夠複製到無參考客觀語音品質評估任務上。對此，本論文首先利用多個分析實驗探討自監督式語音表徵在無參考客觀語音品質評估任務上的可行性，並提出一套簡單但有效的深層類神經網路模型架構將自監督式語音表徵用於無參考客觀語音品質評估。另外，本論文也設計了泛化能力（Generalizability）實驗和探測分析（Probing Analysis）實驗進一步理解本論文提出之模型在不同類型的資料集上的泛化能力以及影響模型表現的因素。

1.2 研究方向

本論文的主要研究方向為，將自監督式語音表徵用於無參考客觀語音品質評估，來進一步提升深層無參考客觀語音品質評估模型的表現。研究方向主要包含：

- 設計多個實驗理解自監督式語音表徵中所含有的資訊種類，並透過降維分析（Dimension Reduction Analysis）觀察自監督式學習預訓練的語音模型能否在潛藏空間（Latent space）上區分不同品質的語音，以探討其在無參考客觀語音品質評估的可行性。

doi:10.6342/NTU202300029

- 基於自監督式語音表徵，設計一套模型架構以將其用於無參考客觀語音品質評估。並比較其與其他基於傳統語音表徵的無參考客觀語音品質評估模型的表現。
- 探討本論文提出之模型在不同資料集之間的泛化能力。
- 利用探測分析實驗進一步理解影響模型表現的因素。

1.3 研究貢獻

本文的研究貢獻如下：

- 可行性分析實驗中發現自監督式語音表徵中同時含有「低層次（聲音強度、音高）」以及「高層次（韻律、語意）」的資訊，而透過繪製二維投影散佈圖（2-D Scatter Plot）以及標準相關分析（Canonical Correlation Analysis）則發現不同品質語音的自監督式表徵在空間上明顯呈現不同分佈，說明自監督式語音表徵在無參考客觀語音品質評估上的潛力。
- 根據可行性分析的結果，本論文基於 HuBERT [13] 表徵，利用前饋式類神經網路（Feed-forward Neural Network）、遞迴式類神經網路（Recurrent Neural Network）及專注合計層（Attentive Pooling）[14] 提出一套簡單但有效的模型架構來執行無參考客觀語音品質評估，並在實驗中證實其表現超越過去使用傳統語音表徵的方法。
- 探討本論文提出之模型在不同資料集之間的泛化能力，實驗顯示本論文提出之模型在不同語言上的泛化能力相較過去的方法更強。
- 探測分析實驗發現模型對於高品質生成語音的預測準確度、相較於其他品質的語音來得差，並藉此反思未來可能的研究方向。

doi:10.6342/NTU202300029

1.4 章節安排

本論文的章節安排如下：

- 第二章：介紹本論文相關的背景知識。
- 第三章：介紹基於深層學習的無參考客觀語音品質評估模型及相關技術。
- 第四章：探討自監督式語音表徵用於無參考客觀語音品質評估的可行性。
- 第五章：提出基於 HuBERT 表徵的無參考客觀語音品質評估模型，並衡量其表現。
- 第六章：本論文的結論與未來方向。

doi:10.6342/NTU202300029

第二章　背景知識

2.1　深層類神經網路

類神經網路（Neural Network）為一種數學模型，通常用來近似或估計現實問題背後所隱含的複雜函式，其構想源自於生物的神經系統，期望藉由模仿人腦的訊息處理方式，使電腦能夠解決現實生活中的應用問題，也就是所謂的「人工智慧（Artificial Intelligence）」。類神經網路的基本單位為神經元（Neuron），最早由麥氏（W. McCulloch）與皮氏（W. Pitts）於 1943 年所提出 [15]，其基本構造如圖 2.1所示。每個神經元可接收多個輸入訊號，經內部的數學運算後可以得到相對應的輸出訊號。完整的數學表示如下：

$$y = \sigma(W^T \cdot X + b) \tag{2.1}$$

其中 X 和 y 分別為神經元的輸入和輸出、W 為神經元的權重參數（Weight）、b 為偏移量（Bias）、而 σ 為激活函數（Activation Function）。

羅氏（F. Rosenblatt）基於上述的神經元，於 1958 年提出感知器（Perceptron）[16]，被視作最早的類神經網路之一。感知機為具有兩層全連接構造的網路，每層含有若干個神經元。羅氏描述了感知機的數學定義與學習演算法，並展示其在圖形識別（Pattern Recognition）上的潛力，在當時造成一大轟動，紐約時報甚至給出了「一台將可以走路、說話、看、複製自己與檢視自己存在的電腦的誕生」

doi:10.6342/NTU202300029

這樣的標題，自此開啟了類神經網路的研究熱潮。

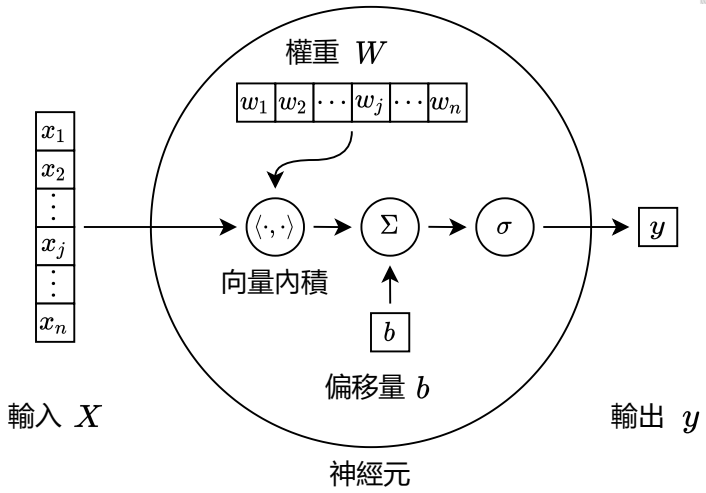


Figure 2.1: 神經元構造示意圖

2.1.1 前饋式類神經網路

雖然感知器在當時引起類神經網路的研究熱潮，但它事實上有著非常致命的缺陷。1969 年，明氏（M. Minsky）與帕氏（S. Parpet）在他們的著作中指出感知器的兩項關鍵問題，一是感知器無法處理非線性可分的問題（如預測 XOR 運算子的輸出）、二是當時的電腦運算能力無法負荷大型感知器所需的龐大運算量，類神經網路的研究也因此進展緩慢。一直到魯氏（D. E. Rumelhart）與辛氏（G. E. Hinton）將反向傳播演算法（Backpropagation）[17] 應用在類神經網路的訓練上 [18, 19]，使得具非線性激活函數的多層網路架構成為可能，第一個問題才得以解決，也確立現今類神經網路的基本訓練方法。並且隨著電腦技術的發展，電腦硬體逐漸可以負荷更深層的網路架構的訓練，而清氏（K. Kawaguchi）在 2000 年提出的多層感知器（Multi-layer Perceptron）[20]，則大致底定了現今深層前饋式類神經網路（Feed-forward Neural Network）的完整架構。

現今常見的深層前饋式類神經網路的模型架構如圖 2.2所示，由多個全連接層所構成，其中包含輸入層、大於一層的隱藏層、及輸出層。模型由輸入層接收訊

doi:10.6342/NTU202300029

號後，訊號經神經元進行計算，得到的輸出則作為下一層網路的輸入，透過此機制的層層傳遞，最後得到相對應的輸出。

一般來說，深層類神經網路主要透過先前提到的反向傳播演算法 [18, 19] 進行訓練。給定一個任務及對應的訓練資料，我們首先將訓練資料輸入至模型，經由層層傳遞後得到輸出，接著某目標函數（Objective Function）會根據模型的輸出與該訓練資料的正確標記計算損失（Loss）$\mathcal{L}$，並透過梯度下降法（gradient descent）更新模型參數：

$$\theta_t \leftarrow \theta_{t-1} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{t-1}} \tag{2.2}$$

其中 θ_t 為模型於時間點 t 的參數、η 為學習率（Learning Rate），用來控制每一次更新的步伐大小。此步驟通常會重複多次，使模型收斂至局部或全域最佳值。此外，在模型的訓練過程中，通常會根據任務類型來選擇目標函數的種類。以分類（Classification）任務來說，普遍使用交叉熵（Cross Entropy）作為目標函數：

$$\mathcal{L}_{CE}(\mathbf{y}, f(x;\theta)) = \sum y_i \log(f(x;\theta)_i) \tag{2.3}$$

其中 $f(x;\theta)$ 為模型在參數為 θ 對於輸入 x 的輸出；而在迴歸（Regression）任務中，則經常使用平均絕對誤差（Mean Absolute Error）及均方誤差（Mean Squared Error）：

$$\mathcal{L}_{MAE}(y, f(x;\theta)) = \sum |y - f(x;\theta)| \tag{2.4}$$

$$\mathcal{L}_{MSE}(y, f(x;\theta)) = \sum (y - f(x;\theta))^2 \tag{2.5}$$

時至今日，最單純的前饋式類神經網路已經較少被單獨使用，但其概念及訓練方法仍深深影響後續各類神經網路的變形。

doi:10.6342/NTU202300029

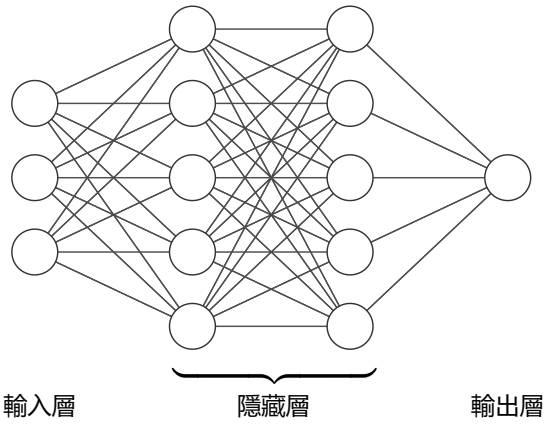


Figure 2.2: 深層前饋式類神經網路示意圖，每個圓圈代表一個神經元。

2.1.2 卷積式類神經網路

卷積式類神經網路（Convolutional Neural Network）為楊氏（Yann LeCun）於 1990 年所提出 [21]，其中，卷積指的是影像處理領域中常用的矩陣卷積運算。卷積式類神經網路主要由卷積層（Convolution Layer）、合計層（Pooling Layer）、及全連接層（Fully-connected Layer）所構成。不同於前饋式類神經網路的全連接性質，在卷積式類神經網路中，卷積層與池化層的神經元只會接收前一層輸出的特定範圍內的訊號，這個範圍被稱為感受域（Receptive Field）。此外，卷積層與合計層通常會串連使用，分別模仿人類視覺皮層（Visual Cortex）中的簡單細胞（Simple Cell）與複雜細胞（Complex Cell）的功能。

我們以圖 2.3來說明卷積式類神經網路的實際運作方式，給定一輸入訊號，卷積層首先使用不同的卷積矩陣（或稱為核心、Kernel）將其轉換為若干個特徵圖（Feature Map），這些特徵圖的值代表著特定區域內的邊界或強度資訊。接著，合計層會以特定模式抽取特徵圖中的數值作為輸出，此步驟是為了降低網路對於細微輸入變化的敏感度，並簡化向量維度。最後，在經過多層的卷積與合計運算後，全連接層會檢視最後一層的所有輸出，以得到整個網路的輸出。

卷積式類神經網路具有局部平移不變性（Shift-invariant property），可用來抽

doi:10.6342/NTU202300029

取訊號中不受空間或時間影響而改變的某些特徵，也因此被廣泛使用在影像辨識 [22–24] 及音素辨識（Phoneme Recognition）任務 [25] 上。

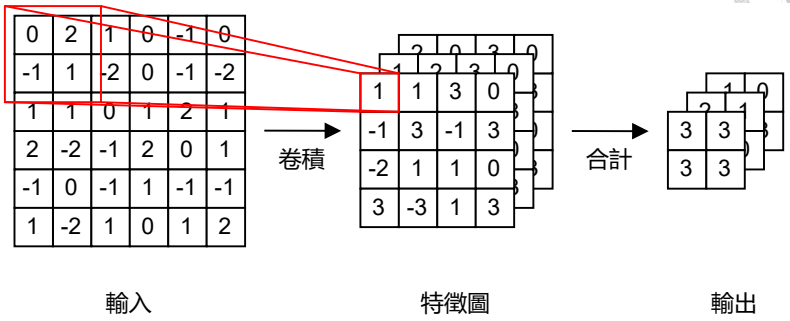


Figure 2.3: 卷積層與合計層運作示意圖

2.1.3 遞迴式類神經網路

遞迴式類神經網路 (Recurrent Neural Network) 為喬式（M. I. Jordan）於 1986 年所提出 [26]，其特點在於，網路中的神經元會與自身連接形成循環，在每一次運算過程中能夠參考先前的輸出以進行推論，這使得網路天生具有「記憶」的能力。也因此，遞迴式類神經網路相當適合用來處理具先後順序的時間序列資料（Time-series Data）。

遞迴式類神經網路的基本架構如圖 2.4所示。給定一筆具有 T 個時間點的序列 $\mathbf{x} = [x_1, x_2, \cdots, x_T]$，將序列輸入至遞迴式類神經網路後，模型會依照先後順序推論各個時間點對應的輸出。對於在時間點 t 的輸入 x_t，其對應的輸出 y_t 為：

$$y_t = \text{RNN}(x_t, h_{t-1}; \theta) \tag{2.6}$$

其中，h_t 為隱藏層於時間點 t 的輸出、θ 為模型的參數。

由於其模型特性，遞迴式類神經網路自提出以來便被廣泛運用在序列至序列（Sequence-to-sequence，輸入與輸出皆為序列）任務上，並衍伸出多種變形，以解決序列至序列任務中常見的長序列問題，例如長短期記憶體（Long Short-term Memroy）類神經網路單元及閥遞迴式類神經網路單元（Gated Recurrent Unit）

doi:10.6342/NTU202300029

等，皆為著名代表之一。而蘇氏（I. Sutskever）基於長短期記憶體與自編碼器（Autoencoder），於 2014 年提出的模型架構，則被視為最經典的序列至序列模型 [27]。其架構如圖 2.5所示，其中，編碼器與解碼器皆為長短期記憶體類神經網路單元構成的遞迴式類神經網路。給定一序列資料，編碼器（Encoder）首先將資料轉換為一固定長度的語境向量（Context Vector）。接著，在接收該語境向量作為初始的隱藏狀態後，解碼器（Decoder）會根據前一個時間點的隱藏狀態及輸出，依序推測各個時間點的輸出。

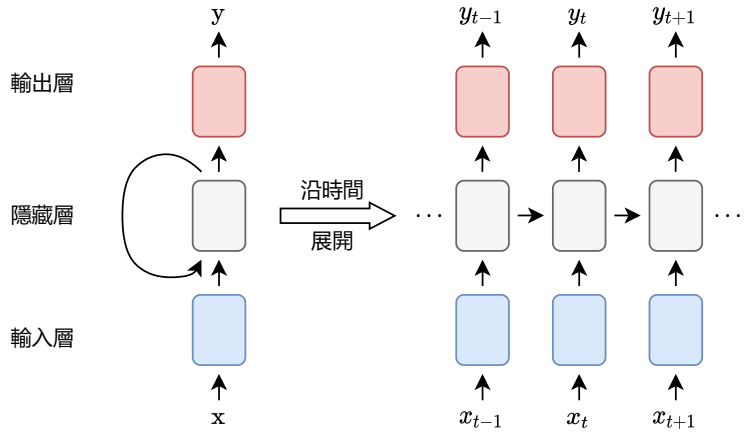


Figure 2.4: 遞迴式類神經網路架構示意圖

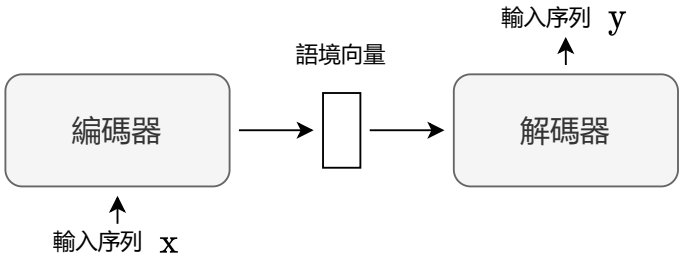


Figure 2.5: 蘇氏（I. Sutskever）所提出的序列至序列模型架構示意圖。

2.1.4 專注機制

上述的架構在多個序列至序列任務表現上取得了重大的突破，然而，幾個明顯的問題仍然存在：一、資料中的序列長度往往不盡相同，將所有序列皆投射到固定長度的向量顯然違背直覺，尤其在長序列的狀況下，其資訊含量會被限制

doi:10.6342/NTU202300029

住。二、解碼器在推論每一個時間點的輸出時只能參考前一時間點的隱藏層狀態（Hidden State），使得越早輸入的內容越容易失去影響力，也因此無法考慮全局的資訊。

為了解決上述的問題，巴氏（D. Bahdanau）提出了一個構想：「那如果讓所有的時間點都擁有自己的語境向量呢？」，也因此，專注機制（Attention Mechanism）便誕生了 [28]。其作法如圖 2.6所示，與一般的遞迴式類神經網路的差異在於，解碼器在解碼時會從編碼器的輸出「搜尋」特定資訊，而非僅依靠編碼器輸出的語境向量。

詳細說明如下，給定一個有 T 個時間點的序列 $\mathbf{x} = [x_1, x_2, \cdots, x_T]$，編碼器首先將序列轉換為一連串的隱藏狀態，其中，在時間點 t 的編碼器輸出為：

$$h_t = \text{Enc}(x_t, h_{t-1}) \tag{2.7}$$

而在解碼階段時，不同於前述的序列至序列模型，這裡不單使用前一個時間點的隱藏狀態作為解碼器的輸入，也同時在各個時間點計算其對應的語境向量：

$$c_t = \sum_i a_{t,i} h_i \tag{2.8}$$

其中，權重 $a_{i,t}$ 是透過另一個對齊模組（Alignment Module）$\mathcal{A}$ 計算得出：

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_i \exp(e_{t,j})} \tag{2.9}$$

$$e_{t,i} = \mathcal{A}(s_{t-1}, h_i) \tag{2.10}$$

可以從上述看出語境向量其實是各個時間點編碼器輸出的加權平均，也就是說，解碼器（這裡的解碼器包含前述的對齊模組）會自己學習如何針對前後文的關係，找出哪個時間點的隱藏狀態對預測輸出是有幫助的；這也使得編碼器的負擔變小，不再需要將整個序列的內容與資訊轉換成固定長度的語境向量。總結上述，對於

doi:10.6342/NTU202300029

在時間點 t 的解碼器輸出變成：

$$y_t = \text{Dec}(y_{t-1}, s_{t-1}, c_t) \tag{2.11}$$

其中，s_t 為解碼器在時間 t 的隱藏狀態。雖然專注機制與傳統的序列至序列模型相比，需要另外計算各個時間點的語境向量，而造成計算量上升，但這樣的機制被證實對於機器翻譯任務有非常大的幫助，並隨後也在多個序列至序列任務超越前述的方法，其中包含語音辨識、語言模型等等。

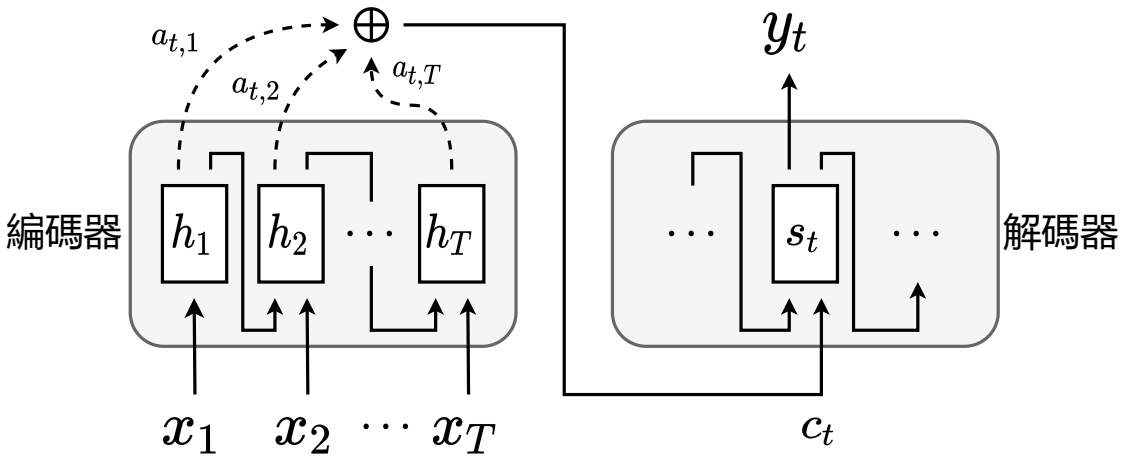


Figure 2.6: 專注機制流程圖

轉換器式類神經網路

由於專注機制在多種序列至序列任務上取得了莫大的成功，瓦氏（A. Vaswani）在 2017 年提出一個完全基於專注機制的網路架構——轉換器類神經網路（Transformer）[29]。轉換器類神經網路的架構如圖 2.7所示，其中包含兩個模組：轉換器編碼器（Transformer Encoder）及轉換器解碼器（Transformer Decoder）。轉換器編碼器能將輸入序列轉換為一連串的隱藏表徵（Hidden Representation），其角色相當於我們在上一小節所提到的語境向量。接著，轉換器解碼器會根據這些隱藏表徵，依序計算每一個時間點的輸出，其中，轉換器解碼器在各個時間點的輸入為前一時間點的輸出序列。

doi:10.6342/NTU202300029

轉換器編碼器為多層架構，每一層由一個多頭自專注（Multi-head Self-attention Mechanism，見圖 2.8）模組及前饋式類神經網路所構成。對於具 T 個時間點的輸入序列 $\mathbf{x} = [x_1, x_2, \cdots, x_T]$，多頭自專注模組會先將輸入序列經線性轉換為三個矩陣，依序為詢（Query）$Q \in \mathrm{R}^{T \times d_k}$、鑰（Key）$K \in \mathrm{R}^{T \times d_k}$ 以及值（Value）$V \in \mathrm{R}^{T \times d_v}$，接著經由縮放點乘積專注（Scaled-dot Product Attention）模組計算在某專注頭（Attention Head）的輸出：

$$\text{Attention}(Q', K', V') = \text{softmax}(\frac{Q'K'^T}{\sqrt{d_k}})V' \tag{2.12}$$

$$\tag{2.13}$$

其中 $\{Q', K', V'\} = \{QW^Q, KW^K, VW^V\}$，而 $\text{softmax}(\cdot)$ 為軟性最大化函數，定義為：

$$\text{softmax}(a_i) = \frac{\exp(a_i)}{\sum_j \exp(a_j)} \tag{2.14}$$

最後，多頭自專注模組的輸出可透過串接多個專注頭輸出，再通過一個線性轉換得出：

$$\text{MultiHeadSelf}(Q, K, V) = \text{Concat}(\text{head}_i)W^O \tag{2.15}$$

$$\tag{2.16}$$

其中，在第 i 個專注頭的輸出為：

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.17}$$

而 $\text{Concat}(\cdot)$ 為向量串接，可以將不同專注頭的輸出合併為單一個矩陣；另一方面，轉換器解碼器同樣具有多層架構，差別在於每一解碼器層除多頭自專注機制外，還包含一多頭跨專注機制（Multi-head Cross-attention Mechanism，見圖 2.8）

doi:10.6342/NTU202300029

模組。多頭跨專注機制與巴氏的專注機制概念相同，唯一不同在於鑰與值是由轉換器編碼器的輸出 $z = [z_1, z_2, \cdots, z_T]$ 經線性轉換得出。

與巴氏的模型架構相比，單純使用專注機制與前饋式類神經網路的轉換器類神經網路在實驗中被證實能夠更好地考慮輸入序列與輸出序列之間的全局資訊，使得在序列至序列任務上的表現進一步地提昇；單純使用前饋式類神經網路也使得模型在計算上更便於平行化，進而讓使用更大的訓練資料集成為可能。由於具備多項優點，轉換器類神經網路自提出以來便迅速替代了基於專注機制的遞迴式類神經網路架構，如今，轉換器類神經網路已然成為處理序列至序列任務的標準作法。此外，多個基於轉換器類神經網路的預訓練模型也被提出，其中包含：BERT [30]、GPT 系列 [31, 32]、Wav2vec 2.0 [12] 等等，這些模型被預訓練在大量無標記的資料集上，而訓練好的模型在經過微調（Fine-tune）後，被證實能增進下游任務（Downstream Task）的表現。

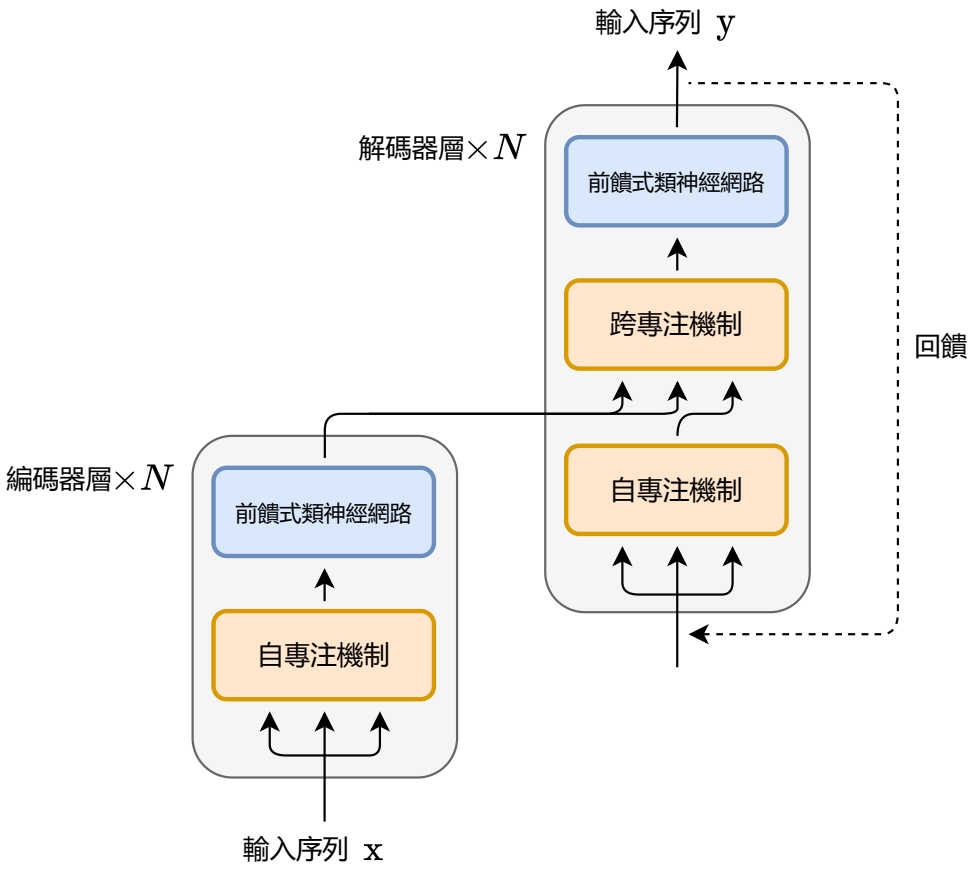


Figure 2.7: 轉換器式類神經網路架構圖

doi:10.6342/NTU202300029

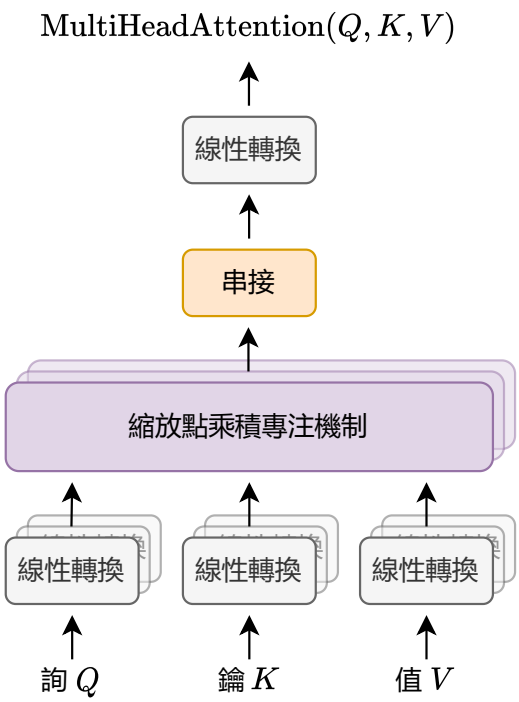


Figure 2.8: 多頭專注機制示意圖，根據鑰與值的來源可進一步分為多頭自專注與多頭跨專注。

專注合計

除了運用在序列至序列模型以外，專注機制也被運用來抽取一連串隱藏表徵的全局資訊，並降低其維度以方便計算。薩氏（P. Safari）首先在其著作提出專注合計（Attention Pooling）[14]，透過計算一連串音框層級（frame-level）表徵的加權平均，將這些表徵合併至語句層級（Utterance-level），以抽取輸入語音的語者資訊。其詳細流程如下所述。給定一具 T 個時間點的表徵序列 $H = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T] \in \mathrm{R}^{T \times d}$，其合併後的表徵向量為：

$$\bar{\mathbf{h}} = \mathrm{Softmax}(W_c H^T) H \tag{2.18}$$

其中 $W_c \in \mathrm{R}^d$ 為可訓練的矩陣。可以看出，上述式子與前述的縮放點乘積專注機制相當類似，差別在於，在專注合計中，僅有詢（在這裡為 W_c）是可以被訓練的。雖然專注合計原先是設計來抽取語音表徵中的語者資訊，但在本論文的實驗中，我們證明專注合計也能用來整合不同時間資訊對於語音品質的影響程度。

doi:10.6342/NTU202300029

2.2 語音表徵

在數位信號處理領域中，我們一般將信號以波形的形式儲存。然而，波形含有的信息太過豐富，單從訊號在每個時間點的能量強弱無法輕易地抽取語音訊號中含有的音素（Phoneme）、音高（Pitch）、及韻律（Prosody）等信息。也因此，在分析時通常會利用短時距傅立葉變換（Short-time Fourier Transform）將波形轉換為時頻譜（Spectrogram），讓訊號有更好的可解釋性。此外，前人更分析人類聽覺感知能力與聲音頻率成份之間的關係，發現人耳對於低頻聲音的變化相對於高頻更加敏感，並根據這項結果提出梅爾刻度（Mel-scale），定義為人耳對於等距音高變化的感知程度。而將時頻譜的頻率轉換為梅爾刻度，即為梅爾時頻譜（Mel Spectrogram），相較於時頻譜，更貼近各頻率區段對於人耳來說的重要程度，至今仍是相當常用的語音表徵之一。

另外，梅爾倒頻譜係數（Mel-frequency Cepstrum Coefficient）也是常見的語音表徵之一，梅爾倒頻譜係數可以透過對梅爾時頻譜做離散餘弦變換（Discrete Cosine Transform），並抽取其低維度的係數得到。與梅爾時頻譜相比，梅爾倒頻譜係數的值之間的相關性較低，在降低維度的同時又能精確地保留語音訊息中的重要資訊，也因此更適合用於早期語音辨識經常使用的高斯混合模型（Gaussian Mixture Model），是為語音表徵工程中的集大成之作。然而，近年來，隨著深層學習的崛起，人們發現相對於高度精緻化的梅爾倒頻譜係數，使用含有更多信息的梅爾時頻譜在多個任務上有較好的表現。

2.2.1 監督式學習之語音表徵

除了上述以聲學知識與表徵工程建構出的語音表徵外，我們也可以先利用監督式學習將模型訓練於特定任務上，而訓練好的模型的輸出則可被作為一種語音表徵。舉例來說，音素後驗機率（Phonetic Posteriorgram）即是透過訓練好的音素

doi:10.6342/NTU202300029

辨識模型，抽取出輸入語音各時間點的音素（Phoneme）的機率分佈；而深層時頻譜（Deep Spectrum）則是將梅爾時頻譜作為輸入，利用訓練好的深層影像辨識模型抽取特定層的隱藏狀態。這些表徵首先被發現能使用在語音轉換 [33, 34]、語音分類 [35] 等語音處理任務上，而近幾年，這些表徵也被證實能夠增進深層無參考客觀語音品質評估模型（請見 2.5.2 小節以及第 3 章）的表現 [36]。

2.2.2 自監督式學習之語音表徵

近年來，由於可利用的無標記語料的大量出現，自監督式預訓練模型（Self-supervised Pre-trained Model）在語音處理領域上獲得了大量的關注。自監督式學習主要透過給定特定部分的資料，讓模型預測資料的另一部分 [37]，並藉此學習該種類資料隱含的內部結構。相較於監督式學習，這種方法大大減輕了人工標記資料的負擔。而透過自監督式學習預訓練的語音模型（以下簡稱自監督式語音模型），可以從原始的語音訊號中，抽取隱含豐富信息的表徵向量，這些表徵被證實能增進多個語音處理任務的表現，其中包含：語音辨識、語者辨識及語音翻譯等 [38]。

常見的自監督式語音模型主要可以透過使用的目標函數分為兩大類，其中包含：遮罩聲學模型 (Masked Acoustic Modeling) 及對比式學習 (Constrastive Learning)。以下將分別為這兩個種類進行詳細的介紹：

遮罩聲學模型

遮罩聲學模型是受到自然語言處理中的遮罩語言模型（Masked Language Modeling）所啟發，其概念如圖 2.9所示，藉由隨機屏蔽訊號中的某一部份，讓模型學習如何從其他未受屏蔽的部分「預測」未知的資訊。舉例來說，Mockingjay [9] 以梅爾時頻譜作為輸入，並屏蔽特定時間比例的音框，迫使模型還原原本的梅爾時頻譜。而 TERA [11] 則以此為基礎，將輸入改為 fMLLR（Feature Space Maximum Likelihood Linear Regression），並額外加入頻率維度上的屏蔽，進一步提升模型的

doi:10.6342/NTU202300029

表現；另一方面，HuBERT [13] 將遮罩運用在潛在空間上（Latent Space），並以輸入訊號的 MFCC 分群（Cluster）作為目標，讓模型學習利用其他時間點的表徵，預測受屏蔽時間點應屬於哪個分群。

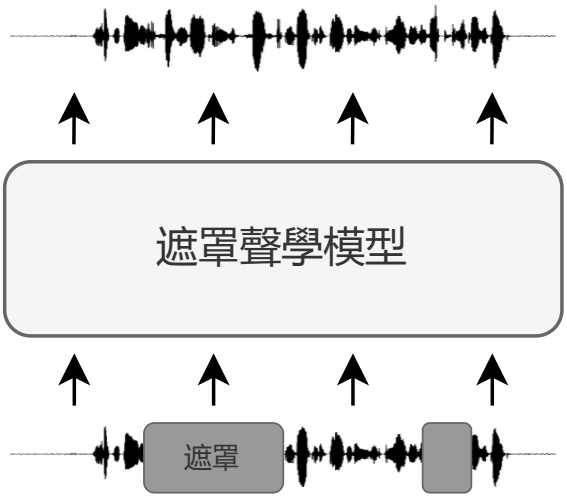


Figure 2.9: 遮罩聲學模型示意圖

對比式學習

雖然基於遮罩聲學模型的自監督式語音模型在多個下游任務上有不錯的表現，然而，由於語音訊號的連續性，讓模型在預測受屏蔽部分時有相當大的不確定性（Uncertainty）[37]，傾向於預測所有看似合理的輸出的平均，進而限制模型的能力。對此，有另一種方法則透過比較不同類別資料之間的相同之處與相異之處來學習，被稱為對比式學習（Contrastive Learning），其基本概念如圖 2.10所示。以影像處理領域的對比式學習為例，在訓練過程中，給定的輸入影像會經由不同的資料增強方法（Data Augmentation）進行變換，這些變換後的圖片被稱作正採樣（Positive Sample），而資料集中的其他影像及對應的變換結果則被稱為負採樣（Negative Sample），模型必須學習在輸出的潛藏空間上拉近正採樣之間的距離，並同時加大與負採樣之間的距離。相似的學習方法也可以用在預訓練自監督語音模型上，例如，最早使用對比式學習的對比預測編碼（Contrastive Predictive Coding，記為 CPC）[10] 即透過在前一時間點的輸出表徵預測編碼器未來的輸出，讓模型理解表徵中的哪些信息是有幫助的。另一方面，wav2vec 2.0 [12] 則在

doi:10.6342/NTU202300029

潛藏空間上屏蔽部分的時間點，讓模型學習利用其他時間點的信息，從一群採樣中正確選出受屏蔽時間對應的表徵。

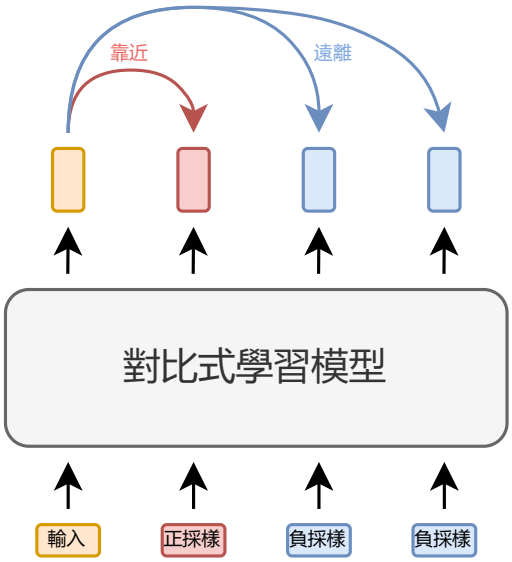


Figure 2.10: 對比式學習模型示意圖

如同前面所述，自監督式語音模型所抽取的語音表徵已被證實能增進多個語音處理任務的表現，而在本論文中，透過實驗結果，我們發現自監督式學習之語音表徵也可以用在無參考客觀語音品質評估任務上。

2.3 語音品質評估

語音品質評估（Speech Quality Assessment），顧名思義指的是衡量語音品質好壞的方法，在語音處理領域中一直是重要的研究主題之一。為了讓人們在生活中聽到品質更好的語音，我們利用不同的語音品質評估方法來衡量所開發的語音生成（Speech Synthesis）系統、語音增強（Speech Enhancement）系統、以及通訊傳輸媒介的好壞。整體而言，目前常用的語音品質評估方法可以分為兩大類：主觀評估方法以及客觀評估方法。以下將分別針對兩者進行詳細的說明。

doi:10.6342/NTU202300029

2.3.1 主觀語音品質評估方法

主觀語音品質評估方法主要透過人類聽覺系統的主觀偏好來衡量語音品質的好壞，為目前所有語音品質評估方法的標竿。常見的主觀語音品質評估方法有 A/B 測試（A/B Preference Test）、平均主觀意見分數（Mean Opinion Score）等等。其中，平均主觀意見分數為最常使用、也最具代表性的主觀評估方法。其定義為所有受試者「在一個預先定義的範圍內，對所分配訊號品質表現的意見」的算術平均數，詳細流程如圖 2.11所示。具體來說，給定待測的語音訊號，該段語音會隨機分配給該次聆聽測驗中的多個受測者，而這些受測者會在聆聽完該段語音後一般常用絕對類別評分（Absolute Category Rating）來表示這個範圍（如圖 2.12所示），其中 1 表示最低感知品質，5 表示最高感知品質。而將多個受試者的評分結果取平均後，即可得到該段語音的平均主觀意見分數。

平均主觀意見分數的優點在於它並不受限於待測的語音種類，幾乎可以使用於各種場景，且其評量結果與人們的普遍喜好直接相關，從可信度（Reliability）上來看是最理想的語音品質評估方法。此外，透過平均主觀意見分數的評量結果我們也可以量化不同待測訊號或系統之間的表現差距，以便於比較。

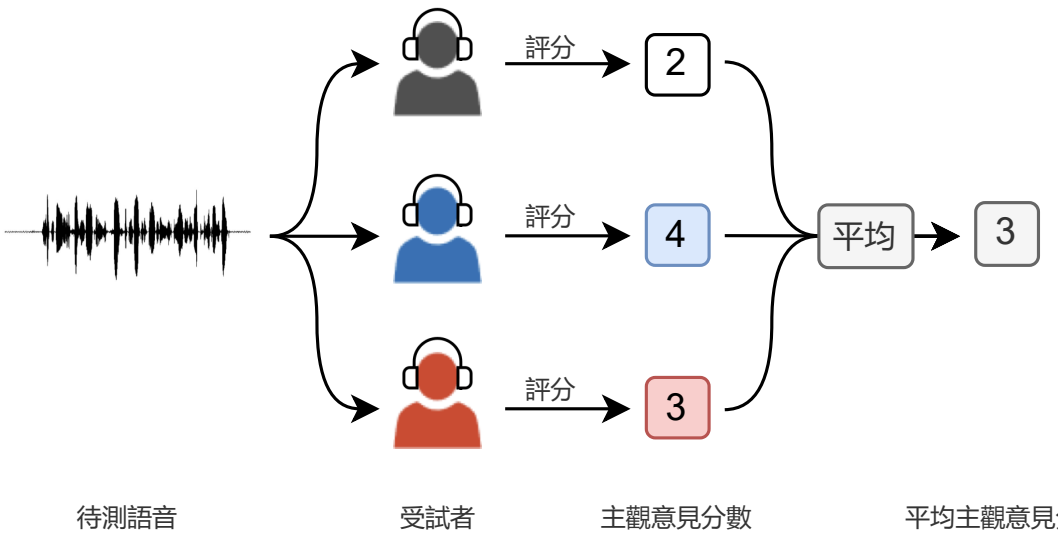


Figure 2.11: 平均主觀意見分數流程圖。

doi:10.6342/NTU202300029

標籤	極差	差	普通	好	極好
評分	1	2	3	4	5

Figure 2.12: 絕對類別評分（Absolute Category Rating）。

2.3.2 客觀語音品質評估方法

雖然主觀評估方法可以提供高可信度的評估結果，然而，評估過程中需收集多個受試者的聆聽測驗結果，使得這些方法在實際應用時相當沒有效率。對此，有多個研究者試圖利用特定演算法來量化語音品質的好壞，即是所謂的客觀語音品質評估方法。理想上來看，完美的客觀評估方法需要整合語音訊號中「低層次（聲音強度、音高）」以及「高層次（韻律、語意）」的資訊，並且預測的結果需要盡可能接近主觀的評估結果 [39]。而依照所需要的訊號來源，我們可以將可將客觀評估方法進一步細分為全參考（Full-reference）評估方法與無參考（No-reference）評估方法，基本概念如圖 2.13所示。

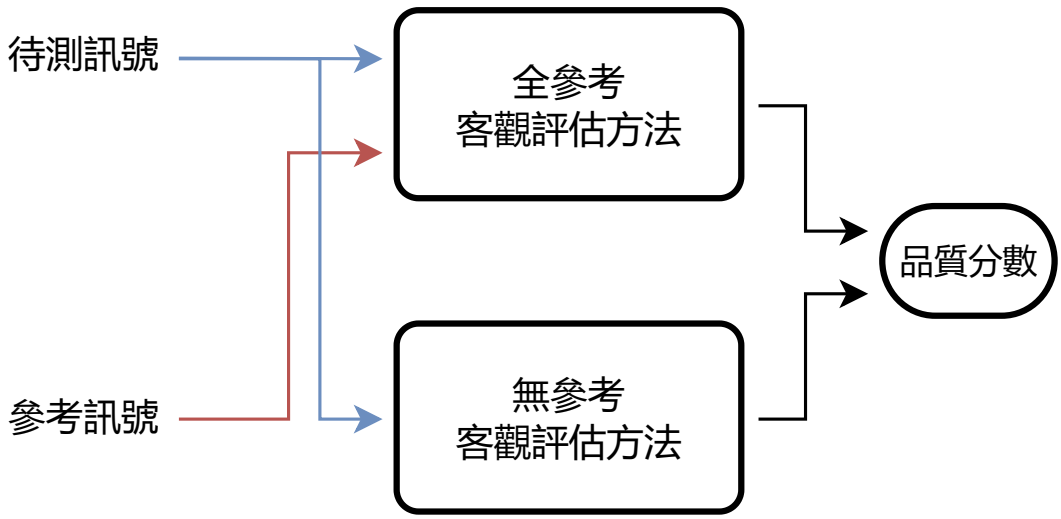


Figure 2.13: 客觀語音品質評估方法概念圖。

doi:10.6342/NTU202300029

全參考客觀評估方法

全參考客觀語音品質評估方法在所有的語音品質評估方法的分類中，為最常使用的一種。這類方法主要透過分析真實乾淨的參考訊號與失真的待測訊號之間的差異，並將這些差異映射到某種指標上，以衡量待測訊號的品質好壞。時至今日，已有多種全參考客觀語音品質評估方法被提出，這些方法根據不同使用情境設計，針對不同種類的語音失真提供客觀的品質指標。

舉例來說，在衡量通訊網路的品質時，經常使用語音品質感知評估（Perceptual Evaluation of Speech Quality）來檢測特定種類的失真——例如，封包丟失（Packet loss）、編解碼器（codec）造成的損失、延遲、中斷等等——所造成的影響程度，語音品質感知評估主要透過在分析待測訊號與參考訊號在時間維度和頻率維度上的差異性，並根據其差異性計算對應的品質分數；此外，短時客觀可理解性（Short-Time Objective Intelligibility）則借助短時距傅立葉變換來衡量待測語音內容的可理解程度；另一方面，對於語音生成任務，例如文句翻語音（Text-to-Speech）及語音轉換上，則經常使用梅爾倒頻譜距離（Mel-cepstral Distance）來計算生成語音與目標間的差異。

全參考客觀語音品質評估方法的優點在於計算速度相對較快，且被證實與主觀的評估結果呈現高度相關，可信度相當高。然而，由於其評估過程中需要清晰的原始輸入訊號作為參考，使得這類方法的使用範圍天生就受到限制。

無參考客觀評估方法

當我們需要即時的語音品質評估結果（意即我們無法使用主觀評估方法），但又無法取得參考訊號時（例如，在文句翻語音與語音轉換任務時，想生成一段目標語者沒有說過的內容；或在通訊時，接收端只收到失真的訊號），無參考的客觀語音品質評估方法便是我們最理想的選擇。無參考客觀語音品質評估方法，顧名思義，我們只需要待測的語音訊號即可預測主觀的品質評估結果。直觀上來看，我們可以透過語音增強技術生成人工的（artifact）參考訊號，並套用前一小節所述的全

doi:10.6342/NTU202300029

參考客觀語音品質評估方法來達成目的；此外，有些方法則是建構在人類的發聲系統與聽覺感知系統的特性上，例如，ITU-T 建議 P.563 [40] 首先藉由對口腔管（vocal tract）發聲系統的建模識別語音訊號中的失真種類，接著評估各種失真所造成的影響，最後利用這些因素的影響程度預測語音的品質分數量化成品質分數。

這些無參考客觀語音品質評估方法，與前述的全參考客觀語音品質評估方法相同，具有計算速度快的優點，此外，不需參考訊號的特性也使得這類方法在使用上具有更大的可能性。然而，傳統的無參考客觀語音品質評估方法在評估的可信度上遠遠落後於其他種類的方法，在某些使用情境下的表現相當糟糕，也因此未被廣泛地使用於現今的語音品質評估的場合中。近年來，有多個研究試圖利用深層學習開發與人類主觀偏好高度相關的無參考客觀語音品質評估方法，我們將在下個章節介紹。

2.4 本章總結

本章依序介紹了深層類神經網路、語音表徵及各種語音品質評估方法的基本概念，在接下來的章節中將繼續介紹基於深層學習的無參考客觀語音品質評估方法，並也將探討自監督語音表徵的各種性質及其是否適用於無參考客觀語音品質評估。

doi:10.6342/NTU202300029

第三章　基於深層學習的無參考客觀語音品質評估

3.1　簡介

在各個種類的語音品質評估方法中，無參考客觀語音品質評估方法（No-reference Objective Speech Quality Assessment）的特點在於，評量過程中無需使用乾淨真實的參考訊號（Reference Signal）作為對照，單單只需分析待測訊號即可預測主觀的品質評估結果。然而，受限於技術發展，傳統的無參考客觀語音品質方法在特定情境以外的評估準確度與可信度相當糟糕，遠不如全參考客觀語音品質評估方法（Full-reference Objective Speech Quality Assessment）。近年來，隨著深層學習（Deep Learning）的崛起，有多項研究試圖讓機器從「資料」中學習人類為語音品質進行評分時隱含的行為準則，來模擬主觀的評估結果 [5, 6, 41–47]。這些方法多半以平均主觀意見分數（Mean Opinion Score）作為模型的預測目標，並被證實與平均主觀意見分數的評估結果有高度的相關性，使人們在語音品質評估方法的發展上看見了新的可能。以下將會介紹深層無參考客觀語音品質評估模型的基本架構與相關技術。

doi:10.6342/NTU202300029

3.2 模型架構

正如我們在 2.3.2小節所說，理想的客觀評估方法為了近似人類聽覺系統評估時的偏好，需要能夠整合語音訊號中「低層次」的資訊以及「高層次」的資訊。其中，低層次的資訊包含聲音的強度、音高等等；而高層次的資訊則包含韻律、語義、語法等等。也因此，如何抽取並分析這些資訊便是深層無參考客觀語音品質評估模型的技術重點。常見的深層無參考客觀語音品質評估模型由一個特徵提取模組（Feature Extractor）、一個時間建模模組（Temporal Modeling Module）以及預測模組（Predictor）構成，並以待測訊號的梅爾時頻圖（Mel-spectrogram）作為輸入。模型示意圖請參見圖 3.1：

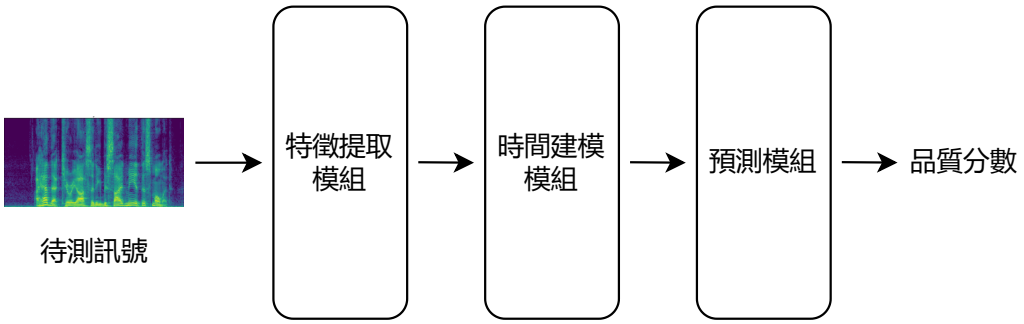


Figure 3.1: 基本的深層無參考客觀語音品質評估模型示意圖

其中，特徵提取模組在接收待測訊號後，會將各音框轉換成一連串的隱藏表徵（Hidden Representations），這些表徵中含有較低層次的語音資訊。接著，時間建模模組，通常由遞迴式類神經網路（Recurrent Neural Network）構成，會對不同時間點的表徵進行建模，從而得到與時間相關（time-dependent）資訊的隱藏表徵，這個步驟使得模型得以抽取語音中與品質相關的資訊，並處理不同時間之間的相依關係。以現實例子做說明，人們在衡量語音品質好壞時，很大一部分會依賴語音的語調變化來協助判斷；此外，對於人類的喜好來說，在話語中出現的噪音明顯比無聲的片段中出現的噪音更為惱人。顯而易見的是，這些資訊無法直接從訊號中取得，若不對低層次的語音資訊進行時間建模，模型將無法準確衡量這些因

doi:10.6342/NTU202300029

素所帶來的影響。最後，預測模組則需要學習辨認高品質語音與低品質語音經過時間建模模組輸出的表徵的差異，並將這些表徵映射為音框層級（frame-level）的品質分數，並透過平均合計（Mean Pooling）預測語句層級（utterance-level）的主觀評估結果（即其平均主觀意見分數）。在訓練時，模型通常透過最小化預測分數與實際分數之間的平均絕對誤差（Mean Absolute Error）或均方誤差（Mean Squared Error）進行學習：

$$\theta = \arg\min_{\theta} \frac{1}{S} \sum_{S} (y_s - F(s;\theta))^2 \tag{3.1}$$

$$\theta = \arg\min_{\theta} \frac{1}{S} \sum_{S} |y_s - F(s;\theta)| \tag{3.2}$$

其中，$F(s;\theta)$ 及 y_s 分別代表第 s 個語音訊號的預測分數及實際的平均主觀意見分數、θ 為模型參數、資料集中的語音總數為 S。

3.3 相關技術

3.3.1 聆聽者相關網路

由於平均主觀意見分數在實施時相對較費時，也因此，目前現存的公開資料集數量較少，且大多數資料集的規模並不大，可能會使得模型的表現受到限制。也因此，如何完整利用資料集中提供的資訊便是一件重要的課題。冷氏（Y. Leng）首先提出聆聽者相關網路（Listener Dependent Network）[5]，其基本概念如圖 3.2所示。在上述模型架構的基礎上，透過額外使用的網路預測特定聆聽者對於待測語句的主觀意見分數（Opinion Score）。具體來說，假設我們將某筆語音交由 5 位受試者評分，並得到 5 筆主觀意見分數 $\{2, 3, 4, 4, 3\}$，則其平均主觀意見分數為 3.2 分。在聆聽者相關網路的訓練中，除了以平均主觀意見分數作為預測目標，也會額外使用各個不同受試者（聆聽者）的評分分佈。這套方法使得模型能夠學習考慮聆聽測驗中各受試者的評分偏好差異，也同時增加可用於訓練的資料

doi:10.6342/NTU202300029

量，進而提升模型表現。而黃氏（W.C. Huang）則進一步引入平均聆聽者（Mean Listener）[6]，藉由對資料集中所有聆聽者的評分偏好的平均進行建模，以解決冷氏的模型中，推論階段（Inference time）無法使用聆聽者相依網路所造成的不匹配（mismatch）問題。

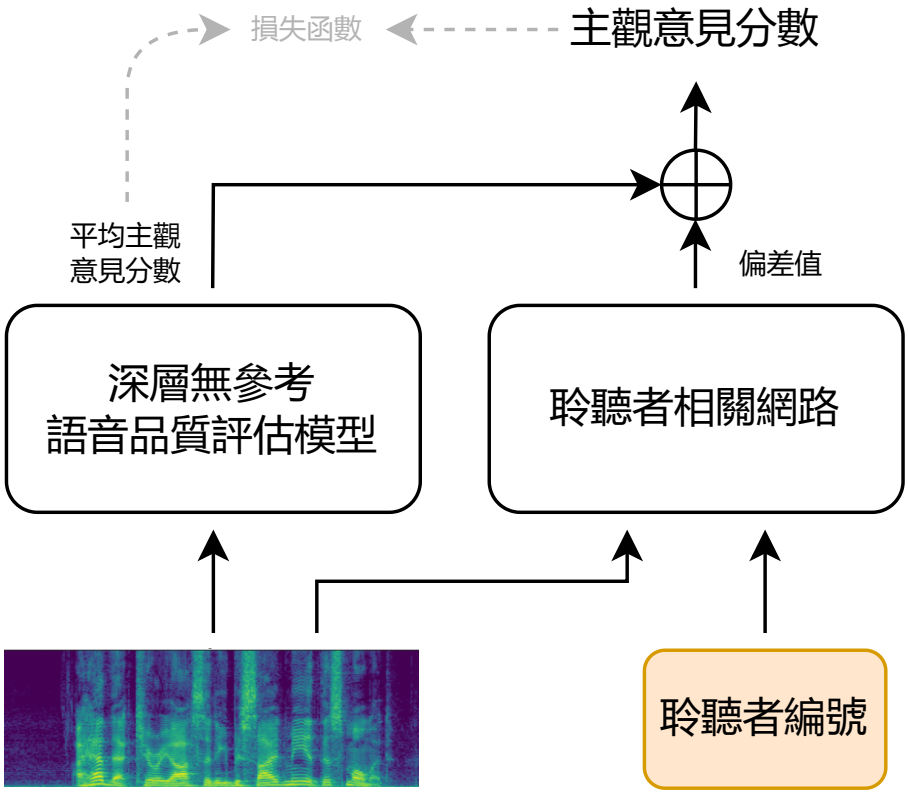


Figure 3.2: 聆聽者相依網路示意圖

3.3.2 轉移學習

如上一小節所述，目前現存的平均主觀意見分數資料集的規模較小，除了語料數量不足的問題外，其涵蓋的語音內容通常豐富度（variation）也較低，若單純將模型從頭訓練（Train-from-scratch）於這些資料集上，可能造成特徵提取模組無法完整抽取語音訊號中的低層次資訊，而時間建模模組也無法對不同時間的相依關係進行準確的建模。除此之外，模型也容易發生過度貼合（Overfitting）的現象，進而降低模型的表現與泛化能力（Generalizability）。有多個研究試圖透過

doi:10.6342/NTU202300029

轉移學習（Transfer Learning）解決上述問題：例如，索氏（M. Soni）與拉氏（A. Ragano）分別基於自編碼器（Autoencoder）以及深層分群演算法（Deep-clustering Algorithm），利用大量無標記語料來預訓練模型中的特徵提取模組 [41, 42]，使抽取的表徵隱含的資訊更為豐富；而米氏（G. Mittag）則額外收集了大量語料的全參考客觀評估結果，將整個模型預訓練在客觀品質分數的預測任務上 [43]。這些方法被證實能更好的初始化（Initialize）無參考客觀語音品質評估模型，進而提升表現。

3.3.3 專注機制

近年來，由於專注機制不停地在多項序列至序列任務——例如，語音辨識（Speech Recognition）、機器翻譯（Machine Translation）、語言模型（Language Model）——上攻城掠地，有數項研究也試圖將專注機制運用在深層無參考客觀語音品質評估模型上。曾氏（W.C. Tseng）首先利用專注合計（Attention Pooling，見2.1.4小節）來學習隱含表徵到品質分數的映射 [44]，並得以解決原先使用平均合計的預測模組無法處理的時近效應（Recency Effect，為心理學名詞，指在多種刺激一次出現時，印象的形成主要取決於後來出現的刺激）；而米氏（G. Mittag）則進一步使用基於專注機制的遞迴式類神經網路 [45]，使得模型進行時間序列上的建模時能考慮更多的全局資訊。

3.3.4 多任務學習

多任務學習（Multi-task Learning），又稱為多目標學習，指的是將多個相關任務放在一起學習的機器學習方法，其概念在於透過多個任務學習共享表徵（shared representation），提高模型的泛化能力，並同時避免模型收斂於局部最佳解（local optimum），其概念如圖 3.3所示。在基於深層學習的無參考客觀語音品質評估模型的研究中，我們可以透過同時預測待測語音的雜訊強度（Noiseness）、語調豐

doi:10.6342/NTU202300029

富程度（Colorization）、不連續性（Discontinuity）及響度（Loudness）[45]，或基於最小可感知差（Just-noticeable difference）及評分的一致與否設計多個損失函數來訓練模型 [46]，不僅可以提升模型表現，還可以透過其預測結果解釋不同面向的好壞所帶來的影響。此外，也有研究借助偽裝檢測（spoofing detection）讓模型學習辨別自然語音與生成語音的差異，使模型更能專注在決策邊界（decision boundary）附近的樣本，進而提升模型在預測自然語音的品質的準確度 [47]。

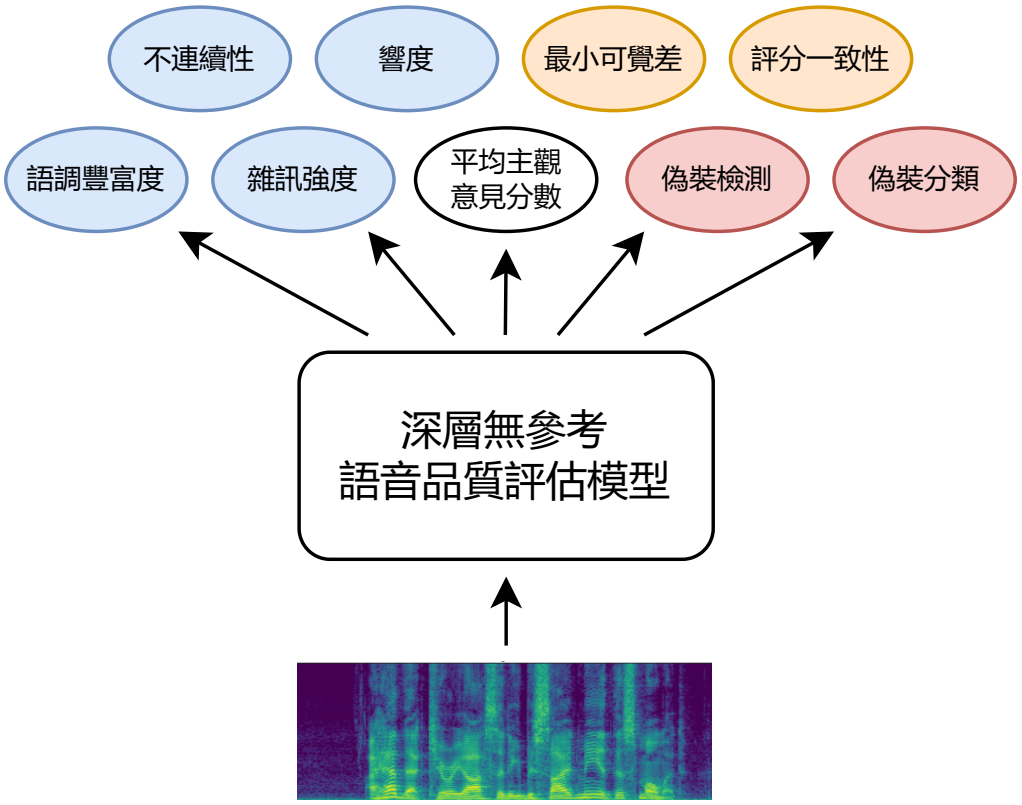


Figure 3.3: 基於多任務學習的深層無參考客觀語音品質評估模型模型

3.4 本章總結

在本章節中，我們首先簡介基於深層學習的無參考客觀語音品質評估。接著描述基本的無參考客觀語音品質評估模型架構，其由特徵提取模組、時間建模模組及預測模組所構成。最後我們介紹近年研究中所使用的相關技術，其中包含，聆聽者相關網路、轉移學習、專注機制及多任務學習。

doi:10.6342/NTU202300029

第四章　自監督式語音表徵用於無參考客觀語音品質評估之可行性分析

4.1　簡介

在前章節介紹了許多基於深層學習的無參考客觀語音品質評估（No-reference Objective Speech Quality Assessment）方法，這些方法被證實能一定程度地學習人類為語音品質進行評分時隱含的行為準則，而其預測結果也與主觀的評估結果呈現高度相關。然而，我們也可以從這些相關技術的介紹中看出，目前基於深層學習的無參考客觀語音品質評估方法所面臨的最大的瓶頸：高品質的平均主觀意見分數（Mean Opinion Score）資料集語料數量較少，豐富程度也不足，造成模型的表現與泛化能力受到限制。

另一方面，近年來，由於可利用的無標記語料大量出現，自監督式預訓練模型（Self-supervised Pre-trained Model）在語音處理領域上獲得了大量的關注。自監督式學習主要透過給定特定部分的資料，讓模型預測資料的另一部分 [37]，並藉此學習該種類資料隱含的內部結構。而透過自監督式學習預訓練的語音模型（以下簡稱自監督式語音模型），可以從原始的語音訊號中，抽取隱含豐富信息的表徵向量（即自監督式語音表徵，Self-supervised Speech

doi:10.6342/NTU202300029

Representation），這些表徵被證實能增進多個語音處理任務的表現，並超越過去單純使用監督式學習的方法，其中包含：語音辨識（Speech Recognition）、語者辨識（Speaker Identification）、語音理解（Spoken Language Understanding）及語音翻譯（Speech-to-speech Translation）等等。

雖然自監督式語音表徵在多個語音處理任務上被證實有良好的表現，但其仍未被運用於無參考客觀語音品質評估之應用上。我們合理地猜測，在其他語音處理任務上的成功有機會能夠複製到無參考客觀語音品質評估任務上。然而，顯而易見的是，自監督式語音模型在預訓練時並沒有看過不同品質的語音，且其預訓練目標也與預測語音品質毫無關聯。為了探討自監督式語音表徵應用於無參考客觀語音品質評估的可行性，在本章的內容中，我們試圖透過回答三個不同面向的問題去發掘其潛力：

- 首先，我們知道，人們在為語音訊號評分時，很自然地會先感知到響度（Loudness）、音高（Pitch）、語調（Prosody）等聲學（Acoustic）資訊 [48, 49]，並得以檢視訊號中是否有突發性的雜訊（Explosive Noise）、不合理的中斷、或是走音等影響品質的因素。我們同樣希望自監督式表徵中具有相同的資訊，因此第一個問題為「自監督式語音表徵中是否含有聲學相關的資訊」。
- 除了上述因素以外，語音訊號的「內容」是否清晰可辨，是否易於理解，對於語音訊號的品質也相當重要。若一段語音訊號中的內容模糊不清，則人們傾向於認為這段語音的品質較差。因此，我們期望得知「自監督式語音表徵中是否含有與語言內容相關的資訊」。
- 自監督式語音模型在預訓練過程中並沒有看過不同品質的語音，且其預訓練目標也與預測語音品質毫無關聯。那麼，我們想知道「自監督式語音表徵是否會受到語音品質的影響」，倘若自監督式語音模型對於不同品質的語音，其抽取的表徵之間有明顯的不同，則代表這些表徵中可能含有與語音品質相關的資訊。

doi:10.6342/NTU202300029

在本章中，我們首先設計兩套實驗來檢視自監督式語音表徵中是否聲學以及語言內容的資訊。接著，我們利用降維分析（Dimension Reduction Analysis）觀察不同品質語音的自監督式表徵在潛藏空間（latent space）上的分布狀況；最後，我們透過標準相關分析（Canonical Correlation Analysis）來分析自監督式語音表徵所含的資訊與語音品質分數的相關性。

此外，在現實生活中，無參考客觀語音品質評估的使用對象大致上包含了文句翻語音（Text-to-speech）系統和語音轉換系統（Voice Conversion）的生成語音、以及通訊傳輸造成的失真語音，而人們在為兩者評分時往往會參考不同的面向。舉例來說，對於生成語音，通常著重於衡量訊號聽起來是否「真實」，也就是所謂的自然度（Naturalness），而自然度可能受到訊號當中的機械音（Robotization）、語調影響；另一方面，對於通訊傳輸過程中造成的語音失真來說，則會著重於衡量訊號聽起來是否「清楚」，也就是所謂的可理解性（Intelligibility），訊號中的內容是否清晰可辨、背景中是否有噪音、傳輸過程中的中斷、抖動都會影響最終的評估結果。此外，人們對於這兩者聽起來的感受也明顯有所不同，生成語音的語調通常較為平淡。有鑑於這些不同之處，本章在探討自監督式語音表徵應用於無參考客觀語音品質評估的可行性時涵蓋了文句翻語音系統和語音轉換系統的生成語音、以及通訊傳輸造成的失真語音，以確保實驗在大多數的使用情境下都具有代表性。

4.2 資料集

本段將介紹實驗所使用到的資料集，其相關資訊與之後會使用的縮寫皆在表 4.1。若未特別說明，各平均主觀意見分數資料集中的評分範圍皆遵從 2.3.1小節中提及的絕對類別評分（Absolute Category Rating）。

- Voice Conversion Challenge 聆聽測驗資料集：Voice Conversion Challenge 聆

doi:10.6342/NTU202300029

聽測驗資料集中的語料收集自各年份的 The Voice Conversion Challenge 語音轉換競賽。資料集中除了包含各參賽系統的生成語音外，也涵蓋原先競賽中提供給參賽者的自然語音。每一筆語音都對應到若干個聆聽者依照自然度（Naturalness）給出的主觀意見分數（Opinion Score），以及最終的平均主觀意見分數（Mean Opinion Score）。

- Blizzard Challenge 聆聽測驗資料集：Blizzard Challenge 聆聽測驗資料集中的語料收集自各年份的 The Blizzard Challenge for TTS 文句翻語音競賽。資料集中除了包含各參賽系統的生成語音外，也涵蓋原先競賽中提供給參賽者的自然語音。每一筆語音都對應到一筆依照自然度評量的平均主觀意見分數。

- BVCC 資料集 [50]：BVCC 資料集為英文多語者的資料集，其語料收集自多個年度的 Voice Conversion Challenge 聆聽測驗資料集 [51–55] 及 Blizzard Challenge 聆聽測驗資料集 [56–61] 收集而成[1]，此外，資料集中也收集了 ESPNet 語音工具包 [62] 中的多個文句翻語音系統的生成語音。所有的語料經過統一的聆聽測驗，並依照其自然度進行評分，確保不同年度語料的品質分數的可比較性。此外，各聆聽者的主觀意見分數也被囊括在內。總語料數量為 7106 筆，音檔的原始取樣率為 16000 赫茲。在所有的實驗中，我們依照官方的建議將資料集分成語料數量為 4974、1066、1066 的三個子集。

- NISQA 資料集 [45]：NISQA 資料集為多語言多語者的資料集，其中包含多個通訊傳輸可能造成的品質損失（編解碼器、封包丟失、背景噪音）及即時通訊軟體錄製（手機、Skype、Zoom、WhatsApp）的失真語音訊號。每一筆語音除了對應的品質分數外，也根據其噪度（Noiseness）、語調豐富程度（Colorization）、不連續（Discontinuity）及響度（Loudness）做範圍為 1-100 的評分。資料集中共有 2 個訓練子集、2 個開發子集及 3 個測試子集。

- LibriSpeech [63]：LibriSpeech 為英文多語者的資料集，其語料收集自 Lib-

[1]BVCC 中使用的 Voice Conversion Challenge 年份為西元 2016、2018 及 2020 年，而 Blizzard Challenge 則為西元 2008 至 2011 年、2013 年及 2016 年

doi:10.6342/NTU202300029

riVox 所提供之有聲書，並有對應的文字標註。資料集被分為訓練集、驗證集與測試集，分別涵蓋了 2338、73 及 73 位語者。而此三個集根據語音的清晰程度，再被分為乾淨集與其他集。其中，960 小時的乾淨訓練集又更進一步劃分為 360 小時及 100 小時的兩個子集。音檔的原始取樣率為 16000 赫茲。LibriSpeech 被廣泛用在語音辨識（Speech Recognition）、音素辨識（Phoneme Recognition）與預訓練自監督式語音模型上。

- Libri-Light [64]：Libri-Light 為英文多語者的資料集，此資料集從 LibriVox 所提供之有聲書收集了大量無標註語料（超過 6000 小時），以及 10 小時的文字標註語料，音檔的原始取樣率為 16000 赫茲。Libri-Light 被廣泛用在預訓練自監督式語音模型上。

- TIMIT 資料集：TIMIT 為英文多語者的資料集，其語料收集自 630 位語者的麥克風錄音。總語料數為 6300 筆，音檔的原始取樣率為 16000 赫茲。TIMIT 被廣泛用在語音辨識、音素辨識任務上。

doi:10.6342/NTU202300029

Table 4.1: 本論文中所使用的資料集。

資料集名稱	縮寫	總語料/時數	語料來源	取樣率
Voice Conversion Challenge 2020 聆聽資料集	VCC2020	20658	語音轉換	24000
Blizzard Challenge 2013 聆聽資料集	BC2013	24196	文句翻語音	44100
BVCC 資料集	BVCC	7106	語音轉換、文句翻語音	16000
NISQA-SIM 訓練與開發子集	NISQA-SIM	12500	模擬通訊失真	48000
NISQA-LIVE 訓練與開發子集	NISQA-LIVE	1220	即時通訊錄製	48000
NISQA-FOR 測試子集	NISQA-FOR	240	模擬通訊失真、即時通訊錄製	48000
NISQA-P501 測試子集	NISQA-P501	240	模擬通訊失真、即時通訊錄製	48000
NISQA-LIVETALK 測試子集	NISQA-LIVETALK	240	即時通訊錄製	48000
LibriSpeech 訓練子集	LS-960	960 小時	LibriVox 有聲書	16000
LibriLight	LL-60k	6000+ 小時	LibriVox 有聲書	16000
TIMIT 資料集	TIMIT	6300	麥克風錄音	16000

doi:10.6342/NTU202300029

4.3 使用之自監督學習語音表徵

在本章的實驗中，我們使用了四種不同的自監督式語音表徵，其中包含：Wav2Vec 2.0、HuBERT、TERA 及 CPC，以下將針對各個模型做更詳細的介紹。

- Wav2Vec 2.0 使用官方提供的預訓練模型[2]，其模型架構包含 7 層的卷積式類神經網路及 12 層的轉換器編碼器。模型的輸出維度為 768。在預訓練過程中，模型透過從一群採樣中正確判別屏蔽時間的表徵進行對比式學習。目標函數為 InfoNCE，預訓練資料集為 LibriSpeech 的所有訓練子集。

- HuBERT 使用官方提供的預訓練模型[3]，其模型架構包含 7 層的卷積式類神經網路及 12 層的轉換器編碼器。模型的輸出維度為 768。在預訓練過程中，模型會在潛在空間上屏蔽部分的時間點，並透過預測屏蔽時間點的語音表徵分群進行遮罩聲學建模。預訓練在 LibriSpeech 的所有訓練子集上。

- TERA 使用官方提供的預訓練模型[4]，其模型架構包含 3 層的轉換器編碼器。模型的輸出維度為 768。在預訓練過程中，輸入語音訊號的 fMLLR（Feature space Maximum Likelihood Linear Regression）會被施以時間及頻率維度上的屏蔽，而模型透過還原屏蔽部分進行遮罩聲學建模。其預訓練資料集為 LibriSpeech 的所有訓練子集。

- 對比預測編碼（Contrastive Predictive Coding，記為 CPC）使用官方提供的預訓練模型[5]，其模型架構包含 5 層的卷積類神經網路加上 1 層長短期記憶（Long Short-Term Memory）類神經網路。模型的輸出維度為 256。在預訓練過程中，模型透過預測未來時間的表徵進行對比式學習。其目標函數為 InfoNCE，而預訓練資料集為 Libri-Light 中所有的語料。

[2]https://github.com/pytorch/fairseq/tree/main/examples/wav2vec
[3]https://github.com/pytorch/fairseq/tree/main/examples/hubert
[4]https://github.com/s3prl/s3prl
[5]https://github.com/facebookresearch/CPC_audio

doi:10.6342/NTU202300029

4.4 自監督式語音表徵中的聲學資訊實驗

在第一個實驗中，我們透過將自監督式語音表徵還原回對數梅爾時頻譜（Log Mel-Spectrogram），試圖檢視自監督式語音表徵中是否含有響度、音高、語調等聲學上的資訊。原因在於人們在為語音訊號評分時，會透過感知這些資訊，來辨別訊號中影響品質的因素。而我們可以透過分析對數梅爾時頻譜來得到這些資訊。因此，若我們能利用簡單的線性轉換，將自監督式語音表徵還原回對數梅爾時頻譜，則這些表徵中較可能含有聲學上的資訊。

4.4.1 實驗設置

實驗被進行在 TIMIT 資料集上。我們將輸入語音降低採樣至 16000 赫茲後經由自監督式語音模型得到音框層級（frame-level）的表徵序列，再訓練一層前饋式類神經網路（即單純線性轉換）將其還原回 80 維的對數梅爾時頻譜。訓練目標為最小化輸出與實際對數梅爾時頻譜間的平均絕對誤差（Mean Absolute Error）。模型訓練透過 Adam 最佳化器，批次大小訂為 16，學習率為 0.0002，一共訓練 50 期。每訓練一期我們就將模型測試在開發子集和測試子集上，最後呈現的為開發子集表現最好的存取點（checkpoint）在測試子集上的結果。

4.4.2 實驗結果

實驗結果列在表 4.2。表中左列所呈現的數字為預測結果與實際的對數梅爾時頻譜的平均絕對誤差，越低則表現越好。從結果中可以觀察出，所有自監督式語音表徵皆可以透過單純的線性轉換，一定程度的還原原始的對數梅爾時頻譜，說明自監督式語音表徵中的確含有聲學上的語音資訊。此外，我們還可以發現，基於遮罩聲學模型（Mask Acoustic Modeling）預訓練的表徵（HuBERT、TERA），其表現普遍好於基於對比式學習的表徵（Wav2vec 2.0、CPC）。

doi:10.6342/NTU202300029

Table 4.2: 自監督式語音表徵用於還原 80 維的對數梅爾時頻譜。

模型	平均絕對誤差
Wav2Vec 2.0	0.272
HuBERT	0.262
TERA	0.044
CPC	0.328

4.5 自監督式語音表徵中的語言內容資訊實驗

在第二個實驗中，我們透過將自監督式語音表徵用於音素辨識（Phoneme Recognition）任務，試圖檢視這些表徵中是否含有與語言內容相關的資訊，以及這些資訊是否貼近人類的語言系統。原因在於人們在為語音訊號評分時，除了考慮聲學上的資訊，其語音內容是否清晰可辨也大大地影響了訊號的品質。我們希望自監督式語音表徵中也同樣含有與語言內容相關的資訊，以用於無參考客觀語音品質評估。

4.5.1 實驗設置

實驗被進行在 LibriSpeech 的 100 小時乾淨訓練子集上。我們將輸入語音降低採樣至 16000 赫茲後經由自監督式語音模型得到音框層級的表徵序列，再訓練一層前饋式類神經網路預測各音框對應的音素。訓練目標為最小化輸出與實際標記間的交叉熵（Cross Entropy）。模型訓練透過 Adam 最佳化器，批次大小訂為 32，學習率為 0.0002，一共訓練 500000 次迭代。每更新 10000 次我們就將模型測試在 LibriSpeech 乾淨開發子集和乾淨測試子集上，最後呈現的結果為開發子集表現最好的存取點在測試子集上的錯誤率（Error rate），越低則表現越好。

doi:10.6342/NTU202300029

4.5.2 實驗結果

實驗結果列在表 4.3。表中的數字為在測試子集上的音素錯誤率（Phoneme Error Rate），代表預測結果中有多少百分比的音素預測錯誤，越低則表現越好。而為了衡量表現，我們另外訓練一個以對數梅爾時頻譜作為輸入的模型當作比較基準。從表中可以觀察出，所有自監督式語音表徵在音素辨識上的結果明顯好於對數梅爾時頻譜，說明自監督式語音表徵中與內容相關的資訊含量相當豐富。其中，Wav2vec 2.0 與 HuBERT 最為貼近人類所能理解的抽象層面（音素）。

Table 4.3: 自監督式語音表徵用於音素辨識。

模型	錯誤率（%）
Wav2Vec 2.0	5.74
HuBERT	5.41
TERA	49.17
CPC	42.54
對數梅爾時頻譜	82.07

4.6 降維分析實驗

從前面兩個實驗中我們可以看出，自監督式語音表徵中含有豐富的聲學資訊與語言內容資訊，顯示對於無參考客觀語音品質評估來說，使用自監督式語音表徵是相當合理的選擇。接下來，我們將更深入地發掘自監督式語音表徵用於無參考客觀語音品質評估的潛力。首先，我們知道，自監督式語音模型在預訓練過程中並沒有看過不同品質的語音，且其預訓練目標也與預測語音品質毫無關聯。然而，我們不知道的是，預訓練好的自監督式語音表徵是否會受到輸入語音的品質影響呢？直觀上來看，若自監督式語音模型針對高品質與低品質的語音，其抽取

doi:10.6342/NTU202300029

的表徵與傳統表徵相比，在空間上的分佈具有明顯的差異，則代表模型在抽取表徵的過程中會受到語音品質的影響，而這些表徵也較可能含有品質相關的資訊。基於上述，在本小節中我們利用降維分析觀察語音品質的好壞，是否影響其自監督式表徵在潛藏空間上的分佈狀況。

4.6.1 實驗設置

基於 4.1節所述，本實驗會分別進行在文句翻語音系統的生成語音、語音轉換系統的生成語音、以及通訊傳輸造成的失真語音上，使用的資料集包含了 VCC2020、BC2013、以及 NISQA 的所有子集[6]。而使用的自監督式語音模型包含了 Wav2Vec 2.0、HuBERT、TERA 以及 CPC。

我們分別從這三個資料集中收集若干筆高品質、中間品質、以及低品質的語音。語音輸入首先會被降低採樣率至 16000 赫茲，並經由自監督式語音模型抽取音框層級的表徵序列。接著，音框層級的表徵序列會經由平均合計（Mean Pooling）轉換為語句層級（utterance-level）表徵。最後，我們利用利用 t-SNE 演算法（t-Stochastic Nearest Neighbor Algorithm）將高維度的表徵向量投影至二維空間，並繪製其散佈圖（scatter plot）。此外，為了進行比較，實驗中還使用了各資料集中的自然語音作為對照。

4.6.2 實驗結果

我們首先觀察在語音轉換系統的生成語音（VCC2020）上的結果。不同品質語音的詳細資料列於表 4.4，而其自監督式表徵在二維空間上的分布情形如圖 4.1所示。

[6]值得注意的是，雖然 BVCC 資料集中已經包含文句翻語音和語音轉換系統的生成語音，然而，我們經初步研究發現文句翻語音和語音轉換的生成語音的自監督式表徵具有明顯的差異性，若使用 BVCC 資料集進行降維分析，其結果可能會同時受到上述的差異性以及語音品質兩項變因所影響

doi:10.6342/NTU202300029

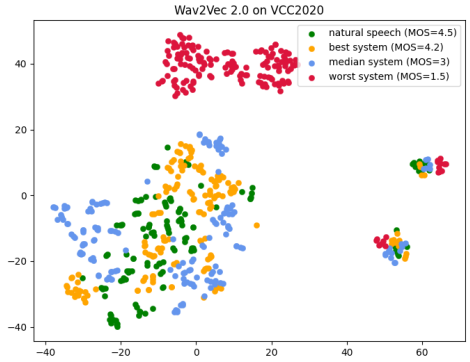
從圖中可以看出，對於所有的自監督式語音模型，低品質生成語音（紅色點）的表徵與其他語音明顯形成不同的分群（Cluster）；此外，根據自然語音（綠色點）、高品質生成語音（黃色點）及中間品質生成語音（藍色點）的分布關係可以發現，Wav2Vec 2.0、TERA 及 CPC 對於高品質生成語音輸出的表徵明顯在空間上更接近自然語音的分佈，說明這些模型可以進一步的區分高品質語音與中間品質語音。而在文句翻語音系統的生成語音（BC2013，請見 4.5和圖 4.2）以及通訊傳輸的失真語音（NISQA，請見 4.6和圖 4.3）上，我們也可以觀察到類似的現象。

整體而言，我們可以得出結論，無論對於生成語音及失真語音，自監督式語音模型在抽取表徵時皆會受到輸入語音品質的影響，造成不同品質語音的表徵在空間上呈現不同分佈。我們認為，其背後的原因在於自監督式語音模型在預訓練過程中僅看過高品質的自然語音，使其在面對較低品質的語音訊號時，所抽取的表徵中的結構性知識（structural knowledge）有所缺失，進而造成上述的現象。從另一個角度來看，這個現象也進一步地顯示自監督式語音表徵中的確含有與語音品質相關的資訊，而透過分析這些表徵的分佈方式，我們可能可以利用這些資訊來衡量語音的品質好壞。

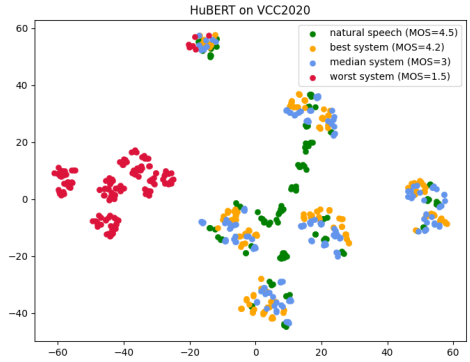
doi:10.6342/NTU202300029

Table 4.4: VCC2020 資料集（語音轉換）中用於繪製二維投影散布圖的語音資訊。

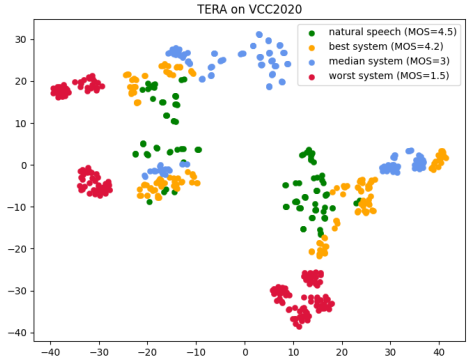
	生成系統	總語音數量	平均主觀意見分數
自然語音	來源語者、目標語者	130	4.8
高品質生成語音	T01、T13	160	4.5
中間生成語音	T24、T12	160	3.0
低品質生成語音	T14、T26	160	1.5



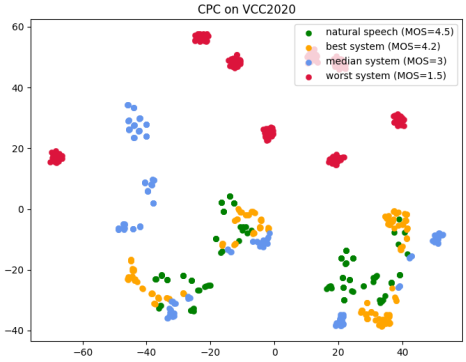
(a) Wav2Vec 2.0



(b) HuBERT



(c) TERA



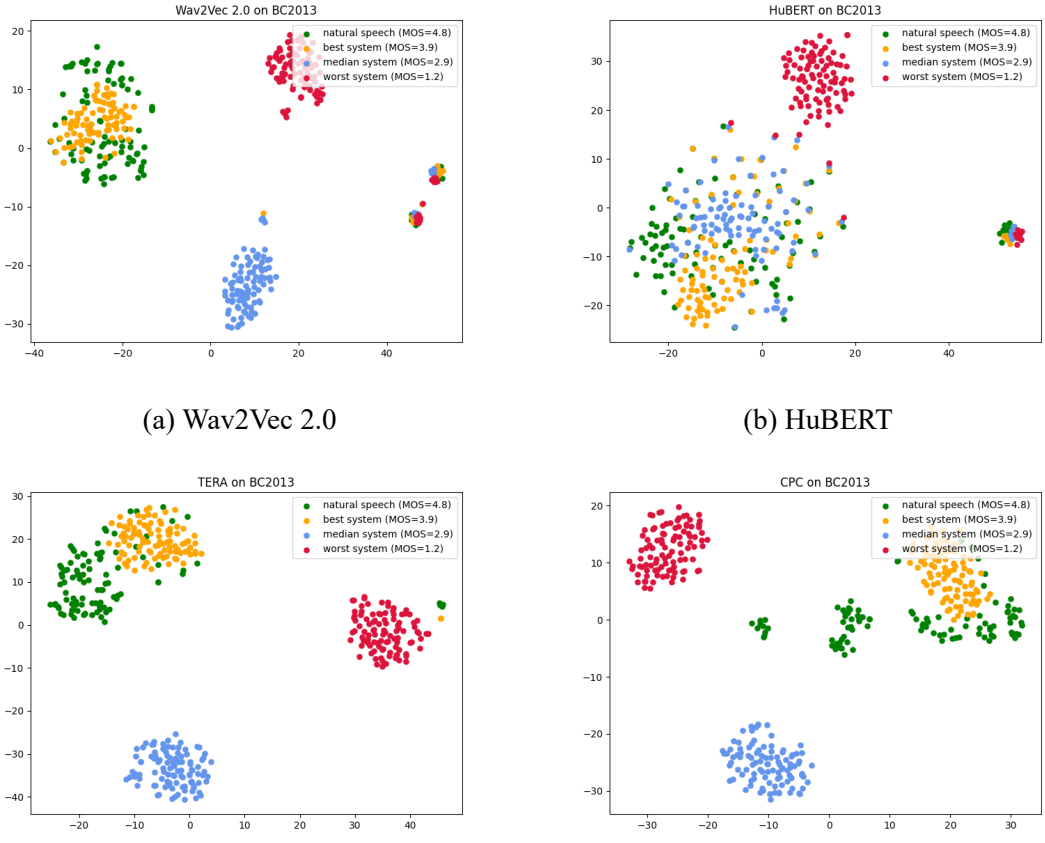
(d) CPC

Figure 4.1: 自監督式語音表徵（Wav2Vec 2.0、HuBERT、TERA、CPC）於 VCC2020 資料集（語音轉換）的二維投影散布圖。其中，綠色點代表自然語音，黃色點代表高品質生成語音，藍色點中間生成語音，而紅色點則代表低品質生成語音。

doi:10.6342/NTU202300029

Table 4.5: BC2013 資料集（文句翻語音）中用於繪製二維投影散布圖的語音資訊。

	生成系統	總語音數量	平均主觀意見分數
自然語音	目標語音	100	4.8
高品質生成語音	M	100	3.9
中間生成語音	C	100	2.9
低品質生成語音	P	100	1.2



(a) Wav2Vec 2.0

(b) HuBERT

(c) TERA

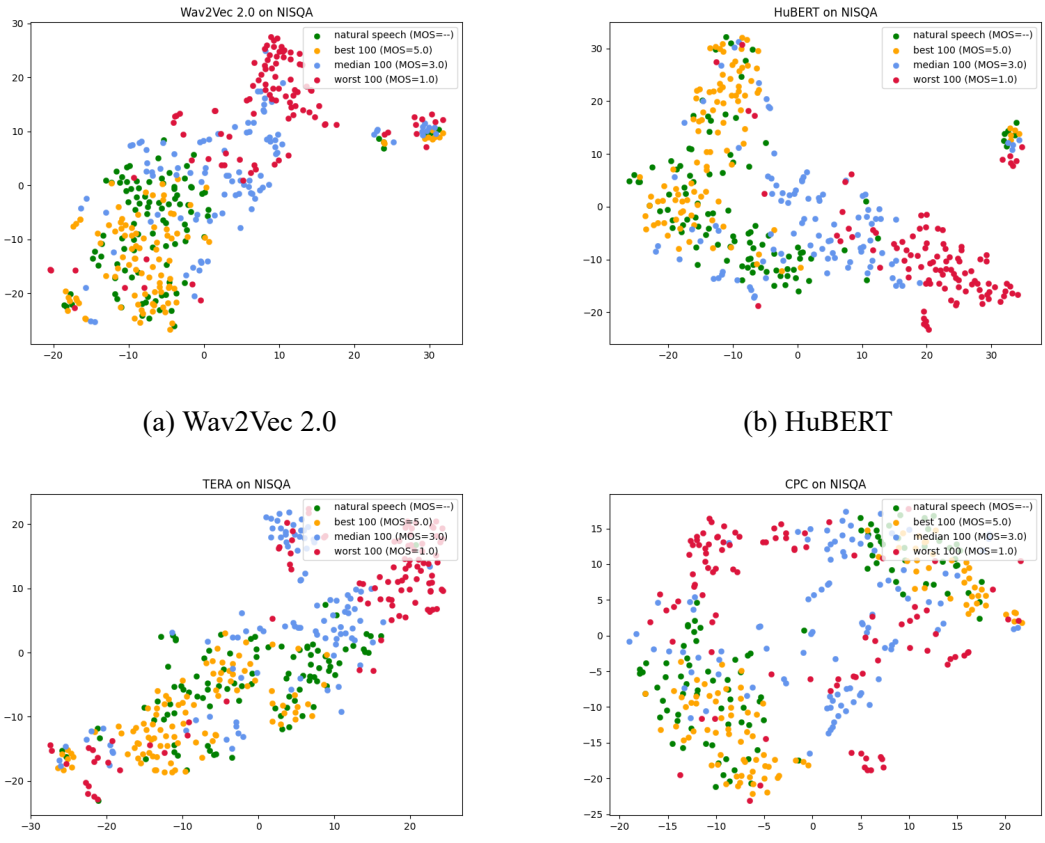
(d) CPC

Figure 4.2: 自監督式語音表徵（Wav2Vec 2.0、HuBERT、TERA、CPC）於 BC2013 資料集（文句翻語音）的二維投影散布圖。其中，綠色點代表自然語音，黃色點代表高品質生成語音，藍色點中間生成語音，而紅色點則代表低品質生成語音。

doi:10.6342/NTU202300029

Table 4.6: NISQA 資料集（通訊傳輸的失真語音）中用於繪製二維投影散布圖的語音資訊。在 NISQA 資料集中，自然語音並沒有相對應的評分。

	總語音數量	平均主觀分數
自然語音	100	–
高品質生成語音	100	5.0
中間生成語音	100	3.0
低品質生成語音	100	1.0



(a) Wav2Vec 2.0 (b) HuBERT

(c) TERA (d) CPC

Figure 4.3: 自監督式語音表徵（Wav2Vec 2.0、HuBERT、TERA、CPC）於 NISQA 資料集（通訊傳輸的失真語音）的二維投影散布圖。其中，綠色代表自然語音，黃色、藍色、紅色則分別代表品質最好、最靠近中位數及最差的 100 筆失真語音。

doi:10.6342/NTU202300029

4.7 標準相關分析實驗

從降維分析實驗的結果中，我們可以看出自監督式語音表徵中的確含有與品質相關的資訊，而不同品質之間的語音在空間上也呈現不同分佈。然而，我們並不清楚這些資訊與人類主觀認定的品質好壞之間具有多大的關聯性。在本小節中，我們使用標準相關分析（Canonical Correlation Analysis）進一步分析自監督式語音表徵中所含的資訊與主觀評估結果（即平均主觀意見分數）之間的相關程度。

標準相關分析為一種多變量統計的分析方法。給定兩組隨機變數向量，標準相關分析可以找出使得隨機變數向量之間的線性相關係數（Pearson's ρ）最大的兩組線性組合，也因此通常用來分析兩組隨機變數向量的相關性。直觀上來看，給定輸入語音的自監督語音表徵及平均主觀意見分數，若經標準相關分析計算出兩者間的相關程度越高，則代表自監督式語音表徵中所含有的資訊與語音品質好壞的關聯性越大，而自監督式語音模型在抽取表徵時的行為也更可能囊括了人類聽覺系統隱含的評分方式。

4.7.1 實驗設置

本實驗分別進行在 BVCC 資料集和 NISQA 資料集上。首先，我們收集了 BVCC 訓練子集、NISQA 的兩個訓練子集（NISQA-SIM 和 NISQA-LIVE）中的語料，語音輸入首先會被降低採樣率至 16000 赫茲，並經由自監督式語音模型抽取音框層級的表徵序列。接著，音框層級的表徵序列會經由平均合計轉換為語句層級表徵。接著，標準相關分析會找出語句表徵與對應的平均主觀意見分數的線性映射。我們使用 Pyrcca 套件[7]來實作標準相關分析，我們在 $[0.1, 10]$ 範圍內找出使線性相關程度最大的正規化係數（Regularization Coefficient）。最後，我們分別將表現最好的線性映射矩陣套用在 BVCC 測試子集和 NISQA 的兩個測試子集

[7] https://github.com/gallantlab/pyrcca

doi:10.6342/NTU202300029

（NISQA-FOR 和 NISQA-P501）中的語料的語句層級表徵上，並計算各語句線性映射後的結果與對應的平均主觀意見分數的線性相關程度，數值越高則代表相關程度越大。

4.7.2 實驗結果

標準相關分析的結果如表 4.7所示。表中的數字代表自監督式語音表徵與平均主觀意見分數的線性相關程度。首先，從 BVCC 測試子集的結果可以看出，各自監督式語音表徵與平均主觀意見分數呈現高度相關（皆大於 0.7），其中以 Wav2Vec 2.0、HuBERT、以及 TERA 的表徵表現最佳；而在兩個 NISQA 的測試子集上，也同樣可以發現所有的自監督式語音表徵都與品質分數呈現高度正相關，其中以 Wav2Vec 2.0 與 HuBERT 的表徵有最大的相關性。也因此，我們認為自監督式語音表徵中所含有的資訊與語音品質好壞具有很大的關連性，透過分析這些表徵中所含的資訊，我們可以為輸入語音進行品質上的排序，且其結果與人類的評分喜好呈現高度相關。此外，這樣的現象也暗示自監督式語音模型在抽取表徵時的行為，無論在生成語音或是失真語音上，與人類為這些語音評分時所考慮的面向有相似之處。

Table 4.7: 利用標準相關分析計算出不同自監督式語音表徵與平均主觀意見分數間的線性相關程度。

	BVCC 測試子集	NISQA-FOR	NISQA-P501
Wav2Vec 2.0	0.750	0.788	0.760
HuBERT	0.752	0.721	0.821
TERA	0.757	0.736	0.798
CPC	0.700	0.745	0.779

doi:10.6342/NTU202300029

4.8 本章結論

在本章中，我們著重於探討自監督式語音表徵於無參考客觀語音品質評估的可行性。我們首先設計兩個實驗，證明自監督式語音表徵中含有豐富的聲學資訊與語言內容資訊，其表現明顯好於實驗中的比較基準。接著，我們進一步深究自監督式語音表徵用於無參考客觀語音品質評估的潛力，我們使用二維投影散布圖觀察不同品質語音的自監督式表徵在空間上的分布狀況，並利用標準相關分析了解自監督式語音表徵中的資訊與語音品質好壞的關聯性。從這兩項實驗我們得知，自監督式語音模型對於不同品質的語音所輸出的表徵在空間具有不同分佈，且其含有的資訊與平均主觀意見分數具有高度的相關性。整體而言，我們可以得出結論，自監督式語音表徵對於無參考客觀語音品質評估之應用是相當理想的輸入表徵選擇。而透過總結所有的實驗結果，我們認為 HuBERT 具有最大的潛力。

doi:10.6342/NTU202300029

第五章　基於自監督式語音表徵的無參考客觀語音品質評估模型

5.1　簡介

在上一章中，我們探討了自監督式語音表徵在無參考客觀語音品質評估上的可行性，而透過實驗結果我們得知，自監督式語音表徵相當適合用於無參考客觀語音品質評估之應用，其中以 HuBERT 具有最大的潛力。有鑑於此，在本章中我們提出一套全新、基於 HuBERT 表徵的深層無參考客觀語音品質評估模型[1]，我們將 HuBERT 表徵作為模型的輸入，期望透過預訓練過程中學習到的隱含的結構性知識（structural knowledge）、以及自監督式表徵中的豐富資訊含量，進一步提升深層無參考客觀語音品質評估模型的預測表現（也就是評估結果的可信度）以及泛化能力（Generalizability）。

而在實驗環節中，我們利用多個指標量化比較語音品質評估的表現，結果證實其表現優於當前表現最好的深層監督式無參考客觀語音品質評估模型 LDNet [6] 以及 NISQAv2 [45]。此外，透過比較模型在不同語言的資料集上的表現，我們也證實使用自監督式語音表徵有助於提升無參考客觀語音品質評估模型的泛化能力。最後，我們使用探測分析（Probing Analysis）深入理解模型的行為，了解影響模型表現的因素，以提供未來可能的研究方向。

[1] 基於作者發表於於 InterSpeech 2021 的論文 [44]

doi:10.6342/NTU202300029

在本章內容中，將依序介紹本論文提出之模型架構、訓練方法、以及下列四項實驗的結果與分析：

- 與其他深層無參考客觀語音品質評估模型之比較實驗。
- 模型在不同語言上的泛化能力實驗。
- 模型面對不同類型的語料時，表現的可轉移性（Transferability）實驗。
- 不同品質的語音對模型表現的影響。

5.2 資料集

本章實驗所使用到的資料集，其相關資訊已在第 4 章中介紹，此處不在詳述。

- Blizzard Challenge 聆聽測驗資料集。
- BVCC 資料集。
- NISQA 資料集。

5.3 本論文提出之方法

5.3.1 模型架構

本論文中提出的深層無參考客觀語音品質評估模型包含了一個時間建模模組（Temporal Modeling Module）、一個專注合計模組（Attention Pooling Module）、以及範圍限幅（range clipping）。這個模型與過去方法最大的差異在於，單純使用自監督式語音表徵作為模型的輸入，期望透過自監督式語音表徵中豐富的聲學以及語言內容資訊，減輕時間建模模組的負擔，並進一步提升模型的表現與泛化能力。模型示意圖請見圖 5.1。

doi:10.6342/NTU202300029

輸入的語音訊號 s 首先會經由 HuBERT 抽取一連串的自監督式表徵序列：

$$\text{HuBERT}(s) = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T] \tag{5.1}$$

其中，T 為序列的總長度，而 $\mathbf{h}_i \in \mathbb{R}^d$ 為音框層級（frame-level）的自監督式表徵向量。接著，這些表徵向量會經由時間建模模組轉換為一連串的品質表徵序列：

$$\text{TemporalModelingModule}(\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T) = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T] \tag{5.2}$$

其中，$\mathbf{q}_i \in \mathbb{R}^m$ 為音框層級的品質表徵向量。具體而言，時間建模模組會從自監督式語音表徵中分離出對語音品質有影響的資訊，並整合這些資訊之間與時間相關（time-dependent）的關係。以語音中突然發生的中斷（Interrupt）為例，人們對於話語當中發生的中斷以及句子之間的中斷的感受明顯有所不同；另一方面，語調（prosody）上的變化也無法單從一個音框中的資訊得知，若不考慮不同時間點的資訊之間的交互關係，則無法進一步衡量這些因素的影響程度。而在抽取品質表徵序列後，專注合計模組會評估不同時間點的品質表徵對於整句語音訊號品質的影響程度，並利用加權平均（weighted-sum）以計算語句層級（utterance-level）的品質表徵，並透過線性轉換得到語句層級的品質分數 Q：

$$\text{AttentionPoolingModule}(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T) = Q \tag{5.3}$$

我們不直接取音框層級的品質表徵的平均的原因在於各時間的影響程度明顯不同，單純取平均明顯不符合人類評分時的行為模式。最後，由於我們並沒有對品質分數的輸出範圍進行限制，可能會得到不合常理的評估結果（例如-1 分），因此我們額外使用範圍限幅來得到最後的平均主觀分數預測結果 $\hat{y}$：

$$\text{RangeClipping}(Q) = \hat{y} \in [1, 5] \tag{5.4}$$

接下來我們將逐一講解各模組的內部細節。

doi:10.6342/NTU202300029

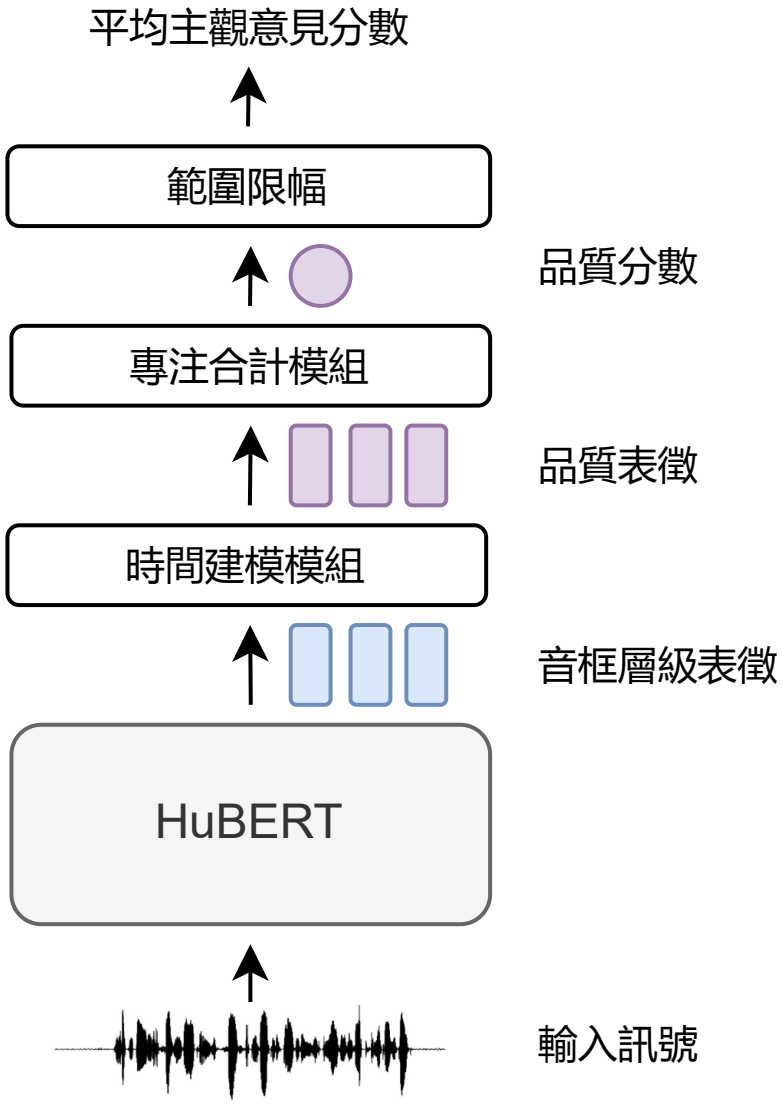


Figure 5.1: 基於 HuBERT 表徵的深層無參考客觀語音品質評估模型

doi:10.6342/NTU202300029

時間建模模組

時間建模模組由一層雙向長短期記憶體（Bi-directional Long Short-term Memory，記為 Bi-LSTM）類神經網路、一層前饋式類神經網路（Feed-forward Neural Network）以及線性整流單元（Rectified Linear Unit，記為 ReLU）組成，其示意圖請見圖 5.2a。其中，Bi-LSTM 的隱藏層維度設為 256，而輸出維度為 2×256=512。前饋式類神經網路的輸出維度為 256，我們透過初步實驗發現經由前饋式類神經網路降低品質表徵的維度能提升模型的表現。

專注合計模組

專注合計模組由專注合計（Attention Pooling，請見第 2.1.4小節）以及一層前饋式類神經網路組成，其示意圖請見圖 5.2b。時間建模模組輸出的品質表徵序列會先經由專注合計計算各時間點的權重，並透過加權平均得到語句層級的品質表徵向量，而前饋式類神經網路（即線性轉換）會將語句層級的品質表徵映射至語句層級的品質分數。

(a) 時間建模模組　　(b) 專注合計模組

Figure 5.2: 時間建模模組與專注合計模組的詳細架構圖。

doi:10.6342/NTU202300029

範圍限幅

範圍限幅為品質分數到絕對類別評分（Absolute Category Rating）範圍內的非線性映射，其計算方式為：

$$\hat{y} = 2 \times \tanh(Q) + 3 \tag{5.5}$$

其中，tanh 為雙曲正切（Hyperbolic Tangent）函數、Q 為專注合計模組輸出的品質分數、而 $\hat{y}$ 為模型的預測結果。可以看出，範圍限幅的輸出範圍在 1 到 5 之間。而在我們先前的研究中，範圍限幅也被發現能提升模型的表現。

5.3.2 訓練方法

模型的實作與訓練皆透過 Python 語言的 PyTorch 套件。在資料的預處理上，語音輸入會被降低採樣率至 16000 赫茲。我們以 S3PRL 工具包[2]來抽取自監督式語音表徵與訓練無參考客觀語音品質評估模型。模型透過最小化為預測分數與實際分數之間的平均絕對誤差（Mean Absolute Error）進行訓練：

$$\theta = \arg\min_{\theta} \frac{1}{S} \sum_{S} |y_s - F(s;\theta)| \tag{5.6}$$

其中，$F(s;\theta)$ 及 y_s 分別代表第 s 個語音訊號的預測分數及實際的平均主觀意見分數、θ 為模型參數、資料集中的語音總數為 S。訓練時使用了 Adam 最佳化器，$\beta_1 = 0.9$、$\beta_2 = 0.999$，初始的學習率設為 0，之後 1000 次迭代中會逐漸提升至最大值，並在達到高峰後以線性逐步遞減回 0，其中，最大值設為 0.0001。批次大小設為 16，一共訓練 50 期，每一期訓練結束後會利用開發子集衡量模型表現，而其中與開發子集間的平均絕對誤差最小的存取點會用於後續的結果討論。

[2]https://github.com/s3prl/s3prl

doi:10.6342/NTU202300029

5.4 基準方法

本章實驗中，比較表現時所用的基準方法，其中包含 LDNet [6] 以及 NISQAv2 [45]，兩者分別為目前 BVCC 資料集（生成語音）以及 NISQA 資料集（通訊傳輸的失真語音）上表現最好的模型。

5.4.1 LDNet

該模型使用編碼器-解碼器（Encoder-Decoder）架構，並基於3.3.1小節所提到的聆聽者相關網路進行變形，其中包含由特徵提取模組構成的編碼器、時間建模模組與預測模組構成的解碼器、以及一個額外的平均網路（Mean Net）。模型接收語音訊號的時頻譜（Spectrogram）作為輸入後，先經由編碼器抽取音框層級的表徵，這些表徵會與特定的聆聽者嵌入（listner-embedding）串接，並經由解碼器預測該聆聽者的主觀意見分數。此外，編碼器的輸出也會經由平均網路預測該段語音的平均主觀意見分數。值得注意的是，在模型訓練中會額外引入一個「平均聆聽者（Mean Listener）」，透過對資料集中所有聆聽者偏好的平均進行建模，使得模型在推論階段時能充分利用解碼器的能力。LDNet 為目前在 BVCC 資料集上的表現最好的模型。本論文的實驗中使用官方提供的預訓練模型[3]。其中，編碼器為 16 層的 MobileNetV3 卷積式類神經網路 [65]、解碼器包含一層雙向長短期記憶體類神經網路及兩層前饋式類神經網路、平均網路為兩層前饋式類神經網路。

5.4.2 NISQAv2

該模型包含特徵提取模組、時間建模模組以及預測模組。其特點在於模型中使用專注機制取代一般的類神經網路。其中，特徵提取模組為 6 層的卷積式類神經網路，時間建模模組為 2 層的轉換器編碼器層（Transformer Encoder），而預

[3]https://github.com/unilight/LDNet/tree/main/exp/Pretrained-LDNet-ML-2337

doi:10.6342/NTU202300029

測模組則為一個專注合計層和一層前饋式類神經網路構成。NISQAv2 為目前在 NISQA 資料集上的表現最好的模型。本論文的實驗中使用官方提供的實作程式[4]。訓練時參考官方的配置，批次大小（batch size）設為 40、學習率（Learning rate）設為 0.001、並使用 NISQA-SIM 及 NISQA-LIVE 訓練子集中的平均主觀意見分數訓練 500 期（epoch），而在每一期訓練結束後會利用開發子集衡量模型表現，並挑選其中表現最好的存取點（checkpoint）。

5.5 評量方法

在本章的實驗中，我們使用四個指標計算模型預測與主觀評估結果之間的誤差與相關程度，以量化模型的表現，其中均方誤差的值越小代表表現越好，而其他三個相關係數指標則是值越大代表表現越好。

- 均方誤差（Mean Squared Error，記為 MSE）：用來衡量兩變數間的誤差。
- 皮爾森動差相關係數（Pearson's r，記為 LCC）[66]：又稱線性相關係數，用來衡量模型兩變數間線性相依的相關程度。其範圍為 $[-1, 1]$，其中，1 代表完全正相關。由於 LCC 假設變數間為線性關係，使其在某些情況下無法正確衡量兩變數間的相關程度。
- 斯皮爾曼等級相關係數（Spearman's ρ，記為 SRCC）[67]：與 LCC 相對，SRCC 被用來衡量兩變數間的非線性相依的相關程度，其範圍為 $[-1, 1]$，其中，1 代表完全正相關。相較於 LCC 可以衡量更多類型的相關性。
- 肯德爾等級相關係數（Kendall's τ，記為 KTAU）[68]：用來衡量兩變數間次序大小的相關性，當兩變數間的次序完全相同時，其值為 1，其與 SRCC 的差異在於其對資料中的離群值較不敏感。

[4]https://github.com/gabrielmittag/NISQA

doi:10.6342/NTU202300029

5.6 與其他無參考客觀語音品質評估模型之比較實驗

首先，我們將本論文提出之模型與當前表現最佳的深層無參考客觀語音品質評估模型進行比較。而誠如 4.1節中所言，在現實生活中，無參考客觀語音品質評估的使用對象大致上包含了文句翻語音系統和語音轉換系統的生成語音、以及通訊傳輸造成的失真語音。為了證實我們的模型在這兩大使用情境下皆能準確地預測主觀的語音評估結果，我們分別使用 BVCC 資料集以及 NISQA 資料集訓練本論文提出之模型。並將訓練好的兩個模型分別與 BVCC 資料集上表現最好的模型——LDNet——以及 NISQA 資料集上表現最好的模型——NISQAv2——進行比較。

5.6.1 實驗結果

圖 5.3呈現了本論文提出之模型在 BVCC 測試子集上的預測表現。圖中的數字代表模型的預測結果和實際的平均主觀意見分數以不同評量指標計算出的量化數值。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。此外，在 BVCC 資料集中，由於我們可以得知各筆語音來自哪個語音生成系統，除了語句層級的預測表現外，系統層級（system-level）的預測表現也會一同被比較（某系統的平均主觀意見分數為該系統的所有生成語音的平均主觀意見分數的平均）。從圖中可以看出，本論文提出之模型無論是在語句層級或系統層級，其預測結果相較於 LDNet 有著更小的誤差（MSE），且進步幅度超過 25%。此外，在其餘三項相關係數指標的衡量上，我們的模型的預測結果明顯與平均主觀意見分數有更高的相關程度。這些結果顯示我們的模型在預測生成語音的主觀評估結果時的準確度與可信度明顯優於 LDNet，並也暗示縱使自監督式語音模型在預訓練過程中即使沒看過任何生成語音，其表徵中所隱含的豐富資訊仍使其在預測生成語音的主觀評估結果時，相較於傳統表徵更具有優勢。

doi:10.6342/NTU202300029

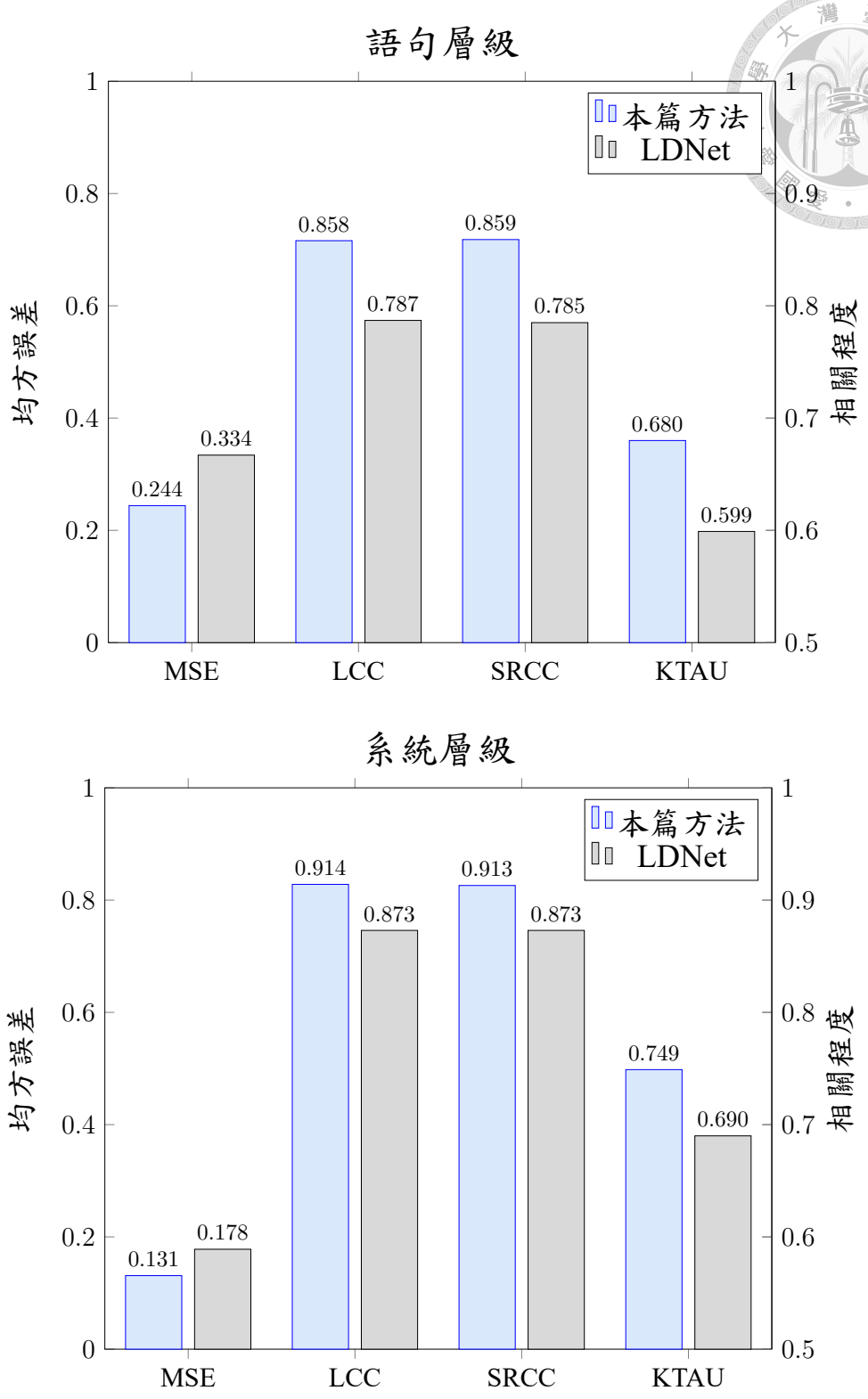


Figure 5.3: 本篇方法在 BVCC 資料集（生成語音）上的語句層級和系統層級表現之比較。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。

doi:10.6342/NTU202300029

接著，圖 5.4則呈現了本論文提出之模型在 NISQA 的兩個測試子集上的預測表現。首先，在 NISQA-FOR 測試子集上，我們可以看出，本論文提出之模型雖在所有指標上皆略優於 NISQAv2，但綜觀來看其表現屬於同個級別。然而，在 NISQA-P501 測試子集上，可以發現我們的模型無論是在預測結果的誤差或是相關程度指標的衡量上，皆明顯地比 NISQAv2 有著更好的表現，其中，在 MSE 上約有 10% 的進步，而 LCC 以及 SRCC 兩項指標更是大於 0.9，說明在預測通訊傳輸的語音品質時，自監督式語音表徵比起傳統語音表徵是更好的選擇。此外，由於 NISQAv2 在模型中使用多個基於專注機制的模組，這裡的實驗結果也可以說明我們的模型的好表現並不歸因於其中的專注合計，而來自於自監督式語音表徵中所隱含的豐富資訊。

整體而言，透過上述實驗結果我們可以合理地認為：由於自監督式語音表徵中含有豐富的聲學資訊和語言內容資訊，且不同品質的語音所抽取的表徵其中隱含的結構性知識與表徵分布方式也有明顯的不同，這些特質使其被運用於無參考客觀語音品質評估之應用時，相較於傳統語音表徵更具有優勢。而本論文中提出的基於 HuBERT 表徵的深層無參考客觀語音品質評估模型，其預測平均主觀意見分數的評估結果的表現全面超越過去使用傳統表徵的所有方法。

doi:10.6342/NTU202300029

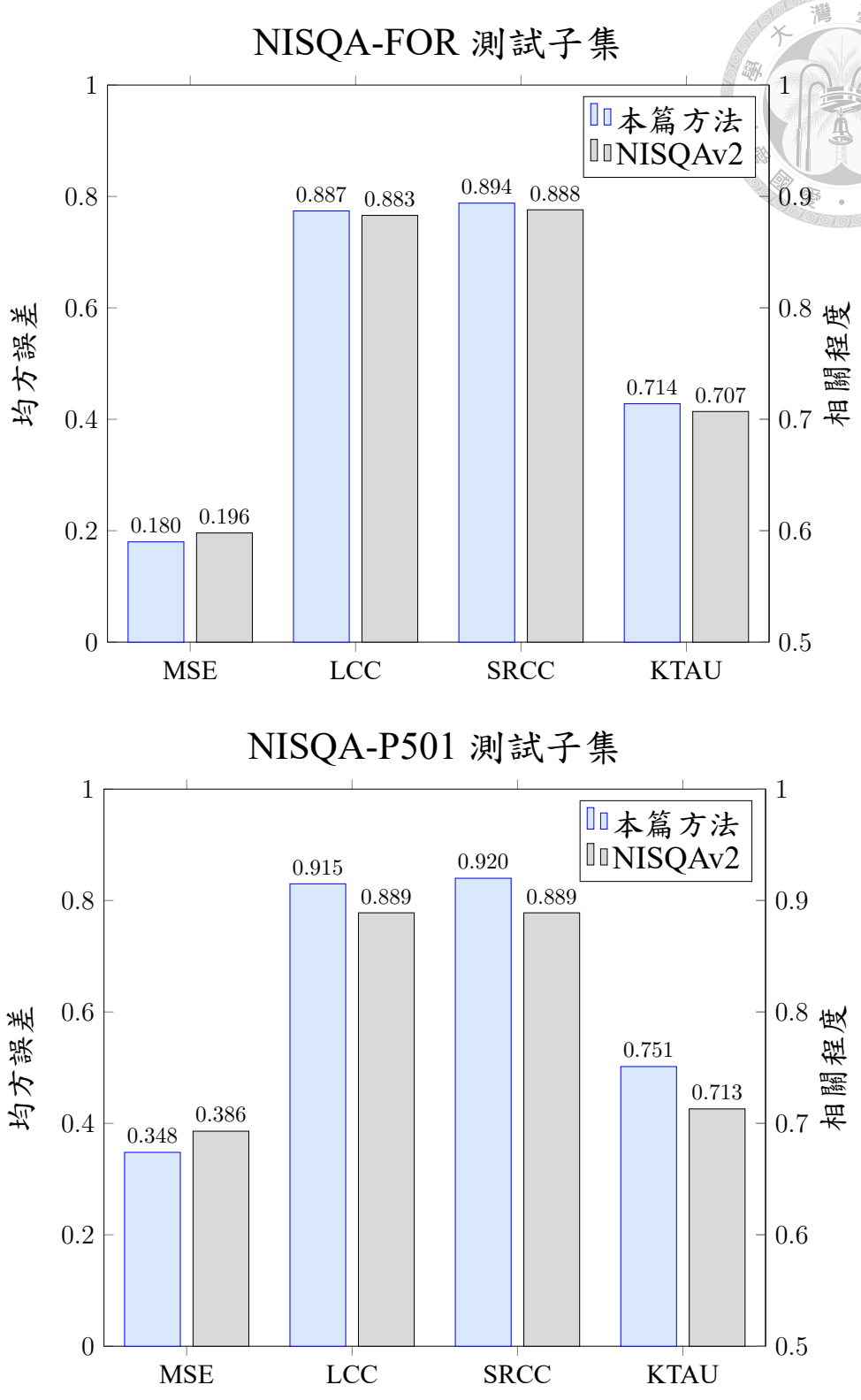


Figure 5.4: 本篇方法在 NISQA 資料集（通訊傳輸的失真語音）上的表現比較。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。

doi:10.6342/NTU202300029

5.7 對於不同語言的泛化能力實驗

當我們開發深層無參考客觀語音品質評估方法時，不僅期望模型在訓練資料集中的特定語言的語料上得到好的表現，我們還需要考量到，在實際應用層面上，使用者仍可能使用其他語言的語料作為待測訊號。也因此，我們將會希望模型在面對不同語言的語料時，也同樣具有良好的預測能力。在本實驗中，我們試圖檢視本論文提出之方法在面對不同語言的語料時的表現，也就是所謂的泛化能力，並與過去的方法進行比較。使用的資料集包含：

- BC2019 聆聽測驗資料集（BC2019）[69]: 收集自 2019 年 blizzard challenge 文字轉語音競賽，為中文多語者資料集。資料集中除了包含各參賽系統的生成語音外，也涵蓋原先競賽中提供給參賽者的自然語音。每一筆語音都對應到一筆依照自然度評量的平均主觀意見分數。
- NISQA-LIVETALK 測試子集（NISQA-LIVETALK）: 為德語多語者資料集，其中包含實際通訊時（Skype、家用電話）的錄音內容。每一筆語音都對應到一筆依照語音品質評量的平均主觀意見分數。

值得注意的是，由於資料型態不同（BC2019 為生成語音，而 NISQA-LIVETALK 為通訊傳輸的失真語音），前項實驗中以 BVCC 資料集訓練的模型會被測試在 BC2019 資料集上；而 NISQA 資料集訓練的模型則會使用 NISQA-LIVETALK 進行測試。

5.7.1 實驗結果

圖 5.5呈現了本論文提出之模型在不同語言上的預測表現。圖中的數字代表模型的預測結果和實際的平均主觀意見分數以不同評量指標計算出的量化數值。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標

doi:10.6342/NTU202300029

（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。首先，從 BC2019 資料集（中文）的結果可以看出，本論文提出之模型的預測結果與 LDNet 相比有著較大的誤差（MSE）以及較高的相關性，說明我們的模型雖然在數值上較無法貼近主觀的評估結果，但相較於 LDNet 更能反映不同品質語音之間的好壞關係。接著，在 NISQA-LIVETALK 測試子集（德文）上，可以看出，我們的模型不僅與主觀評估結果有更小的誤差，在相關程度上也遠勝於 NISQAv2。整體而言，我們認為，無論對於何種語言，使用自監督式語音表徵皆有助於提升深層無參考客觀語音品質評估模型的泛化能力，且其進步幅度相當巨大。

另一方面，藉由比較模型在兩個資料集上的表現，我們可以進一步發現其在 BC2019 資料集（中文）上的表現遠低於在 NISQA-LIVETALK（德文）上的表現，我們認為可能的原因有二：一、通訊傳輸過程中影響語音品質的因素對於不同語言的語音來說是共通的。例如，背景噪音、封包丟失等等。然而，對於生成語音來說，其中影響品質的因素則會因語言而異，不同語言的語調以及說話方式都不盡相同；二、相較於德文，中文與英文（HuBERT 的預訓練語料）在語言結構上明顯有較大的差異，無論是高品質或低品質的中文語音，其透過 HuBERT 抽取出來的表徵，當中所隱含的結構性知識很可能都有所缺失，進而造成後續的無參考客觀語音品質評估模型無法同樣有效地區分不同品質的語音。

doi:10.6342/NTU202300029

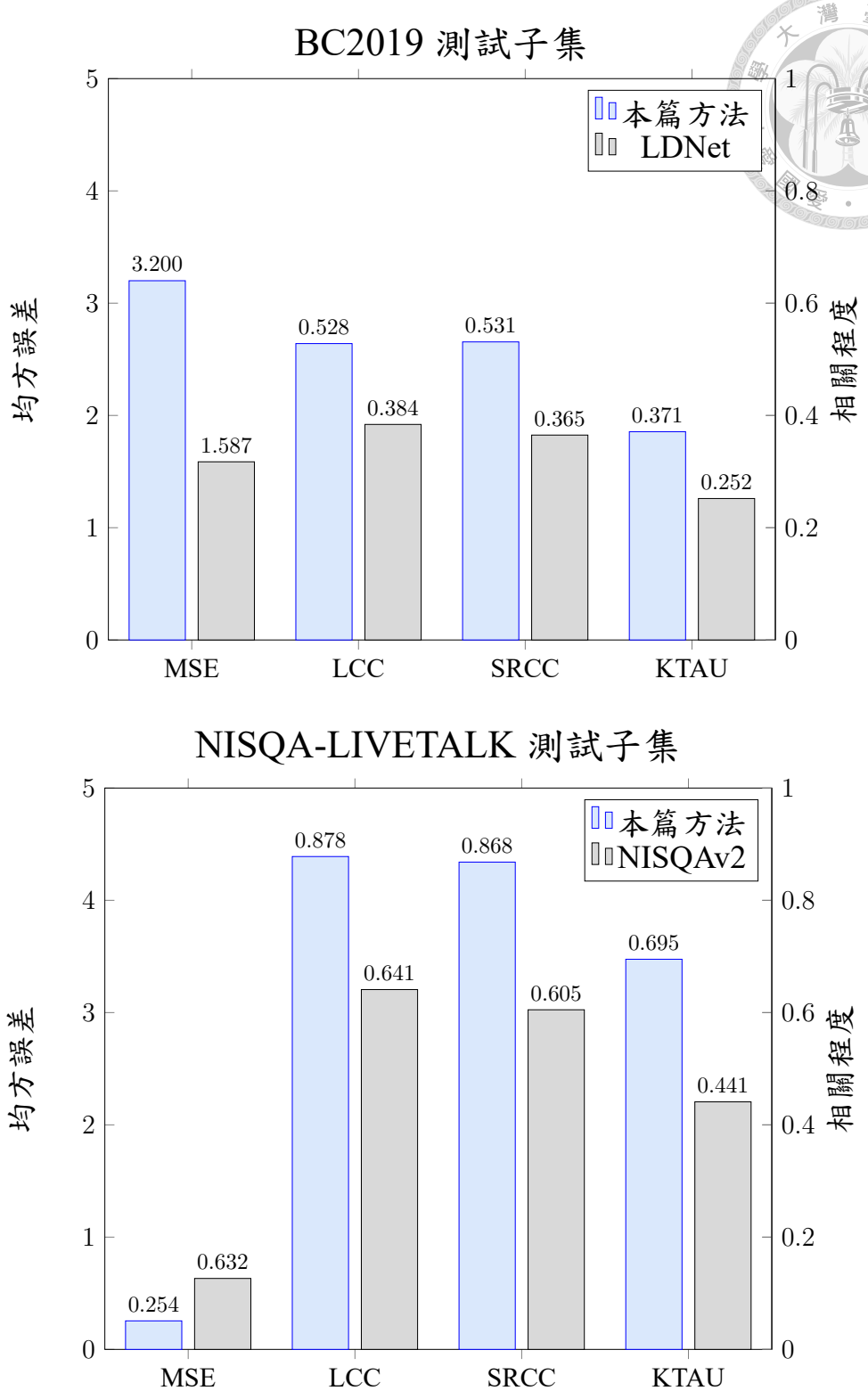


Figure 5.5: 本篇方法在 BC2019 資料集（中文）、以及 NISQA-LVETALK 測試子集（德文）上的泛化能力比較。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。

doi:10.6342/NTU202300029

5.8 對於不同類型語料的可轉移性實驗

在前兩項實驗中，我們著重於探討模型的表現，並證實我們提出的基於 HuBERT 表徵的無參考客觀語音品質評估模型，相較於過去使用傳統語音表徵的方法，其預測結果更接近主觀的評估結果，且在不同語言上的泛化能力也更好。接下來，我們則試圖透過探測實驗來理解模型的行為，進而為未來提供可能的研究方向。更精確地說，模型到底學到了什麼？模型對於生成語音以及通訊傳輸的失真語音，其學到的隱含的評分準則是否相同？我們能否開發更通用（universal）的無參考客觀語音品質評估模型？上述這些問題將會是我們的關注重點。

本實驗中使用的資料集為 BVCC 資料集和 NISQA 資料集，在前面實驗中訓練的兩個模型，會被測試在另一個資料集上。具體而言，以 BVCC 資料集訓練的模型會被測試在 NISQA 測試子集上；反之，以 NISQA 資料集訓練的模型則會被測試在 BVCC 測試子集上。我們透過這樣的方式衡量模型表現對於域不匹配（Domain Mismatch）的語料的可轉換性，進而探討模型在預測過程中的行為。

5.8.1 實驗結果

圖 5.6呈現了本論文提出之模型其原始的預測表現，以及轉換至不同類型語料上的結果（這裡稱為轉移表現）。圖中的數字代表模型的預測結果和實際的平均主觀意見分數以不同評量指標計算出的量化數值。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。而為了方便起見，我們將 NISQA-FOR 和 NISQA-P501 資料集上相同評量指標的數值取平均，概括為 NISQA 資料集上的表現。這裡我們同時觀察模型轉移至 BVCC 資料集和 NISQA 資料集上的結果，可以看出，模型在轉換至不同類型語料時，其預測結果的誤差明顯變大，且相關程度也有所下降，代表模型的轉移表現相較於原始表現來得差。這不僅顯示了這兩

doi:10.6342/NTU202300029

種資料在分布特性上的不同，也再次驗證人們在衡量生成語音以及通訊傳輸的失真語音的品質時，其隱含的評分準則是有所不同的（見 4.1節）。然而，我們也可以發現，模型的轉移表現仍舊維持著相當高的水準，不僅未出現不合理的誤差值（在兩個資料集上的 MSE 皆小於 0.7），其預測結果也與主觀評估結果依然有相當高的相關程度（在兩個資料集上，LCC 和 SRCC 兩項指標皆約為 0.8，而 KTAU 皆約為 0.6），這說明對於這兩類語音，其中影響語音品質的因素在某種程度上是共通的，而模型在訓練過程中也傾向於學習如何處理這些資訊。以現實的例子做佐證，無論在生成語音或通訊傳輸的語音，都可能受到背景噪音和突發性的噪音影響；此外，語音的內容是否清晰可辨識，對於這兩類語音來說也同等重要。

此外，我們也可以從另一個角度來思考上述結果，若我們能更深入的理解生成語音以及通訊傳輸的語音，其在資料分布特性上的不同（舉例來說，通訊傳輸語音其普遍的語調豐富程度明顯要優於生成語音），以及人們對於為兩類語音評分時，其隱含的評分準則之間的相同與相異之處，試圖讓模型學習考量更多不同面向的資訊，則更有可能得到一個通用的無參考客觀語音品質評估模型，也更接近所謂的「人工智慧」。

doi:10.6342/NTU202300029

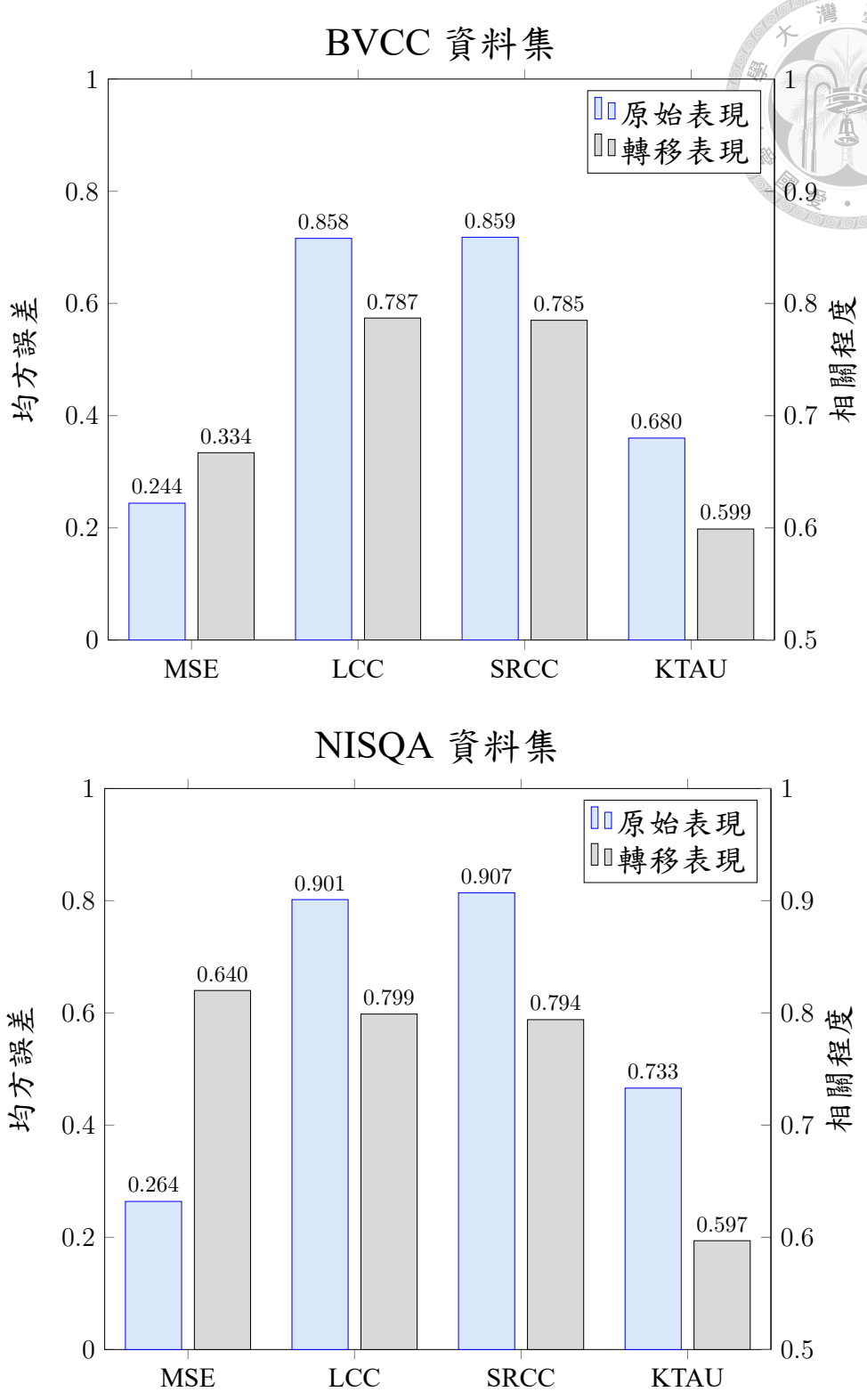


Figure 5.6: 本篇方法在 BVCC 資料集（生成語音）和 NISQA 資料集（通訊傳輸的失真語音）上的表現的可轉移性。其中，MSE 的數值對應到圖左側「均方誤差」的刻度，而其餘三項相關係數指標（LCC、SRCC、KTAU）則對應到圖右側「相關程度」的刻度。

doi:10.6342/NTU202300029

5.9 語音品質對模型的預測表現影響實驗

最後，我們試圖理解語音品質是否會對模型的預測表現有所影響。理想上來說，我們希望模型對於不同品質的語音皆有相同的預測表現，如此一來，其預測結果的可信度會大大增加；反之，則模型的預測結果不可盡信。舉個例子，倘若某個無參考客觀語音品質評估模型僅在預測低品質的語音時具有相當大的誤差，當這個模型預測一筆語音的平均主觀意見分數為 3 時，我們將無法判斷這筆語音是否真的為中間品質的語音。而在本實驗中，我們將語料依照對應的平均主觀意見分數分為低品質、次低品質、中間品質、次高品質、以及高品質語音[5]，分別計算這些語音的預測結果與主觀評估結果之間的平方差（Squared Error），並繪製成箱型圖（box plot）以衡量語音品質對模型的預測表現影響。使用的資料集為 BVCC 資料集和 NISQA 資料集。

5.9.1 實驗結果

圖 5.7呈現了本論文提出之模型對於 BVCC 資料集中不同品質區間的語音對表現的影響、以及預測結果與主觀預測結果的散佈圖（scatter plot）。其中，箱型圖中的每個箱子（box）共有五條橫線，由下往上分別代表該品質區間的所有預測結果的平方差的最小值、第一四分位數、中位數、第三四分位數、以及最大值[6]，若箱型圖整體越靠近下方，則說明該品質區間的預測表現較佳；而在散佈圖中，橫軸為各個語音訊號實際的平均主觀意見分數，縱軸則為模型的預測結果。首先，我們觀察箱型圖可以看出，模型在面對高品質的語音時明顯有較差的預測表現，而以散佈圖做為對照則可以發現模型的預測結果往往較實際的平均主觀意見分數來得低，說明模型無法正確顯示高品質語音的主觀評估結果。而這個現象在過去

[5]各品質區間語音的平均主觀意見分數範圍為：低品質 =[1,1.5]、次低品質 =(1.5,2.5]、中間品質 =(2.5,3.5]、次高品質 =(3.5,4.5]、高品質 =(4.5,5]。

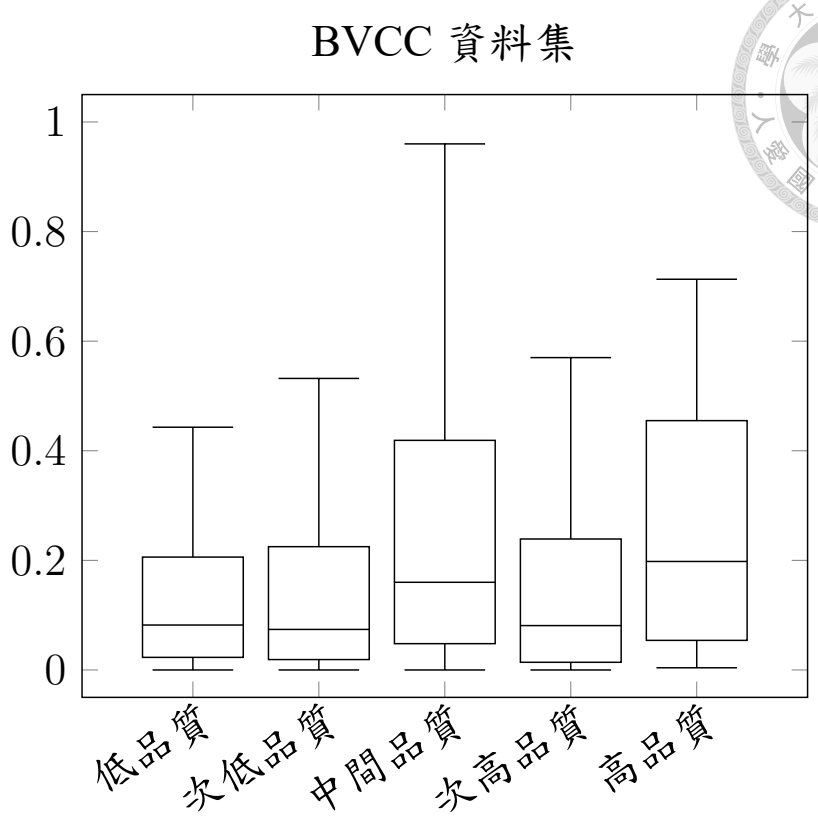
[6]以 BVCC 資料集的中間品質的箱子為例，在這個品質區間中，模型的預測結果與實際平均主觀意見分數的平方差的最小值約為 0、第一四分位數約為 0.05、中位數約為 0.16、第三四分位數約 0.42、而最大值約為 0.96

doi:10.6342/NTU202300029

的研究中也曾被發現 [47]，我們認為其主要原因為，在 BVCC 資料集中大多為語調豐富程度較低的生成語音，這使得模型在面對高品質的自然語音時無法正確衡量語調對語音品質帶來的影響。然而，從箱型圖我們也可以看出，模型在中間品質語音上的表現並不是太好，其預測結果的誤差度浮動很大，暗示了影響模型對於特定品質區間語音的預測表現的因素並不單純在於其語料數量（從散佈圖中可以看出中間品質的語音的數量明顯多於高品質語音的數量），試圖透過增加特定品質區間的語料數量來提升模型在該品質區間的預測表現可能不會是個有效率的解決方法。

另一方面，圖 5.8則呈現了在 NISQA 資料集上的結果，為了方便起見，這裡同時使用 NISQA-FOR 以及 NISQA-P501 測試子集中的語料。從箱型圖中可以看出，模型對於高品質的通訊傳輸語音有非常好的預測表現，對照 BVCC 資料集上的結果，我們認為這不僅顯示生成語音與自然語音的自監督式表徵在分布特性上的不同，也再次暗示了分析訊號中的語調對於無參考客觀語音品質評估的重要性。此外，我們也同樣發現模型在在中間品質語音上的表現並不是太好，我們猜測其原因來自於人們對於何謂（普通）品質的語音，其考量的層面過於豐富，且因人而異，使得模型較難以同時學習這些行為與偏好。而在未來的研究中，分析上述兩現象將可能是提升無參考客觀語音品質評估模型表現的重點。

doi:10.6342/NTU202300029



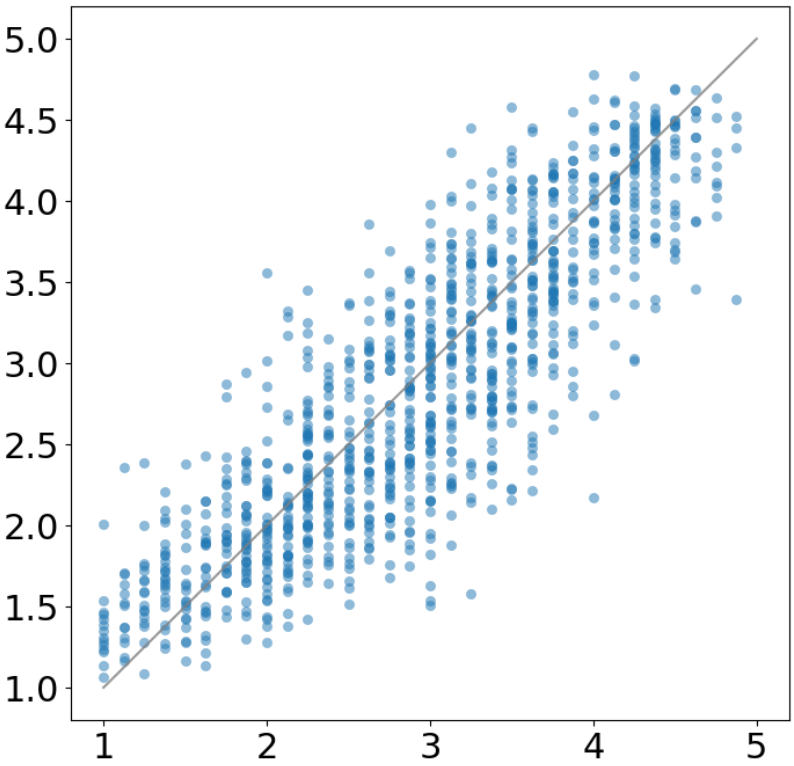


Figure 5.7: 在 BVCC 資料集（生成語音）中，語音品質對於模型預測準確度的影響。箱型圖的縱軸為平方差，而散佈圖的橫軸與縱軸分別代表實際分數與模型預測結果。

doi:10.6342/NTU202300029

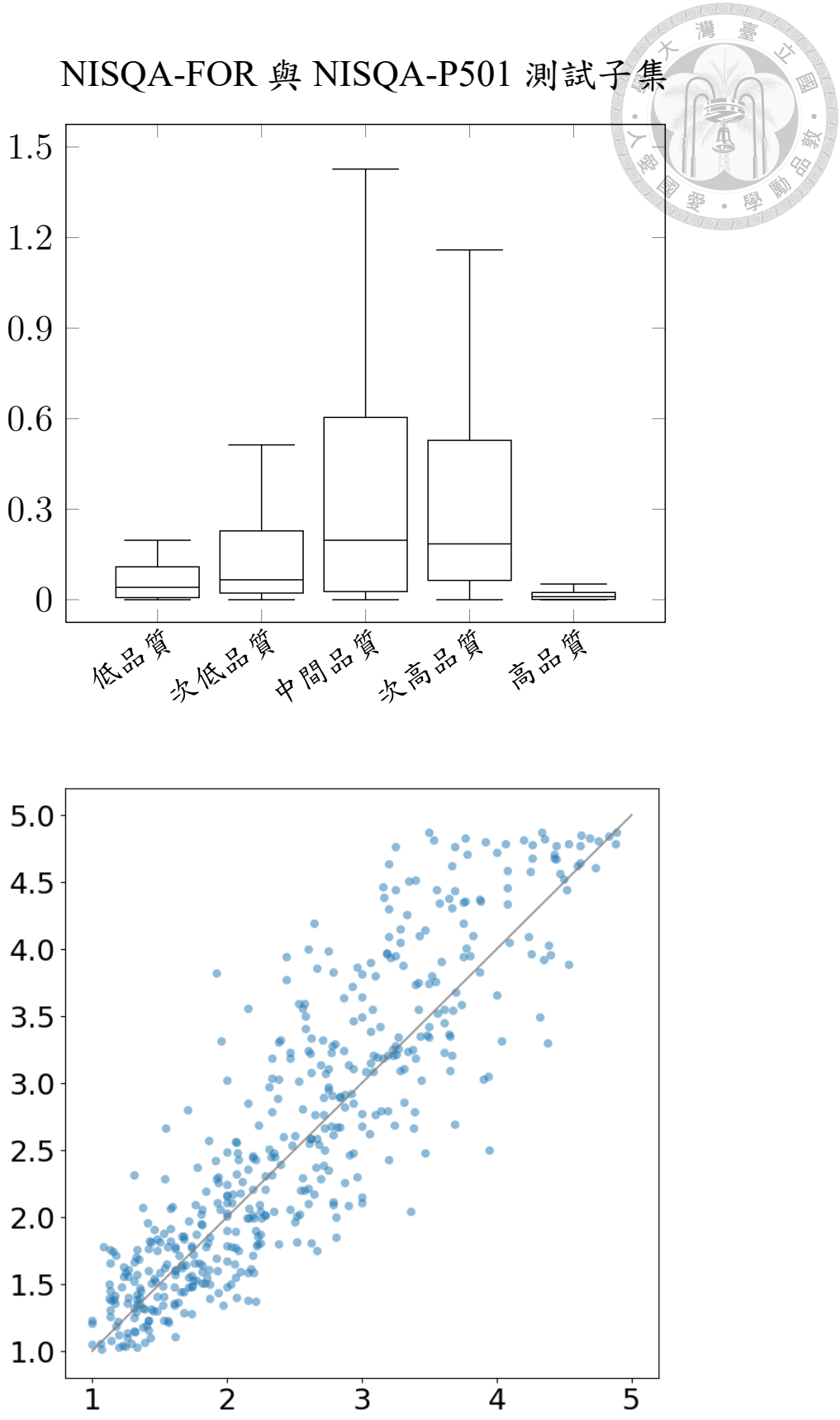


Figure 5.8: 在 NISQA 資料集（通訊傳輸的失真語音）中，語音品質對於模型預測準確度的影響。箱型圖的縱軸為平方差，而散佈圖的橫軸與縱軸分別代表實際分數與模型預測結果。

doi:10.6342/NTU202300029

5.10 本章總結

在本章中，我們提出一套基於 HuBERT 表徵的深層無參考客觀語音品質評估模型，並在實驗中發現，自監督式語音表徵中所含有的豐富訊息有助於提升無參考客觀語音品質評估的預測表現以及不同語言上的泛化能力，並全面超越當前表現最好的深層監督式無參考客觀語音品質評估模型 LDNet 以及 NISQAv2。此外，透過探測分析，我們發現人類對於生成語音與通訊傳輸語音具有某些共通的評分準則，而模型傾向於學習這些資訊。最後，我們發現模型對於中間品質的語音，以及高品質的生成語音具有最差的預測表現，在未來我們可以從此處切入，進一步提升模型的表現。

doi:10.6342/NTU202300029

第六章　結論與展望

本論文著重於無參考客觀語音品質評估，從分析自監督式語音表徵在語音品質評估上的可行性，接著提出一套基於微調自監督式語音表徵的深層無參考客觀語音品質評估模型，到最後對模型的表現進行泛化能力分析與探測分析，都是希望能為此研究做出貢獻及帶來新的發展方向。

6.1　研究貢獻與討論

本論文第 3 章中首先介紹基於深層學習的無參考客觀語音品質評估模型，詳細描述目前常見的模型架構，其中包含特徵提取模組、時間建模模組與預測模組，接著我們進一步介紹目前在無參考客觀語音品質評估模型所使用到的相關技術，並從中了解到，過去的方法多數使用傳統語音表徵作為輸入，其中隱含的資訊與結構性知識較為隱晦，可能會限制模型的表現。

本論文第 4 章中轉而分析自監督式語音表徵用於無參考客觀語音品質評估的可行性，透過還原梅爾時頻譜以及音素辨識實驗我們發現自監督式語音表徵同時含有豐富的聲學資訊與語言內容資訊。而透過繪製二維投影散佈圖與標準相關分析，我們進一步發現，自監督式語音表徵中含有相當高層次，與品質相關的資訊。足見自監督式語音表徵用於無參考客觀語音品質評估的潛力。而在所有的自監督式語音表徵中，以 HuBERT 的表現最為突出。而這也是首次在公開文獻中分析自監督式表徵運用於無參考客觀語音品質評估的可行性。

doi:10.6342/NTU202300029

在第 5 章中，我們基於 HuBERT 表徵，提出一套全新的深層無參考客觀語音品質評估模型，並在實驗中發現，自監督式語音表徵中所含有的豐富訊息有助於提升無參考客觀語音品質評估的預測表現以及不同語言上的泛化能力，並全面超越當前表現最好的深層監督式無參考客觀語音品質評估模型 LDNet 以及 NISQAv2。此外，透過探測分析，我們發現人類對於生成語音與通訊傳輸語音具有某些共通的評分準則，而模型傾向於學習這些資訊。最後，我們發現模型對於中間品質的語音、以及高品質的生成語音具有最差的預測表現，在未來我們可以從此處切入，進一步提升模型的表現。而這也是首次在公開文獻中將自監督式表徵用於無參考客觀語音品質評估之應用。

6.2 未來展望

在本論文的結果中，我們透過簡單的模型架構以及訓練方式，就能在預測表現上超越先前所有使用傳統表徵的所有方法，在未來，我們期望結合聆聽者相依網路（Listner-dependent Network）[5, 6]、多任務學習（Multitask Learning）[45–47] 等相關技術進一步提升模型的預測表現。除此之外，我們也希望利用預微調（pre-fine-tuning）[70]、多語言學習（Multi-lingual Learning）以及資料增強（Data Augmentation）等方法增進模型的泛化能力。最後，我們則希望能收集同時包含自然語音、生成語音以及失真語音的品質分數資料集，以利後續研究中，開發更全面（universal）的無參考客觀語音品質評估模型。

doi:10.6342/NTU202300029

參考文獻

[1] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), volume 2, pages 749–752. IEEE, 2001.

[2] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing, pages 4214–4217. IEEE, 2010.

[3] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, volume 1, pages 125–128. IEEE, 1993.

[4] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352, 2019.

[5] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. Mbnet: Mos prediction for synthesized speech with mean-bias network. In

doi:10.6342/NTU202300029

ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 391–395. IEEE, 2021.

[6] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 896–900. IEEE, 2022.

[7] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm. arXiv preprint arXiv:1808.05344, 2018.

[8] Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A Saurous, and D Sculley. Automos: Learning a non-intrusive assessor of naturalness-of-speech. arXiv preprint arXiv:1611.09207, 2016.

[9] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6419–6423. IEEE, 2020.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[11] Andy T Liu, Shang-Wen Li, and Hung-yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. arXiv preprint arXiv:2007.06028, 2020.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 2020.

doi:10.6342/NTU202300029

[13] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021.

[14] Pooyan Safari, Miquel India, and Javier Hernando. Self-attention encoding and pooling for speaker recognition. Proc. Interspeech 2020, pages 941–945, 2020.

[15] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133, 1943.

[16] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386, 1958.

[17] Paul John Werbos. The roots of backpropagation: from ordered derivatives to neural networks and political forecasting, volume 1. John Wiley & Sons, 1994.

[18] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[19] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. nature, 323(6088):533–536, 1986.

[20] Kiyoshi Kawaguchi. A multithreaded software model for backpropagation neural network applications. 2000.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

doi:10.6342/NTU202300029

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[25] Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. Analysis of CNN-based speech recognition system using raw speech as input. In Proc. Interspeech 2015, pages 11–15, 2015. doi: 10.21437/Interspeech.2015-3.

[26] Michael I. Jordan. Chapter 25 - serial order: A parallel distributed processing approach. In John W. Donahoe and Vivian Packard Dorsel, editors, Neural-Network Models of Cognition, volume 121 of Advances in Psychology, pages 471–495. North-Holland, 1997. doi: https://doi.org/10.1016/S0166-4115(97)80111-2. URL https://www.sciencedirect.com/science/article/pii/S0166411597801112.

[27] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

doi:10.6342/NTU202300029

[28] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2015.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30:5998–6008, 2017.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[31] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[32] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[33] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:1717–1728, 2021.

doi:10.6342/NTU202300029

[34] Songxiang Liu, Yuewen Cao, Na Hu, Dan Su, and Helen Meng. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation. In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2021.

[35] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. Snore sound classification using image-based deep spectrum features. In Interspeech 2017, pages 3512–3516. ISCA, August 2017.

[36] Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King. Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis.

[37] Yann LeCun. Self-supervised learning, 2020. URL https://vimeo.com/390347111.

[38] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051, 2021.

[39] Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, and Haohong Wang, editors. Multimedia Analysis, Processing and Communications, volume 346. 2011. ISBN 978-3-642-19550-1. doi: 10.1007/978-3-642-19551-8. URL https://doi.org/10.1007/978-3-642-19551-8.

[40] L. Malfait, J. Berger, and M. Kastner. P.563—the itu-t standard for single-ended speech quality assessment. IEEE Transactions on Audio, Speech, and Language Processing, 14(6):1924–1934, 2006. doi: 10.1109/TASL.2006.883177.

doi:10.6342/NTU202300029

[41] Meet H. Soni and Hemant A. Patil. Novel deep autoencoder features for non-intrusive speech quality assessment. In 2016 24th European Signal Processing Conference (EUSIPCO), pages 2315–2319, 2016. doi: 10.1109/EUSIPCO.2016.7760662.

[42] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. More for less: Non-intrusive speech quality assessment with limited annotations. In 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), pages 103–108. IEEE, 2021.

[43] Gabriel Mittag and Sebastian Möller. Deep learning based assessment of synthetic speech naturalness. arXiv preprint arXiv:2104.11673, 2021.

[44] Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Yist Y Lin, and Hung-yi Lee. Utilizing self-supervised representations for mos prediction. arXiv preprint arXiv:2104.03017, 2021.

[45] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494, 2021.

[46] Joan Serrà, Jordi Pons, and Santiago Pascual. Sesqa: Semi-supervised learning for speech quality assessment. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 381–385, 2021. doi: 10.1109/ICASSP39728.2021.9414052.

[47] Yeunju Choi, Youngmoon Jung, and Hoirin Kim. Neural mos prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification. In 2021 IEEE Spoken Language Technology Workshop (SLT), pages 462–469, 2021. doi: 10.1109/SLT48900.2021.9383533.

doi:10.6342/NTU202300029

[48] Ute Jekosch. Assigning Meaning to Sounds — Semiotics in the Context of Product-Sound Design, pages 193–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-27437-7. doi: 10.1007/3-540-27437-5_8. URL https://doi.org/10.1007/3-540-27437-5_8.

[49] Ute Jekosch. Voice and speech quality perception: assessment and evaluation. Springer Science & Business Media, 2006.

[50] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. Generalization ability of mos prediction networks. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8442–8446. IEEE, 2022.

[51] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, F. Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. The voice conversion challenge 2016. In INTERSPEECH, 2016.

[52] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. Analysis of the voice conversion challenge 2016 evaluation results. In Interspeech 2016, Interspeech, pages 1637–1641. International Speech Communication Association, September 2016. doi: 10.21437/Interspeech.2016-1331. URL http://www.interspeech2016.org/. Interspeech 2016 ; Conference date: 08-09-2016 Through 12-09-2016.

[53] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. arXiv preprint arXiv:1804.04262, 2018.

[54] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge

doi:10.6342/NTU202300029

2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527, 2020.

[55] Rohan Kumar Das, Tomi Kinnunen, Wen-Chin Huang, Zhen-Hua Ling, Junichi Yamagishi, Zhao Yi, Xiaohai Tian, and Tomoki Toda. Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions. In Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, pages 99–120, 2020. doi: 10.21437/VCC_BC.2020-15. URL http://dx.doi.org/10.21437/VCC_BC.2020-15.

[56] Simon King, Robert AJ Clark, Catherine Mayo, and Vasilis Karaiskos. The blizzard challenge 2008. 2008.

[57] Simon King and Vasilis Karaiskos. The blizzard challenge 2009. In The Blizzard Challenge 2009 Workshop, 2009.

[58] Simon King and Vasilis Karaiskos. The blizzard challenge 2010. 2010.

[59] Simon King and Vasilis Karaiskos. The blizzard challenge 2011. 2011.

[60] Simon King and Vasilis Karaiskos. The blizzard challenge 2013. 2014.

[61] Simon King and Vasilis Karaiskos. The blizzard challenge 2016. 2016.

[62] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015, 2018.

[63] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE

doi:10.6342/NTU202300029

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[64] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7669–7673. IEEE, 2020.

[65] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1314–1324, 2019.

[66] Karl Pearson. Note on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London Series I, 58:240–242, January 1895.

[67] C. Spearman. The proof and measurement of association between two things. The American Journal of Psychology, 15(1):72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

[68] M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. Biometrika, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

[69] Simon King Zhizheng Wu, Zhihang Xie. The blizzard challenge 2019. 2019.

[70] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964, 2020.

doi:10.6342/NTU202300029