國立臺灣大學電機資訊學院暨中央研究院
資料科學學位學程

碩士論文

Data Science Degree Program
College of Electrical Engineering and Computer Science
National Taiwan University and Academia Sinica

Master's Thesis

部分偽造語音中偽造片段偵測：從資料建構到模型設計

Detecting Spoofed Segments in Partially Spoofed Audio: From Dataset Construction to Model Design

郭恒成

Heng-Cheng Kuo

指導教授: 李宏毅 博士、曹昱 博士

Advisor: Hung-yi Lee, Ph.D., Yu Tsao, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202501466

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

部分偽造語音中偽造片段偵測：從資料建構到模型設計

Detecting Spoofed Segments in Partially Spoofed Audio:
From Dataset Construction to Model Design

本論文係 郭恒成（R11946023）在國立臺灣大學資料科學學位學程完成之碩士學位論文，於民國 114 年 06 月 23 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 Data Science Degree Program on 23 June 2025 have examined a Master's thesis entitled above presented by HENG-CHENG KUO (R11946023) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李宏毅	曹昱	王新民
（指導教授 Advisor）	（指導教授 Advisor）	
賴穎暉	蔡宗翰	李琳山

學位學程主管 Director: 謝育昀

中文摘要

本論文針對部分偽造語音的生成與偵測展開研究。我們提出了一種基於流匹配的非自回歸語音編輯模型 VoiceNoNG，該模型以神經編解碼器 Descript Audio Codec 輸出的量化前表徵作為符元化表徵，並同時條件於文字稿與語音上下文，實現高品質、低延遲的語音補全。基於 VoiceNoNG，我們進一步建構了「語音補全編輯資料集」，用以克服傳統「半真實語音偵測資料集」中因剪貼式流程導致的偽造片段邊界訊號不連續問題，提供更貼近實務場景的部分偽造語音評測基準。

在生成端實驗中，VoiceNoNG 相較於既有的流匹配與自回歸模型，在字詞錯誤率、訊噪失真比與主觀聆聽測試等多項指標上均取得顯著提升；在偵測端，我們以四種最先進的防偽偵測模型進行多場景（真實-補全／真實-剪貼／重合-補全）、跨場景與跨編輯模型（VoiceCraft）的實驗，驗證「偽造片段邊界不連續」及「編解碼處理差異」成為模型容易採取捷徑學習的關鍵因素。唯有在統一編解碼處理的「重合-補全」場景下訓練的模型，才能真正聚焦於偽造片段的本質特徵。而跨編輯模型實驗則顯示現有的防偽偵測模型在測試資料與訓練資料存在生成演算法與編解碼器不匹配時，仍難以實現普適化。

關鍵字：部分偽造語音、流匹配、神經編解碼器、語音編輯、偽造語音偵測

doi:10.6342/NTU202501466

英文摘要

This thesis investigates the generation and detection of partially spoofed audio. We propose VoiceNoNG, a non-autoregressive speech editing model based on flow matching. VoiceNoNG leverages pre-quantization representations extracted from a neural codec, Descript Audio Codec, as tokenized representations and is conditioned on both the transcript and surrounding audio context, enabling high-quality and low-latency speech infilling. Built on VoiceNoNG, we further construct the Speech Infilling Edit dataset, which overcomes the signal discontinuity at spoofed segment boundaries caused by the cut-and-paste process in the traditional Half-Truth Dataset, thereby providing a more realistic benchmark for evaluating partially spoofed audio detection.

In generation experiments, VoiceNoNG outperforms existing flow-matching and autoregressive models across multiple evaluation metrics, including word error rate, signal-to-noise distortion ratio, and subjective listening tests. In detection experiments, we evaluate four state-of-the-art anti-spoofing detectors under multiple scenarios (real-infill, real-

doi:10.6342/NTU202501466

paste, resynthesis-infill), as well as cross-scenario and cross-generator (VoiceCraft) settings. Results confirm that discontinuity at spoofed segment boundaries and differences in codec processing are critical cues that lead detectors to rely on shortcut learning. Only detectors trained in the "resynthesis-infill" scenario—where all frames uniformly undergo the neural codec—can truly focus on the intrinsic features of spoofed segments. Cross-generator experiments further demonstrate that existing anti-spoofing detectors still struggle to generalize under mismatches in synthesis algorithm and codec processing between training and test data.

Keywords: Partially Spoofed Audio, Flow Matching, Neural Codec, Speech Editing, Spoof Detection

doi:10.6342/NTU202501466

目次

doi:10.6342/NTU202501466

doi:10.6342/NTU202501466

doi:10.6342/NTU202501466

圖次

doi:10.6342/NTU202501466

表次

doi:10.6342/NTU202501466

第一章　導論

1.1　研究動機

近年來，隨著深度神經網路（Neural Network）的蓬勃發展，語音合成（Speech Synthesis）與語音克隆（Voice Cloning）技術突飛猛進，並廣泛應用於影片配音、虛擬助理與口語語言模型（Spoken Language Model）等領域，顯著提升了使用者體驗。然而，這些高品質的合成語音同時帶來潛在風險：使用者愈來愈難以分辨真實語音與合成語音，進而對多種以「聲音可信度」為基礎的安全與信任機制造成威脅，其中之一便是生物特徵識別系統 [1]。過去，聲紋因其個體特異性與難以複製性而被視為高度安全，如今已能被攻擊者合成，用以欺騙基於聲紋辨識的驗證系統，使其誤將攻擊者視為合法使用者而取得受保護的資訊或貴重資產。因此，如何有效偵測並防範這類威脅，已成為亟待解決的重要課題 [2, 3]。

為了降低偽造語音帶來的安全風險，山岸順一（Junichi Yamagishi）等人自 2015 年起發起「自動說話人驗證防偽挑戰」（Automatic Speaker Verification Spoofing and Countermeasures Challenge）[4–7]。該系列挑戰旨在推動偽造語音偵測技術的發展，同時評估防偽對策與自動說話人驗證系統的協同效能。「自動說話人驗證防偽挑戰」主要分為兩個賽道：邏輯通路（Logical Access）與物理通路（Physical Access）。邏輯通路假設攻擊者直接輸入由語音合成或語音轉換（Voice Conversion）模型生成的語音，著重衡量系統在無物理失真條件下抵禦生成式攻擊

doi:10.6342/NTU202501466

的能力；物理通路則假設攻擊者掌握目標說話者的真實錄音，先用喇叭播放，再由麥克風收錄，將空間殘響與裝置通道響應等物理因素納入考量，從而評估系統在面對重播攻擊時的穩健性。競賽要求參賽者同時計算防偽模型與自動說話人驗證模型的聯合錯誤率，藉此全面評估整體語音驗證流程的抗攻擊能力。該挑戰自舉辦以來吸引了眾多國際研究團隊參與，並催生出多種高效能的防偽模型以對抗上述兩類攻擊 [8–10]。同時，圍繞防偽模型與自動說話人驗證模型的對抗式攻擊（Adversarial Attack）策略也已獲得廣泛而深入的探討 [11–13]。

在「自動說話人驗證防偽挑戰」奠定的基礎上，易江燕（Jiangyan Yi）等人發起「語音深度合成偵測挑戰」（Audio Deep Synthesis Detection Challenge），以因應真實場景中更加複雜的攻擊情境 [14, 15]。該挑戰補足了「自動說話人驗證防偽挑戰」未涵蓋的多種場景，將各種背景噪聲干擾下的低品質偽造語音，以及來自更先進語音合成與語音轉換模型的對抗式攻擊一併納入評估。值得注意的是，主辦方首次提出「部分偽造語音偵測」賽道，其攻擊方式僅替換真實語音中少數關鍵詞，卻足以顛覆整句語意，因而隱蔽性極高、風險巨大。「語音深度合成偵測挑戰」成為全球第一個針對此新型攻擊提出解決方案的競賽，而本論文正是以此議題為核心。

1.2 研究方向

在「語音深度合成偵測挑戰」結束後，我們進一步意識到部分偽造語音偵測的重要性，也發現該挑戰所使用的「半真實語音偵測資料集」（Half-Truth Audio Detection Dataset）存在侷限 [16]。「半真實語音偵測資料集」採用剪貼（Cut-and-Paste）式編輯：系統先利用文字轉語音（Text-to-Speech）依照修改後的文字稿生成完整的合成語音，再將對應的偽造片段剪下並貼回原始語音。此流程

doi:10.6342/NTU202501466

在偽造片段邊界不可避免地引入訊號不連續，導致許多頂尖偵測模型傾向於依賴此類邊界不連續進行捷徑學習（Shortcut Learning），反而難以偵測偽造片段與原始語音無縫融合的語音編輯。為克服此侷限，我們提出 VoiceNoNG ——一種同時條件在目標文字稿與周遭語音上下文的語音編輯模型，能直接在時間域中補全被遮蔽的偽造片段，避免邊界不連續——並建立「語音補全編輯資料集」（Speech Infilling Edit Dataset）[17, 18]。「語音補全編輯資料集」為研究真實且難檢測的部分偽造語音提供了更貼近實務場景的評測基準。

本論文首先回顧「半真實語音偵測資料集」的生成流程，以及 VoiceBox [19] 與 VoiceCraft [20] 兩種演算法迥異的上下文感知語音編輯模型。我們藉此說明「半真實語音偵測資料集」為何會在偽造片段邊界引入訊號不連續，並指出 VoiceBox ／ VoiceCraft 雖能同時條件於文字稿與語音上下文，從而避免訊號不連續，但仍分別存在抹除背景聲與注意力幻覺的侷限。為克服上述問題，我們提出 VoiceNoNG，在保留背景聲與語義一致性的同時，實現高品質、低延遲的語音編輯。隨後，我們重現「語音深度合成偵測挑戰」中表現最佳的四種防偽偵測模型，並在自建的「語音補全編緝資料集」上評估其效能。實驗結果顯示，相較於「半真實語音偵測資料集」，大多數模型在「語音補全編輯資料集」上的性能均明顯下降，驗證了在缺乏邊界不連續作為學習捷徑的的情況下，現有模型仍難以準確偵測與定位偽造片段。而當以自監督語音基石模型（Self-Supervised Speech Foundation Model）提取的高階表徵作為輸入時，偵測準確率可顯著提升。

進一步地，我們提出一個新穎的評測場景：將經過神經編解碼器重合成的語音獨立分類，不再與偽造語音劃為同類。此定義基於兩項考量：(1) 隨著神經編解碼器技術成熟，未來其壓縮格式可能如 MP3 般普及，大量合法語音勢必經過神經編解碼器處理；(2) 除剪貼式編輯造成的邊界不連續外，「經編解碼」與「未編解碼」片段之間的聲學差異亦會形成另一條捷徑，使模型無法真正學習偽造片段的

doi:10.6342/NTU202501466

本質特徵。藉由納入此場景，我們期望推動開發在更貼近實務條件下仍具魯棒性，且對捷徑訊號不敏感的部分偽造語音偵測方法。

1.3 研究貢獻

本論文的主要貢獻包含：

- 提出語音編輯模型 VoiceNoNG：將流匹配生成框架與神經編解碼器 Descript Audio Codec 相結合，設計出可同時條件於文字稿與語音上下文的非自回歸語音編輯模型，並證明其在字詞錯誤率、訊噪失真比與主觀聆聽測試等多項指標上均優於既有流匹配與自回歸方法。
- 構建「語音補全編輯資料集」：基於 VoiceNoNG 生成不含邊界不連續的部分偽造語音樣本，為後續防偽偵測研究提供更貼近實務場景的基準。
- 揭示偵測模型捷徑學習問題：透過多場景、跨場景與跨編輯模型的實驗，系統性地驗證「偽造片段邊界不連續」及「編解碼處理差異」皆成為現有防偽偵測模型易於利用的學習捷徑，並指出僅有移除這些捷徑後訓練的模型才能真正聚焦於偽造片段的本質特徵。
- 區分重合成語音與偽造語音：首次主張將經神經編解碼器重合成的合法語音獨立於「真實／偽造」之外，並通過實驗證明，這一新類別不僅更符合未來壓縮傳輸與神經編解碼器技術普及的趨勢，也能幫助開發對編解碼差異不敏感的通用偽造語音偵測方法。
- 提出強化策略與未來方向：探討三元分類、加噪增強等方法以提升防偽偵測模型對跨編解碼器與跨編輯模型的魯棒性，並指明「跨編輯模型泛化能力」將是未來防偽偵測技術成熟度的重要衡量指標。

doi:10.6342/NTU202501466

1.4 章節安排

本論文的章節安排如下：

- 第二章 背景知識
- 第三章 語音編輯模型 VoiceNoNG
- 第四章 語音補全編輯資料集
- 第五章 重合成語音的影響
- 第六章 結論與展望

doi:10.6342/NTU202501466

第二章　背景知識

2.1　半真實語音偵測資料集

「半真實語音偵測資料集」是專為部分偽造語音任務設計，旨在評估與分析偽造片段的偵測與定位方法，並在「語音深度合成偵測挑戰」中提供給參賽隊伍訓練防偽偵測模型。此資料集以 AISHELL-3 中文語料庫 [21] 為基礎，透過僅替換一句話中的少數詞彙來生成偽造片段。部分偽造語音的生成流程分為五個步驟，如圖 2.1 所示：(1) 依預定義的文本替換策略，在真實語音文字稿中選定並替換關鍵詞；(2) 利用端到端文字轉語音模型生成完整的偽造語音；(3) 擷取替換後關鍵詞所對應的偽造片段；(4) 正規化真實語音與偽造片段音量；(5) 以偽造片段替換真實語音中關鍵詞對應的時間區段，得到部分偽造語音。

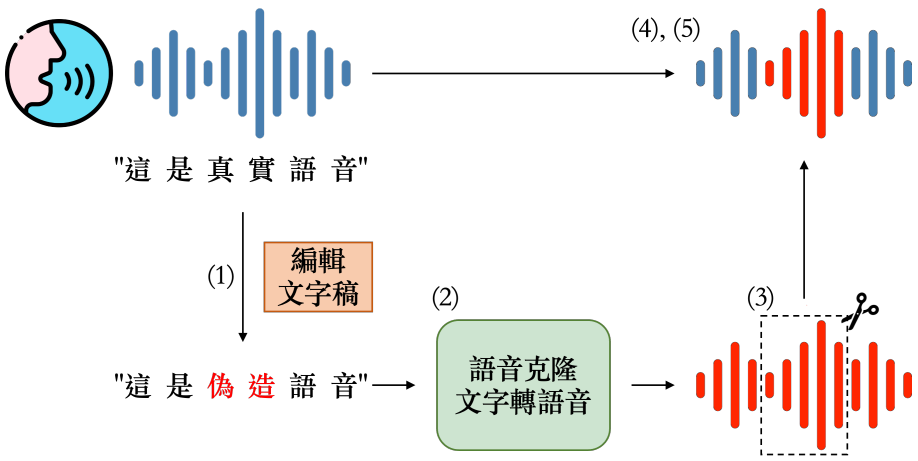


圖 2.1: 半真實語音偵測資料集生成流程

doi:10.6342/NTU202501466

文本替換策略透過修改真實語音文字稿中的關鍵詞，改變整句語意，主要分為：(1) 隨機替換命名實體：在句中將人名、地名、組織名或時間替換為同類別的其他實體；(2) 替換態度詞為反義詞：選擇能改變語句立場或情感的關鍵詞，以其反義詞取代。資料集建立者使用結巴分詞（Jieba）[22] 進行中文斷詞與命名實體辨識，每句僅替換一次。表 2.1 列出各實體類別的「出現次數」與「替換池大小」，反義詞配對則收錄 181 組。

表 2.1: 半真實語音偵測資料集各實體類別統計資料

	人名	地名	組織名	時間	總數
出現次數	4232	2731	2401	1329	10693
替換池大小	5154	3307	3270	1381	13112

完整偽造語音生成則使用 AISHELL-3 訓練集訓練基於全域風格向量（Global Style Token）與韻律轉移（Prosody Transfer）的端到端文字轉換語音模型 Tacotron [23, 24]。Tacotron 先生成符合目標說話者風格（音調、語速等）的時頻圖，再透過神經聲碼器（Vocoder）LPCNet [25] 轉為語音。接著，以語音辨識系統對原始語音與完整偽造語音進行強制對齊（Forced Alignment），取得字級時間戳記，以便自動定位並替換偽造片段，完成部分偽造語音的生成。

2.2 神經編解碼器 EnCodec

根據思科系統公司（Cisco Systems, Inc.）2021 年的報告 [26]，串流音訊與影像已占全球網路流量 82 %，且預估未來五年仍將維持雙位數的年增長。其中，音訊流量因多軌、高解析度而迅速膨脹，迫使雲端服務在成本與使用者體驗之間尋求最佳平衡；同時，智慧眼鏡與車載系統等邊緣裝置受限於功耗與頻寬，亟須低位元率（Bitrate）、低延遲且低計算量的壓縮演算法。

在有損壓縮中，核心目標是在盡可能降低位元率的同時，依據與人類聽覺高

doi:10.6342/NTU202501466

度相關的評量指標，將失真降至最低。傳統音訊編解碼器（Codec）以精心設計的訊號處理流程結合編碼器（Encoder）與解碼器（Decoder）：先對輸入訊號施行變換以去除冗餘，再輸出精簡的位元流（Bitstream），並對較不影響聽覺的頻段做品質取捨 [27]。然而，傳統編解碼器已逼近效能極限，進一步降低位元率往往伴隨明顯失真。

近年來，研究轉向以神經編解碼器取代手工設計的變換，透過端到端學習自動尋找最佳表徵，展現了在超低位元率下仍能維持高音質的潛力，典型架構由編碼器 E、量化器 Q 與解碼器 D 三部分組成。輸入語音波形或時頻圖 x，先經編碼器映射為連續潛在向量

$$z = E(x)$$

為在傳輸或儲存時生成可離散化的位元流，量化器採用 K 層殘差向量量化（Residual Vector Quantization）。第 k 層量化器自碼簿（Codebook）$B_k = \{b_{k,1}, ..., b_{k,M}\}$ 中選出最接近殘差 r_{k-1} 的離散索引

$$i_k = \text{argmin}_j \left\| r_{k-1} - b_{k,j} \right\|_2, \quad r_k = r_{k-1} - b_{k,i_k}$$

其中 $r_0 = z$。累計各層碼字（Codeword）得到量化潛在向量

$$q = \sum_{k=1}^{K} b_{k,i_k}$$

最後將量化潛在向量輸入解碼器還原為重建訊號

$$\hat{x} = D(q)$$

模型以失真量 R 與平均位元率 S 為目標聯合優化，其中 λ 控制壓縮比與音質間的

doi:10.6342/NTU202501466

權衡

$$\mathcal{L} = R(x, \hat{x}) + \lambda S(q, \{i_1, ..., i_k\})$$

另一方面，生成式人工智慧也將語音／音樂符元（Token）納入整體的生成流程。神經編解碼器會把連續語音波形映射為每秒僅數十個離散符元，使語音序列長度與文字相當，從而可直接沿用大型語言模型（Large Language Model）的技術進行生成與編輯。在此基礎上，衍生出了神經編解碼器語言模型（Neural Codec Language Model）系列的語音生成模型，如 VALL-E [28] 與 VoiceCraft。這些模型在推論時可同時條件於文字稿與參考語音：文字稿決定內容，參考語音提示語者特徵與韻律，從而在語音編輯及零樣本（Zero-Shot）文字轉語音等任務上取得最先進表現。

EnCodec [29] 是目前最常用的神經編解碼器之一，也是本文後續比較的基線 VoiceCraft 所採用的符元化器（Tokenizer）。如圖 2.2 所示，它延續了「編碼器–量化器–解碼器」的簡潔架構：首先，編碼器以一層核大小 7、通道數 32 的一維卷積層（1D Convolutional Layer）起始，隨後串接四個重複的卷積區塊。每個區塊先經由兩層核大小 3 的卷積層與跳接（Skip Connection）組成的殘差單元抽取特徵，再依序利用步幅 $S \in \{2, 4, 5, 8\}$、核大小 $2S$ 的下採樣（Down-sampling）卷積層，同步完成時間下採樣並將通道數加倍。卷積主幹之後，編碼器接上兩層長短期記憶網路（Long Short-Term Memory）以捕捉長距時序依賴，最後透過核大小 7、輸出通道數 512 的一維卷積層，將隱層嵌入向量（Hidden Embedding）投射到量化器的輸入空間。整個卷積部分均採用 ELU 非線性函數並配合層正規化（Layer Normalization），以穩定訓練並提升表徵能力。

EnCodec 的量化器採用殘差向量量化將編碼器輸出的連續潛在向量離散化：首先，將連續潛在向量映射到第一個碼簿中，距離最近的碼字；接著計算殘差，

doi:10.6342/NTU202501466

並依序使用第二個、第三個……第 K 個碼簿對殘差再次量化，以逐步細化表徵。透過在訓練過程中隨機選擇不同數目的碼簿，單一模型即可支援多種頻寬設定。具體而言，24 千赫茲（kHz）版本最多使用 32 個碼簿，而 48 千赫茲版本最多使用 16 個，每個碼簿包含 1024 個碼字（相當於 10 位元）。訓練時隨機以 4 的倍數選取碼簿數量，24 千赫茲音訊下，便可對應 1.5、3、6、12 及 24 千位元每秒的多檔位元率。

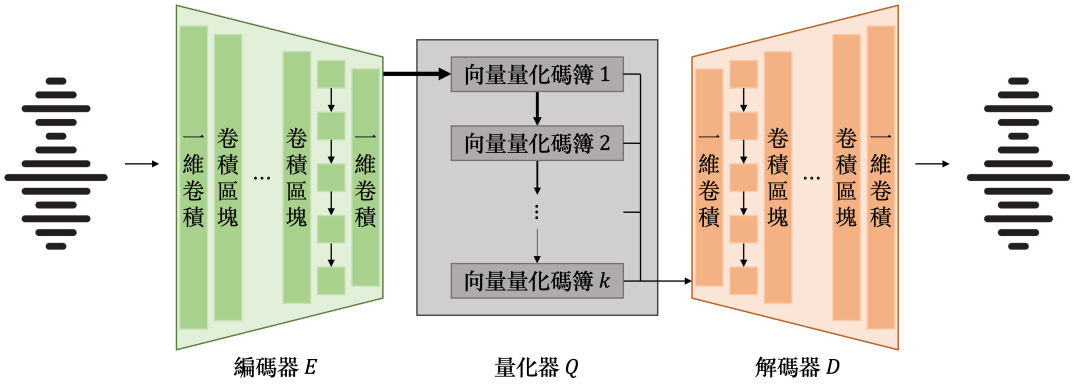


圖 2.2: 神經編解碼器 EnCodec 模型架構

2.3 自回歸的語音編輯模型 VoiceCraft

VoiceCraft 是一個語音編輯模型，亦可視為「符元補全神經編解碼器語言模型」(Token Infilling Neural Codec Language Model)。它在語音編輯與零樣本文字轉語音任務上都取得了最先進的成果。模型架構採用僅含解碼器的 Transformer [30]，並透過預測自回歸序列的方式訓練。其關鍵創新在於透過「因果遮蔽」與「延遲堆疊」對輸入符元重新排列，使模型在保持自回歸機制的同時，仍能同時利用遮蔽片段前後的雙向語音上下文，從而在續接或補全情境下都能生成語義連貫且音質自然的語音。

因果遮蔽的流程如圖 2.3.1 所示。首先，VoiceCraft 的輸入為連續語音波形，先經過 EnCodec 的編碼器將語音量化為大小為 $T \times K$ 的矩陣 X，其中 T 為時間

doi:10.6342/NTU202501466

幀數，K 為殘差向量量化的碼簿數。令 $X = (X_1, ..., X_T)$，其中 X_t 是長度為 K 的矩陣，代表在時間步 t 由各碼簿取出的碼字。在訓練時，會隨機選定連續的一段符元 $(X_{t_0}, ..., X_{t_1})$ 作為遮蔽區段，然後讓模型在給定其餘未遮蔽符元的條件下，自回歸地預測這些被遮蔽的符元。但若 $t_1 < T$，在直接遮蔽的情況下，模型無法條件於「未來」的符元。因此，為保持因果結構，VoiceCraft 先將被遮蔽區段用遮蔽符元 $\langle M \rangle$ 取代，並且移至序列末端。同時，在被遮蔽區段起始以及結束位置會分別插入遮蔽符元 $\langle M \rangle$ 以及區段結束（End-of-Span）符元 $\langle EOS \rangle$，在原句結束位置亦會插入語句結束（End-of-Utterance）符元 $\langle EOU \rangle$。以 $T = 5$ 的符元序列為例，假設 $X = (X_1, X_2, ..., X_5)$，欲遮蔽 X_2 到 X_3，重新排列後得到

$$Y = (Y_1; \langle M \rangle; Y_2; \langle EOU \rangle; \langle M \rangle; Y_3; \langle EOS \rangle)$$

其中，$Y_1 = (X_1)$、$Y_2 = (X_4, X_5)$ 為未遮蔽區段，$Y_3 = (X_2, X_3)$ 為遮蔽區段。經此因果遮蔽處理後，模型輸入 Y 便同時包含遮蔽區段前後的上下文，而預測仍遵循自回歸順序。

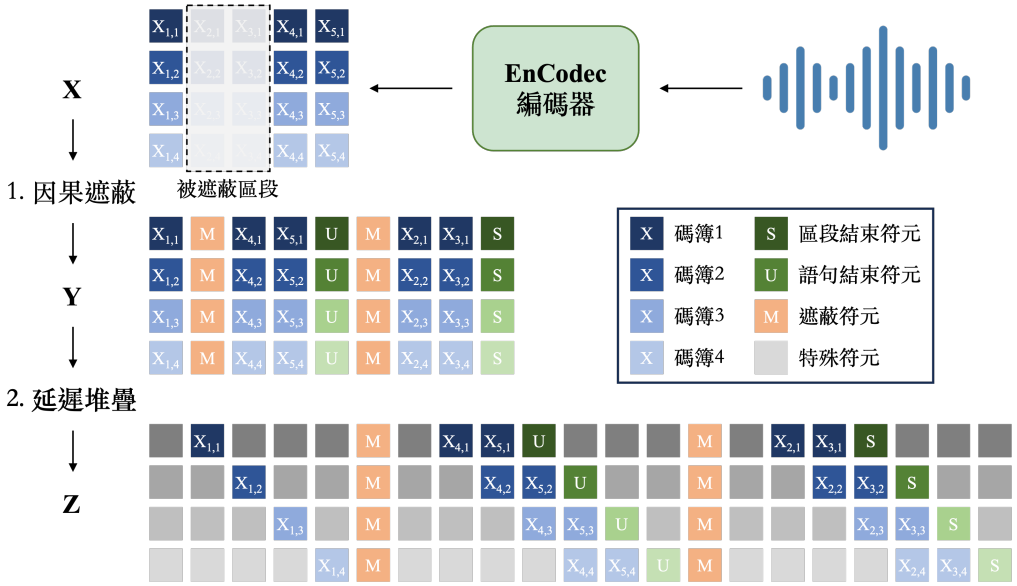


圖 2.3: 因果遮蔽與延遲堆疊

在完成因果遮蔽並重新排列取得矩陣 Y 之後，序列中每一個時間步都是由 K

doi:10.6342/NTU202501466

個殘差向量量化碼字組成。柯佩（Copet）等人發現：對這種堆疊式殘差向量量化符元做自回歸生成時，若能讓第 t 步、第 k 個碼簿的預測，能夠條件在同一步第 $k-1$ 個碼簿的輸出，生成品質會顯著提升 [31]。VoiceCraft 因此採用「延遲堆疊」的機制，以下說明具體做法（圖 2.3.2）。假設 Y' 為某個大小為 $L \times K$ 的區段，延遲堆疊後得到序列 $Z = (Z_0, Z_1, ..., Z_{L+K-1})$，其中對任何 $t \in [0, L+K-1]$，向量 Z_t 定義為

$$Z_t = (Y'_{t,1}, Y'_{t-1,2}, ..., Y'_{t-K+1,K})$$

亦即，Z_t 的第 k 個分量取自 Y' 中「時間 $t-k+1$、碼簿 k」的符元。換言之，同一時間步的高階碼簿 k 會延遲 $k-1$ 個時間步才被讀取，如此模型在預測 (t, k) 的碼字時，就能看到 $(t, k-1)$ 的輸出。若索引超出範圍，即 $t-k+1 < 1$ 或 $t-k+1 > L$，矩陣中並不存在對應符元，則以一個可學習的特殊符元 $\langle EMPTY \rangle$ 代替。公式寫成

$$Y'_{t-k+1,k} = \langle EMPTY \rangle, \quad \forall t-k+1 \notin [1, L]$$

值得一提的是，這使得開頭的 Z_0 皆由 $\langle EMPTY \rangle$ 填充，相當於自然引入一個區段起始標記，以提示模型開始生成。另需注意，遮蔽標記 $\langle M \rangle$ 並不屬於任何區段，其位置在延遲堆疊過程中保持不變，不參與上述重排。

此外，VoiceCraft 自回歸地建模 Z 時，需條件在音素化文字稿（Phonemized Transcript）W。因此，Transformer 的完整輸入組成為 $[W; Z]$，其中「;」表示將兩者串接。在語音編輯任務的推論中，假設有一段語音 X 及其文字稿 W，目標是只修改 X 中與改動相關的片段，使其內容符合目標文字稿 W'。這裡 W' 是 W 的編輯後的版本，部分詞語可能被插入、替換或刪除。整體流程與模型的訓練任務幾乎相同，但有兩點差異：（1）訓練時模型看到的是原始文字稿 W，推論時則提供已修改的 W'；（2）訓練階段隨機選擇遮蔽片段，推論階段則先比對 W 與 W'，

doi:10.6342/NTU202501466

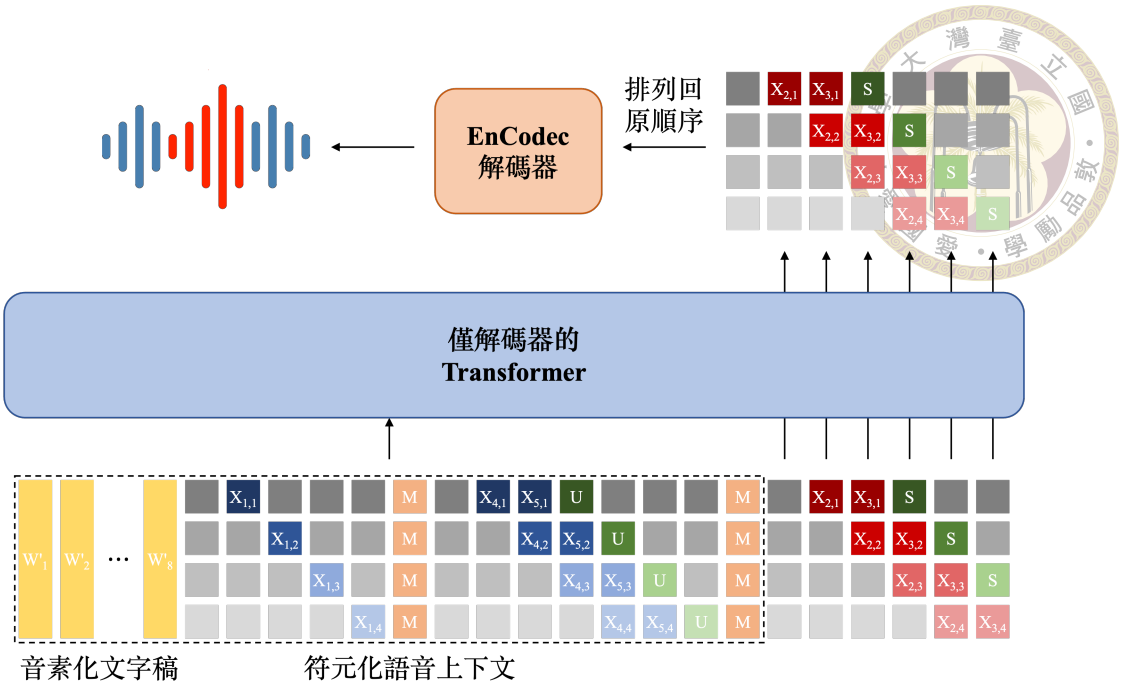


圖 2.4: VoiceCraft 推論流程

找出需要修改的詞，利用詞級強制對齊標定對應的符元區段並加以遮蔽。如圖 2.4 所示，推論時模型收到的輸入序列包含「編輯後文字-未遮蔽片段」，然後以自回歸方式填補遮蔽片段。生成的符元依序嵌回原位置，再交由 EnCodec 的解碼器還原為完整語音波形。而在零樣本文字轉語音任務，可視作在原句末端做一次「插入式編輯」。具體做法是：輸入一段參考語音及其文字轉錄稿，並附上欲生成的目標文字稿，將「參考文字-目標文字-參考語音」串接後送入模型。模型隨後自回歸地生成目標文字稿對應的符元序列，即可得到與參考語音聲線一致的全新語音。

2.4 流匹配的語音編輯模型 VoiceBox

VoiceBox 是第一個採用流匹配（Flow Matching）[32] 生成框架，達到跨任務泛化的語音編輯模型，其推論速度可達基於自回歸模型的 10～20 倍。在訓練階段，VoiceBox 把文字引導（Text-Guided）的語音補全作為核心任務，給定文字稿與待補全的語音上下文，模型學習在雙向條件下生成缺失語音。推論時，這一能力自然遷移到多種語音任務（續講、編輯、去噪、風格轉換）。

doi:10.6342/NTU202501466

流匹配是一種單步（或極少步）求解常微分方程（Ordinary Differential Equation）的生成框架。其核心概念是對時間區間 $[0,1]$ 上的每一點參數化一個隨時間變化的向量場

$$v_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$$

並以此建構出一條流（Flow），把樣本從高斯的先驗分佈 p_0 逐步推送至真實資料的目標分佈 q。向量場與其流之間的關係可透過下列常微分方程來定義：

$$\frac{d}{dt}x_t = v_t(x_t;\theta); \quad x_0 \sim N(\mathbf{0},\mathbf{I})$$

在理想情況下，終點 x_1 會滿足 $x_1 \sim p_1 \approx q$。為使參數化向量場 v_t 收斂到真實向量場 u_t，可以透過最小化目標函式

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}\left\|u_t(x_t) - v_t(x_t;\theta)\right\|^2$$

來進行訓練。而對於如何選定這個流，利普曼（Lipman）等人提出了最佳傳輸（Optimal Transport）路徑作為流的構造方式 [33]。在此設定下，對於任一對樣本 $(x_0, x) \sim p_0 \times q$

$$u_t(x_t) = x - x_0, \quad x_t = (1-t)x_0 + tx$$

即樣本沿固定方向、等速地從 x_0 往 x 移動，路徑上每一點透過線性插值即可完成轉換。實務上，為降低神經網路的擬合難度，VoiceBox 會保留一小比例高斯雜訊，將路徑改寫為

$$u_t(x_t) = x - (1-\sigma)x_0, \quad x_t = (1-(1-\sigma)t)x_0 + tx$$

在終點 x_1 仍殘留 σx_0 的微量噪聲，此舉可顯著平滑向量場，降低神經網路的擬合難度，同時不破壞其最佳傳輸特性。

doi:10.6342/NTU202501466

VoiceBox 為滿足精細對齊的需求，將整體系統拆分為拆分為「語音模型」與「時長模型」兩個子模型。設 $x = (x^1, x^2, ..., x^N)$ 為長度 N 的語音序列；$y = (y^1, y^2, ..., y^M)$ 為長度 M 的音素序列；$l = (l^1, l^2, ..., l^M)$ 為每個音素對應的幀數，其中 l^i 表示音素 y^i 佔用多少語音幀，並滿足 $\sum_{j=1}^{M} l^j = N$。定義 $z = \text{rep}(y, l) = (z^1, z^2, ..., z^N)$ 為逐幀的音素標籤，方法是將每個 y^i 重複 l^i 次。給定一組 (x, y)，可利用語音辨識模型做強制對齊來估計時長 l 與逐幀音素 z。

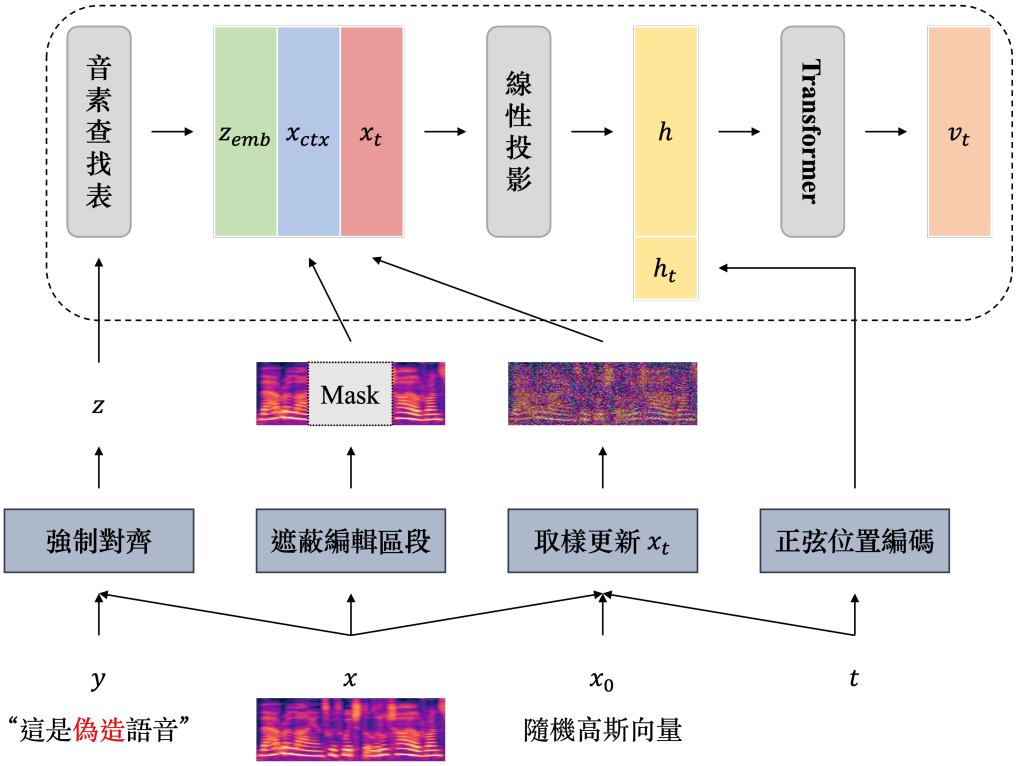


圖 2.5: VoiceBox 前向傳播流程

VoiceBox 的語音模型也是採用僅含解碼器的 Transformer 架構，其流匹配的單步前向傳播流程如圖 2.5 所示。首先，原始語音 x 為 80 維的對數梅爾頻譜圖，模型的主要輸入包含 x_t、x_{ctx}、z_{emb} 和 h_t。x_t 在 $t = 0$ 時，由與 x 同樣大小的高斯向量隨機初始化，之後每一步皆依據模型預測的向量場 v_t 更新。x_{ctx} 為被遮蔽的對數梅爾頻譜圖，被遮蔽的語音幀會設為零向量，未被遮蔽的語音幀則保持跟 x 相同。音素序列 z 是以音素類別做獨熱編碼（One-Hot Encoding），可根據可學習式查找表（Learnable Lookup Table）得到音素的嵌入向量 z_{emb}。接著，將

doi:10.6342/NTU202501466

x_t、x_{ctx}、z_{emb} 串接，經線性投影為潛在向量序列 h。時間步 t 經正弦位置編碼（Sinusoidal Positional Encoding）轉換為 h_t，並插入序列 h 的起始位置，得到 h'，送入 Transformer，以求得條件向量場 $v_t(x_t, x_{ctx}, z; \theta)$。而從向量場取樣更新 x_t 方式為

$$x_{t+\frac{1}{N}} = x_t + \frac{v_t(x_t, x_{ctx}, z; \theta)}{N}$$

其中，N 為向量場評估次數。當 N 越大，x_1 對真實語音 x 的近似越精確，但運算量亦隨之增加，使用者可據此在速度與品質間調節。實驗顯示，VoiceBox 在 $N < 10$ 時，即可輸出極高品質的語音，推論延遲遠低於自回歸語音編輯模型。時長模型則跟語音模型架構完全相同，只是將（x_t, x_{ctx}, z_{emb}）換成（l_t, l_{ctx}, y_{emb}）。

以語音編輯為例，先以編輯後的音素序列 $\hat{y}$ 取代原始音素 y，並遮蔽編輯音素對應的的時長區段。接著，用時長模型預測新時長 $\hat{l}$，以此建構新的序列 $\hat{x}_{ctx}$。最後，再令語音模型根據 $(\hat{x}_t, \hat{x}_{ctx}, \hat{z}_{emb})$ 產生向量場並更新，來合成最終的編輯語音。而零樣本文字轉語音，則跟 VoiceCraft 一樣，可視作在原句末端做一次語音編輯。

2.5 本章總結

本章首先回顧了「半真實語音偵測資料集」與其剪貼式生成流程，該流程先以端到端文字轉語音模型生成完整偽造語音，再透過強制對齊將偽造片段剪貼回原始語音，從而在偽造片段邊界不可避免地引入訊號不連續。接著，介紹神經編解碼器，說明這類模型可將連續語音壓縮成長度與文字近似的離散符元，既保留語音細節又大幅降低序列長度，為後續生成框架奠定基礎。隨後，本章比較了兩種可同時條件在文字稿與雙向語音上下文的語音編輯模型：一種是 VoiceCraft，僅含解碼器的自回歸 Transformer，可透過因果遮蔽與延遲堆疊達成語音補全以及零

doi:10.6342/NTU202501466

樣本文字轉語音，但推論延遲仍受逐步生成所限；另一種是 VoiceBox，以流匹配為核心的平行化生成方法，藉由學習向量場來求解極少步的常微分方程，兼具高品質與低延遲。

透過上述脈絡，本章串連出本論文提出的 VoiceNoNG 技術路徑：先利用神經編解碼器將語音波形符元化，再以流匹配框架訓練一個語音編輯模型，既克服自回歸方法的速度瓶頸，又保持語音與文字序列長度相當的便利性。憑藉此模型，我們得以生成不含邊界不連續的部分偽造語音樣本，從而構建新的資料集，以推動對部分偽造語音更嚴謹且更貼近真實場景的研究。

doi:10.6342/NTU202501466

第三章　語音編輯模型 VoiceNoNG

3.1　先前研究的侷限

為了建立一個在部分偽造語音任務下具有高度擬真的新資料集，首要條件是擁有一個精準且自然的語音編輯模型——在完全遵循目標文字稿的同時，不改變語者音色與韻律，並且盡量保留原有背景聲與上下文一致。在前章節，本論文介紹了兩種目前最具代表性的語音編輯模型：流匹配的 VoiceBox 與自回歸的 VoiceCraft。兩者雖具有領先性能，但仍有各自的侷限，難以直接滿足我們的資料集需求。

VoiceBox 依賴流匹配進行平行生成，在零樣本文字轉語音任務中達到基於自回歸模型 10～20 倍的生成效率。然而，當編輯語音含有環境音或背景噪聲時，它常出現失真或抹除背景的情形。原因主要有二：(1) VoiceBox 僅使用無背景噪聲的 LibriLight [34] 有聲書語料庫訓練，缺乏多樣化的錄音場景；(2) VoiceBox 採用對數梅爾頻譜圖作為語音表徵，其輸出需要再交由神經聲碼器 HiFi-GAN [35] 重新還原為語音波形。既有研究指出，此組合對非語音成分（環境音、音樂等）普遍泛化能力較差 [36, 37]。

VoiCraft 雖然主觀聽感優於 VoiceBox，然而其字詞錯誤率（Word Error Rate）反而較高（參考表 3.2）。深入檢視後發現，高字詞錯誤率主要源自注意力錯亂：

doi:10.6342/NTU202501466

當目標文字稿含重複或複雜的片段時，模型有時會出現漏讀、重讀或錯讀等「幻覺」（Hallucination）現象，破壞語義準確度。

總結而言，VoiceBox 缺乏背景聲魯棒性，而 VoiceCraft 則受注意力錯亂困擾。為同時克服這兩大侷限，我們提出 VoiceNoNG：一方面沿用 VoiceBox 的流匹配生成框架，以維持高速且可控的語音編輯；另一方面則採用 VoiceCraft 的神經編解碼器符元化策略，以提升對多樣背景聲的適應力。透過這種流匹配加上神經編解碼器的融合框架，VoiceNoNG 兼具背景聲保留能力與語義精準度，為我們的新資料集提供了更可靠的支援。

3.2 VoiceNoNG 技術設計

由於 VoiceBox 尚未開源，我們首先完整重現了其流匹配生成框架，作為 VoiceNoNG 的基礎。語音表徵方面，VoiceNoNG 採用神經編解碼器 DAC 作為符元化器。與 EnCodec 相比，DAC 的預訓練語料同時涵蓋語音、音樂與環境音，更契合「保留上下文背景聲」的需求。兩者在架構與推論流程上大體一致，差別在於 DAC 使用 9 層殘差向量碼簿，而 EnCodec 僅採 4 層，因而能捕捉更精細的聲學細節。除了受到 VoiceCraft 以 EnCodec 作為語音符元化器的啟發外，採用神經編解碼器的另一個關鍵在於其向量量化機制，該機制將連續表徵量化至離散潛在空間，只要模型的預測誤差未跨越量化決策邊界，即便數值略有偏移，也會被量化至正確符元，從而提升整體魯棒性。

具體來說，VoiceNoNG 是使用將語音波形通過 DAC 編碼器得到的「量化前（Pre-quantization）表徵」作為語音表徵。單次的前向傳播流程與 VoiceBox 相同（參考圖 2.5），僅將對數梅爾頻譜圖替換為量化前表徵。訓練模型時，主要的目標

doi:10.6342/NTU202501466

函數是流匹配損失：

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}\left\|u_t(x_t) - v_t(x_t, x_{ctx}, z; \theta)\right\|^2$$

與原始 VoiceBox 不同，我們將「僅遮蔽區重建」改為「整句重建」。如此可同時確保編輯片段與上下文流暢銜接，且未編輯片段仍維持高音質重建。

為進一步提升符元預測的精確度，我們仿效 NaturalSpeech 2 [38] 的做法，在訓練過程中額外加入一項交叉熵（Cross Entropy）損失。具體而言，假設殘差向量量化由 K 層碼簿組成，每層含 N 個碼字，以 $b_{k,n}$ 表示第 k 層的第 n 個碼字；令 M 為所有被遮蔽位置的集合，$Q_k(x)_i$ 表示模型在位置 i 預測的量化前表徵經過第 k 層殘差碼簿量化後所選取的碼字，$c_{k,i}$ 則為該位置的正確碼字。在將所有向量先行經過二範數（L2 normalization）標準化後，交叉熵損失定義為

$$\mathcal{L}_{CE} = -\frac{1}{|M||K|}\sum_{i\in M}\sum_{k=1}^{K}\log\frac{\exp(-\left\|Q_k(x)_i - c_{k,i}\right\|_2)}{\exp(\sum_{n=1}^{N} -\left\|Q_k(x)_i - b_{k,n}\right\|_2)}$$

這項損失鼓勵模型縮短與正確碼字的距離，同時拉開與其他候選碼字的距離，從而提升量化器輸出的可靠性。

而條件式語音生成中，模型通常必須在「完美地遵循條件」與「語音自然度／多樣性」之間取得平衡。VoiceNoNG 延續了無分類器引導（Classifier Free Guidance）的策略，利用同一組參數權重 θ 同時得到有條件與無條件的兩種向量場。在推論時，我們對同一雜訊樣本，執行兩次的前向傳播，一次保留條件（x_{ctx}, z）得到有條件向量場 $v_t(x_t, x_{ctx}, z; \theta)$；另一次則把（$x_{ctx}, z$）設為零向量，得到無條件向量場 $v_t(x_t; \theta)$。再以係數 α 將兩者線性混合

$$\tilde{v}_t(x_t, x_{ctx}, z; \theta) = (1+\alpha)\cdot v_t(x_t, x_{ctx}, z_{emb}; \theta) - \alpha \cdot v_t(x_t; \theta)$$

doi:10.6342/NTU202501466

便能按需求強化（$\alpha > 0$）或弱化（$\alpha < 0$）條件訊息。此方法既避免了引入額外分類器帶來的計算與不穩定性，又保留了語義精準度，對流匹配生成框架尤為友好，在極少步積分下仍能產生流暢且低失真的語音。

3.3 訓練資料與測試資料

我們選用 GigaSpeech [39] 作為 VoiceNoNG 的訓練語料庫。GigaSpeech 是一套涵蓋多領域的英語語音識別資料集，錄音場景包括朗讀、對話、採訪、脫口秀、線上課程、廣播與音樂評論等，較 LibriLight 更貼近真實環境中的噪音與口音變化。該資料集總量超過 10,000 小時，所有文字稿先由自動語音識別（Automatic Speech Recognition）系統生成初稿，隨後經人工與半自動校正。在過濾階段，主集的字詞錯誤率上限設為 4%，其他較小子集則嚴格控制至 0%。此外，GigaSpeech 還提供強制對齊的時間標籤，可直接用於切段後的模型訓練。

測試資料方面，由於 VoiceCraft 並非自行訓練，且其自回歸生成流程與基於流匹配的 VoiceBox ／ VoiceNoNG 截然不同，故採用 VoiceCraft 團隊釋出，含有 VoiceCraft 編輯後語音的「真實編輯資料集」（RealEdit Dataset）進行評測，而非後續會介紹的「語音補全編輯資料集」。「真實編輯資料集」共收錄 310 組經人工設計的測試樣本，每組包含原始語音、原始文字稿與編輯後文字稿，並確保語法正確、語意流暢。該資料集由 100 組 LibriTTS [40] 語音、100 組 YouTube 語音 [39] 及 110 組 Spotify 播客片段 [41] 組成，所有文字稿均由母語英語者校對，編輯操作涵蓋插入、刪除與替換，字數編輯自 1～2 字至 7～12 字不等，且可包含多處待編輯區段，表 3.1 展示部分樣本。對於「真實編輯資料集」中的每組〈原始語音、原始文字稿、編輯後文字稿〉三元組，我們讓所有語音編輯模型生成對應的編輯後語音，並依後述評估指標進行性能比較。

doi:10.6342/NTU202501466

表 3.1: 真實編輯資料集樣本

編輯類型	原始文字稿	編輯後文字稿
刪除	I wrote the title **of the course many years ago, ah,** when I created this course.	I wrote the title when I created this course.
插入	And **we're at** this point.	And we're **all extremely excited** at this point.
替換、替換	See why it's extremely **valuable to it's kind of like having a wall hack** to watch a demo.	See why it's extremely **important right? it's kind of like having a rough time** to watch a demo.

接下來，我們藉由與 VoiceBox 及 VoiceCraft 的比較，來評估所提出的 VoiceNoNG 語音編輯模型之能力，如未特別註明模型參數量大小則預設為 330M。我們比較了兩種 VoiceCraft 模型（分別為 330M 與 880M 參數量版本）以及兩種 VoiceBox 模型（分別在 LibriLight 與 GigaSpeech 上訓練）。

3.4 實驗結果

3.4.1 字詞錯誤率

字詞錯誤率是評估文字轉語音與語音編輯系統能否準確對照文字稿逐字生成語音的常用指標。表 3.2 列出在使用 Whisper-large-v3 [42] 作為自動語音辨識模型時，各語音編輯方法在「真實編輯資料集」上的字詞錯誤率。字詞錯誤率透過將語音辨識轉錄結果與資料集給定的目標文字稿比較計算得出。

如表所示，所有基於流匹配的方法均顯著優於 VoiceCraft，這一結果與 VoiceCraft 原論文中的觀察相符。我們推論這與 VoiceCraft 易因注意力幻覺而出現漏詞或重詞相關，此現象將在第 3.4.5 章中深入分析。此外，當將 VoiceBox 的訓練語料庫從無背景噪音的 LibriLight 更換為涵蓋多種場景的 GigaSpeech 後，所有來源的字詞錯誤率均有不同程度的下降；再進一步以神經編解碼器 DAC 提取的

doi:10.6342/NTU202501466

量化前表徵取代對數梅爾頻譜圖作為語音表徵（即 VoiceNoNG），字詞錯誤率又獲得了進一步的改善。

我們也驗證了利用量化前表徵作為輸出以及加入交叉熵損失對提升性能的作用。若 VoiceNoNG 輸出量化前表徵，需經向量量化及 DAC 解碼器才能還原波形，此流程對於細微預測誤差展現更強的魯棒性；反之，若直接輸出量化後的表徵，雖僅需 DAC 解碼器，但平均字詞錯誤率會顯著提高。而加入交叉熵損失可提升符元預測精度，進而進一步降低字詞錯誤率。

表 3.2: 字詞錯誤率各語音編輯模型比較

	字詞錯誤率 ↓ (%)			
模型	LibriTTS	YouTube	Spotify	平均
編輯前語音	1.84	6.13	5.87	4.65
VoiceCraft (330M)	3.72 (0.13)	7.41 (0.38)	**4.63** (0.14)	5.23 (0.16)
VoiceCraft (830M)	3.77 (0.41)	7.36 (0.35)	5.43 (0.33)	5.52 (0.19)
VoiceBox (LibriLight)	3.64 (0.19)	6.26 (0.19)	5.45 (0.19)	5.13 (0.08)
VoiceBox (GigaSpeech)	3.48 (0.16)	6.03 (0.16)	5.36 (0.13)	4.97 (0.12)
VoiceNoNG	**2.82** (0.20)	5.84 (0.21)	4.92 (0.21)	**4.54** (0.14)
無交叉熵損失	2.94 (0.23)	5.81 (0.18)	5.07 (0.13)	4.62 (0.09)
改量化後表徵	3.16 (0.19)	**5.59** (0.24)	5.38 (0.22)	4.73 (0.18)

3.4.2 訊噪失真比

為了客觀評估生成語音的品質，我們採用無參考的尺度不變訊噪失真比（SI-SDR）[43] 估計方法。表 3.3 中列出了各模型在不同語料上訊噪失真比的分數。結果顯示，LibriTTS 原始語音的訊噪失真比高於 YouTube 與 Spotify，這反映了後兩者錄音常含環境背景聲的特性。雖然在 YouTube 與 Spotify 上，訊噪失真比分數對編輯語音的區分度較低，但我們仍將其納入比較，以提供更全面的參考。

在 LibriTTS 資料上，VoiceNoNG 的訊噪失真比得分最高，最貼近原始語音。而 VoiceBox（GigaSpeech）的分數最低，可能源於其依賴對數梅爾頻譜圖與聲碼

doi:10.6342/NTU202501466

器 HiFi-GAN 的還原所致。進一步地，我們發現若將 VoiceNoNG 改為直接輸出量化後表徵，訊噪失真比會大幅下降。同時，VoiceBox（GigaSpeech）及 VoiceNoNG（量化後表徵）模型的分數標準差較大，與我們的主觀聆聽經驗相符：這些模型魯棒性較低，在不同隨機種子下品質波動顯著。上述實驗結果進一步驗證了預測量化前表徵並倚賴殘差向量量化模組糾正誤差的策略，在保持音質穩定性方面的優勢。

表 3.3: 訊噪失真比各語音編輯模型比較

	訊噪失真比↑			
模型	LibriTTS	YouTube	Spotify	平均
編輯前語音	23.94	19.55	19.22	20.85
VoiceCraft (330M)	22.22 (0.08)	19.97 (0.08)	20.15 (0.16)	20.76 (0.04)
VoiceCraft (830M)	22.53 (0.07)	20.04 (0.06)	20.26 (0.13)	20.92 (0.08)
VoiceBox (LibriLight)	19.49 (0.13)	17.96 (0.16)	16.32 (0.10)	17.88 (0.03)
VoiceBox (GigaSpeech)	18.10 (0.41)	16.86 (0.32)	15.83 (0.17)	16.90 (0.20)
VoiceNoNG	23.15 (0.09)	19.29 (0.05)	19.04 (0.08)	20.44 (0.05)
無交叉熵損失	23.36 (0.09)	19.30 (0.05)	18.98 (0.06)	20.50 (0.04)
改量化後表徵	20.80 (0.26)	18.28 (0.11)	17.81 (0.10)	18.93 (0.05)

3.4.3 主觀評估

為了主觀評估各語音編輯方法的音質，我們設計了一項五點制聆聽測試。受測者會隨機聆聽真實語音或編輯後的語音，並依照自然度在 1 至 5 分量表上打分：5 分代表高度自然且未經修改的語音，1 分則表示強烈判定該語音已被部分編輯，完整的聆聽測試指令如圖 3.1 所示。我們分別從 LibriTTS、YouTube 和 Spotify 中各隨機挑選 8 段真實語音，再對應產生由 VoiceCraft、VoiceBox、VoiceNoNG 以及 VoiceNoNG（量化後表徵）所生成的編輯語音，共計 $(8\times3)\times5 = 120$ 段語音供每位受測者評分。共有 15 位受測者參與此研究，實驗結果如圖 3.2 所示。VoiceNoNG 與 VoiceCraft 的編輯語音在自然度上與真實語音相當（由於聆聽測試只聚焦音質，忽略語意和語法，因此 VoiceCraft 的幻覺問題並未顯現）。另一方面，VoiceNoNG

doi:10.6342/NTU202501466

（量化後表徵），其平均分數顯著下降，反映出直接輸出量化後表徵對音質造成的影響。

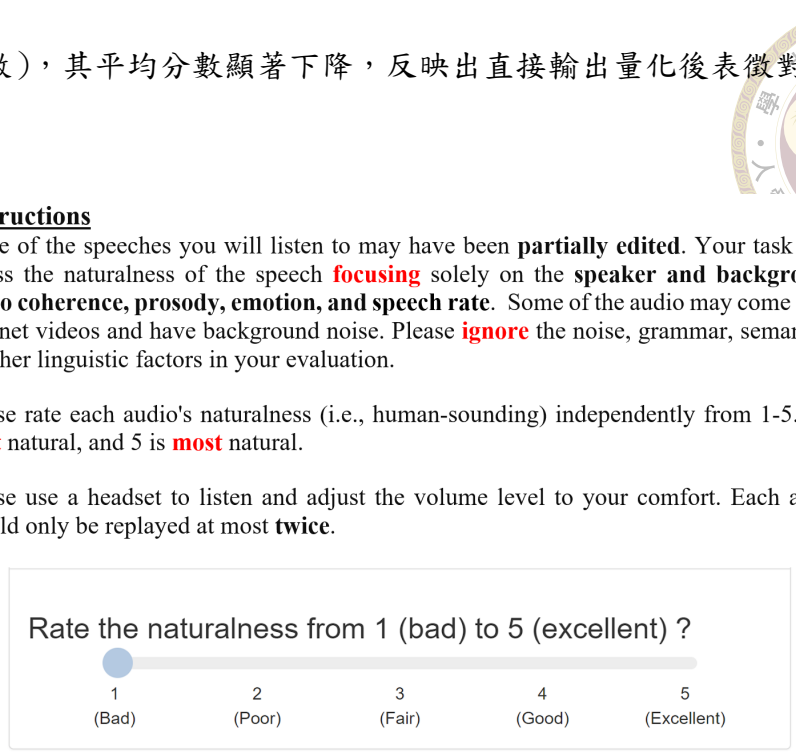


圖 3.1: 聆聽測試指令

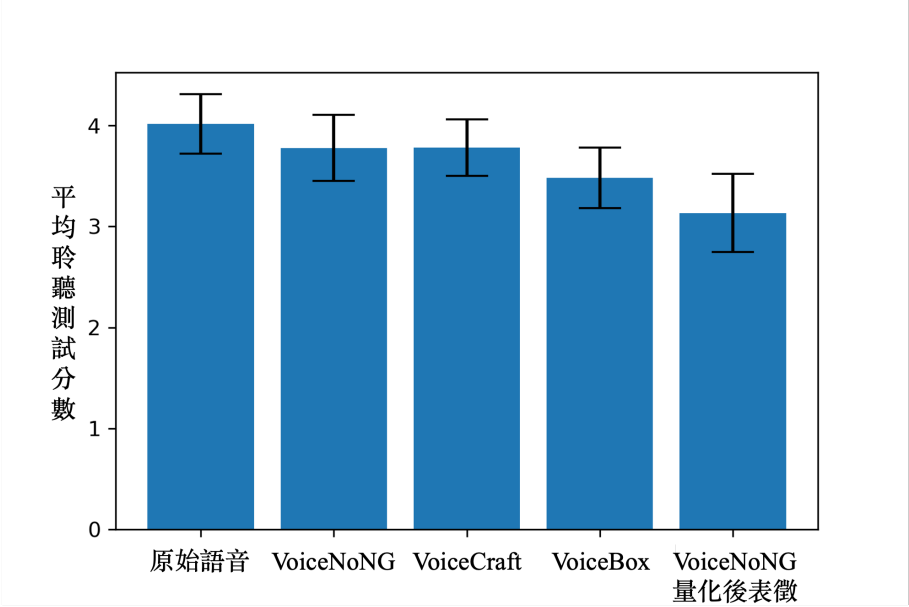


圖 3.2: 聆聽測試分數

3.4.4 量化前後表徵

如前述實驗所示，當以量化後表徵作為學習目標時，VoiceNoNG 的性能顯著下降。為驗證預測量化前表徵所帶來的優勢，我們在這兩種表徵上添加不同強度

doi:10.6342/NTU202501466

的高斯雜訊，模擬語音編輯模型輸出可能產生的預測誤差。本實驗採用了「真實編輯資料集」中來自 LibriTTS 的子集，並將受雜訊影響的 DAC 量化前後表徵重新合成為語音，最終以無參考的語音品質感知評估（Perceptual Evaluation of Speech Quality）[43] 指標評分。圖 3.3 顯示，在不同雜訊強度（雜訊量基於輸入表徵的動態範圍正規化）下，向量量化機制仍能保持卓越的魯棒性：在中度雜訊條件下，量化前表徵的語音品質感知評估得分約提升 0.2，進一步證明預測量化前表徵不僅穩健且有助於改進音質。

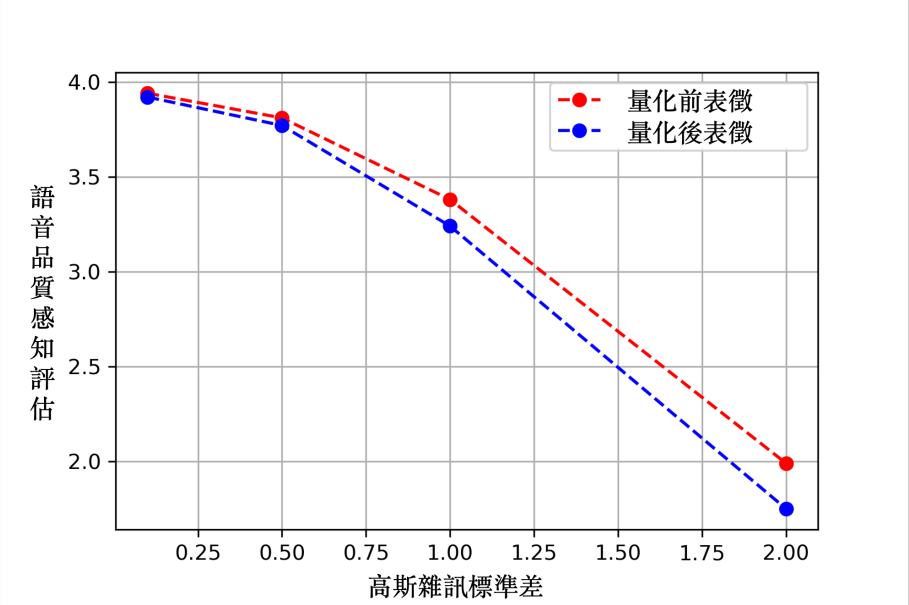


圖 3.3: 量化前後表徵在高斯雜訊影響下重合成語音品質評估

3.4.5 VoiceCraft 的幻覺現象

前文中指出 VoiceCraft 會出現類似大型語言模型的幻覺問題，例如意外生成過長的靜默、語速過慢，或漏字、重複字等現象。這一現象顯著的影響了 VoiceCraft 在字詞錯誤率上的表現。表 3.4 與表 3.5 分別列舉了兩個具體例子：在範例一中，VoiceCraft 漏生成了長詞組“from the bank for the bail”；在範例二中，系統則無中生有地插入“we are denounced as gooders”等多個詞彙。

doi:10.6342/NTU202501466

表 3.4: VoiceCraft 幻覺現象範例一

	Whisper-large-v3 語音辨識結果
目標文字稿:	”promise that you will not ask me to borrow any money from the bank for the bail of you for mister van brandt she rejoined and i accept your help gratefully”
VoiceCraft:	”promise that you will not ask me to borrow any money **-from the bank for the bail** of you -for +from mister van -brandt she +branch you re-joined and i accept your help gratefully”
VoiceNoNG:	”promise that you will not ask me to borrow any money from the bank for the -bail +money of you for mister van -brandt +brant she rejoined and i +will accept your help gratefully”

表 3.5: VoiceCraft 幻覺現象範例二

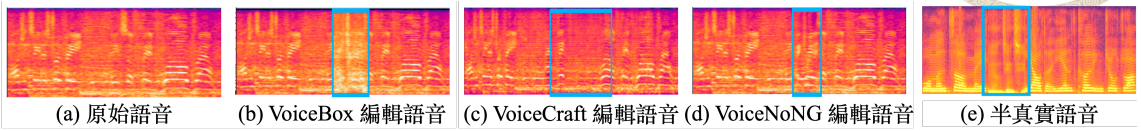
	Whisper-large-v3 語音辨識結果
目標文字稿:	“yet anytime you and i question the schemes of the dogooders or dare to dig into any of their motives were denounced as being against their humanitarian goals they say we are always against things we are never for anything”
VoiceCraft:	“yet anytime you and i question the schemes of the -dogooders +dog or dare to dig into any of their motives **-were +we are denounced as gooders we are** denounced as being against their humanitarian goals they say we are always against things we are never for anything”
VoiceNoNG:	“yet anytime you and i question the schemes of the -dogooders +dog eaters or dare to dig into any of their -motives were +we are denounced as being against their humanitarian goals they say we are always against things -we are +were never for anything +and”

3.4.6 部分偽造語音可視化

我們將不同語音編輯方法生成的部分偽造語音以對數梅爾頻譜圖形式可視化，藉以直觀比較不同語音編輯方法在同一組原始語音及目標文字稿上的表現。實驗結果如圖 3.4 所示，原始語音是帶有背景噪聲的音檔，藍色區塊則標示各模型生成的偽造片段。VoiceBox 的輸出在偽造片段呈現響亮而模糊的聲音，VoiceCraft 則出現不自然的靜默。唯獨 VoiceNoNG 生成的片段既音質清晰，又與原始背景聲保持一致，邊界過渡最為平滑。

doi:10.6342/NTU202501466

圖中亦包含「半真實語音偵測資料集」的偽造樣本（即半真實語音）。由於該資料集是唯一的中文資料集，所以其並非基於相同的原始語音進行編輯，但仍可明顯觀察到偽造片段邊界的訊號不連續，進一步凸顯剪貼式流程帶來的缺陷。



(a) 原始語音 (b) VoiceBox 編輯語音 (c) VoiceCraft 編輯語音 (d) VoiceNoNG 編輯語音 (e) 半真實語音

圖 3.4: 不同語音編輯方法的部分偽造語音可視化

3.5 本章總結

本章首先回顧了現有語音編輯模型的不足：VoiceBox 雖具備流匹配與並行生成的速度優勢，卻在含背景聲的場景下泛化不佳；VoiceCraft 則因自回歸注意力機制易產生幻覺錯誤，導致漏字、重字或靜默過長等問題。為同時克服這兩大侷限，我們提出 VoiceNoNG——在流匹配生成框架中引入 DAC 的量化前表徵，並額外加上交叉熵損失以強化符元預測精度；同時保留無分類器引導機制，讓使用者可透過一個超參數 α 自由調節「遵循度」與「自然度／多樣性」之間的平衡。

在實驗設計上，我們採用「真實編輯資料集」進行客觀與主觀評測。客觀評估中，使用 Whisper-large-v3 計算字詞錯誤率，並以無參考的尺度不變訊噪失真比評估音質。結果顯示，VoiceNoNG 在字詞錯誤率、訊噪失真比兩項指標上均優於或持平於 VoiceBox 與 VoiceCraft，特別在含背景聲的場景下仍能保持高準確度與高音質。主觀聆聽測試亦印證了 VoiceNoNG 與原始語音及 VoiceCraft 同級的自然度，而直接輸出量化後表徵的版本則表現明顯下滑。研究進一步確認，使用量化前表徵連動殘差向量量化可對抗細微誤差，添加交叉熵損失則有效降低字詞錯誤率。整體而言，VoiceNoNG 成功融合了流匹配和神經編解碼器的優勢，為高品質、背景保留型語音編輯提供了一套穩健且高效的解決方案。

doi:10.6342/NTU202501466

第四章　語音補全編輯資料集

4.1　本章簡介

在完成對 VoiceNoNG 在多指標上的驗證後，我們進一步構建了專門用於部分偽造語音任務的新資料集「語音補全編輯資料集」。此資料集由可依語音上下文條件生成的 VoiceNoNG 產出，能夠生成不含邊界不連續的部分偽造語音樣本。以下章節將依序闡述本資料集的設計原則、生成流程、統計特徵，以及各類防偽偵測模型在該資料集上的實驗表現。

4.2　文字稿編輯

我們沿用與「半真實語音偵測資料集」中對整句語義進行微調的思路：從每個句子中選取一個命名實體，並以同類實體或其反義詞替換，從而改變整句的語義。不同於「半真實語音偵測資料集」，我們並未事先準備固定的命名實體或反義詞配對庫，而是希望產生更多元的編輯結果。為此，我們透過提示（Prompt）zephyr-7B-beta [44] 大型語言模型生成編輯後的文字稿。當大型語言模型輸出不符合預期格式，則以字級萊文斯坦距離（Levenshtein Distance）比對候選替換詞，並從中隨機選取一個替換詞完成替換，使編輯結果兼具多樣性與可控性。

doi:10.6342/NTU202501466

4.3 資料集生成流程

為確保「語音補全編輯資料集」的品質，並蒐集足夠樣本量，我們選用 LibriLight 的中等規模子集作為真實語音來源，並僅保留時長介於 6 至 8 秒的片段，最終獲得約 87,000 段錄音。由於 LibriLight 並未附帶文字稿，我們採用第三方整理的 LibriHeavy 提供所需的文字稿。依照 VoiceBox 論文中的流程，我們使用蒙特利爾強制對齊器（Montreal Forced Aligner）對每段語音與其文字稿進行強制對齊，從而獲取包括音素化逐字稿及對應音素時長在內的關鍵資料。隨後，對每段原始文字稿調用 zephyr-7B-beta 生成編輯後文字稿，識別並替換關鍵詞，再應用訓練完成的 VoiceNoNG，以語音補全技術合成最終語音。

在訓練／驗證／測試集的切分上，我們先依據每位語者在整體資料中的語音數量排序，先將語音數量最多的語者優先分配到訓練集，再依序分配至驗證集與測試集，並以約 6：2：2 的比例維持三者之間的語音數量。此策略不僅保證驗證集與測試集規模與「半真實語音偵測資料集」相當，且因基於語者劃分，使測試集中包含大量未在訓練集中出現的語者，可有效評估偵測器的語者層面泛化能力。「語音補全編輯資料集」的統計資料詳列於表 4.1。

語音類別	子集	樣本數	語者數	總時長（小時）	語音長度（秒）	
					最短	最長
真實語音	訓練集	26,547	70	51.82	6.00	8.00
	驗證集	8,676	100	16.98	6.00	8.00
	測試集	8,494	900	16.60	6.00	8.00
部分偽造語音	訓練集	26,546	70	51.98	5.40	9.08
	驗證集	8,686	100	16.99	5.45	8.76
	測試集	8,493	903	16.64	5.49	8.85

表 4.1: 語音補全編輯資料集的統計資料

doi:10.6342/NTU202501466

4.4 防偽偵測模型

我們使用目前最先進的四種部分偽造語音偵測模型來驗證我們提出的「語音補全編輯資料集」，分別是 2022 年「語音深度合成偵測挑戰」的前兩名 [45, 46]，以及 2023 年「語音深度合成偵測挑戰」的第二、第三名 [47, 48]。其中 2023 年第一名並未在競賽結束後提交論文，因此未納入本次比較。四個模型簡要的架構列於表 4.2。我們將從兩個維度評估它們的性能：一是判斷整句語音中是否包含偽造片段，二是對偽造片段具體位置的定位準確度。

表 4.2: 防偽偵測模型的模型架構

模型	前端表徵	骨幹架構	幀級分類器	句級分類器
模型一	Wav2Vec 2.0	線性投影	線性投影	注意力統計池化 ＋線性投影
模型二	對數梅爾頻譜圖	SENet ＋ Transformer	線性投影	注意力統計池化 ＋線性投影
模型三	對數梅爾頻譜圖	一維殘差卷積層 ＋雙向長短期記憶層	線性投影	偽造幀數＜3
模型四	對數梅爾頻譜圖	二維卷積層 ＋雙向閘門循環單元	線性投影	線性指數歸一化

4.4.1 模型一

模型一是 2022 年「語音深度合成偵測挑戰」第一名的防偽偵測模型，採用了 Wav2Vec 2.0 [49] 的微調架構。該方法首先以 Wav2Vec 2.0 提取語音表徵，語音表徵會經過一層線性投影映射為潛在向量，然後分為兩條分支執行偵測：在幀級（Frame-level）分支中，每個時間步的潛在向量會再經過一層線性投影，用以判斷該幀是否由語音編輯模型生成，從而精確定位偽造片段；在句級（Utterance-level）分支中，則在潛在向量後堆疊一層注意力統計池化（Attentive Statistic Pooling），先透過小型注意力網絡學習各幀的重要性，再依據這些權重計算加權平均與加權

doi:10.6342/NTU202501466

標準差，最終輸出一個固定維度的向量，並經一層線性投影以預測整段語音是否包含偽造內容。訓練過程中，Wav2Vec 2.0 也一併微調。為避免與 LibriLight 資料交疊，我們在實驗中以僅於 LibriSpeech [50] 上預訓練的 wav2vec2-base-960 替代原論文使用的 wav2vec2-xls-r。

4.4.2 模型二

模型二是使用 SENet [51] 作為基礎模型進行修改，SENet 是一種在 ResNet 上引入壓縮與激勵（Squeeze-and-Excitation）模組的變體，壓縮與激勵模組會為卷積層的各通道產生權重，對原始各通道特徵進行縮放，強化對最關鍵通道的響應，抑制不重要或冗餘的特徵。模型輸入為對數梅爾頻譜圖，通過 SENet 得到潛在向量，潛在向量會先經過單層的 Transfomer 做注意力的強化。接著，注意力強化的潛在向量會送入跟模型一完全相同的兩條分支分別執行幀級預測以及句級預測。

4.4.3 模型三

模型三以對數梅爾頻譜圖為輸入，首先經過七層一維殘差卷積堆疊，提取出潛在特徵向量。接著，該向量分別輸入兩個並行模組：一是雙向長短期記憶層，用以捕捉時間上的前後語境；二是多幀檢測模組（Multi-Frame Detection），由兩層一維卷積構成，透過設計適當的步幅實現下採樣，以獲得更大範圍的跨幀特徵。跨幀特徵經過重複（Repeat）上採樣後，與雙向長短期記憶層的輸出相加，再經一層線性投影，對每一語音幀進行真偽預測。至於句級判斷，若預測為偽造的連續幀數少於三幀，則將整段語音視為真實。

doi:10.6342/NTU202501466

4.4.4 模型四

模型四同樣以對數梅爾頻譜圖為輸入，首先通過五層的 VGG 式區塊，每個區塊採「二維卷積 → 批次正規化（Batch Normalization）→ 二維卷積 → 整流線性單位（ReLU）」堆疊。接著將所得特徵輸入兩層雙向閘門循環單元（Gated Recurrent Unit）與一層線性投影，對每個語音幀進行真偽分類。而句級的預測結果則是基於所有幀級的預測結果通過線性指數歸一化（Linear-softmax）池化為整句語音的預測。

4.5 實驗結果

4.5.1 評估在半真實語音偵測資料集

由於在「語音深度合成偵測挑戰」舉辦時，「半真實語音偵測資料集」尚未公開，且競賽的測試集混入大量背景噪音，再加上多數參賽組別會採用集成（Ensumble）多個模型的策略，以提升預測預測準確度。所以我們首先在「半真實語音偵測資料集」上，對這些重新實作的防偽偵測模型進行評測，以驗證它們的效能。在幀級我們使用 F1 分數作為評估指標，在句級則採用準確率（Accuracy），評估結果見表 4.3。

表 4.3: 語音防偽偵測模型在半真實語音偵測資料集上的評估分數

模型	幀級 F1（%）		句級準確率（%）	
	訓練集	測試集	訓練集	測試集
模型一	99.99	99.90	100.00	99.94
模型二	97.59	94.14	99.91	96.79
模型三	99.43	95.39	99.99	96.21
模型四	99.88	99.82	99.19	99.99

在「半真實語音偵測資料集」上的評測結果顯示，無論是在幀級 F1 分數還是

doi:10.6342/NTU202501466

句級準確率方面，四個防偽偵測模型幾乎都取得了 95% 以上的優異表現，其中模型一與模型四更高達 99–100% 的分數。這表明，在剔除「語音深度合成偵測挑戰」測試集中故意加入大量背景噪音所造成的領域不匹配（Domain Mismatch）影響後，偽造片段邊界不連續成為模型容易利用的強力線索，導致其採取捷徑學習。但同時，這也證實了這四個防偽偵測模型在現有資料集上的卓越性能，足以作為後續實驗的驗證基準。

4.5.2 評估在語音補全編輯資料集

接下來我們將四個語音防偽偵測模型應用在我們提出的「語音補全編輯資料集」上進行評估，其結果彙整於表 4.4。從中可以看到，模型一在各項評估指標上依然取得卓越表現，顯示其對於偽造片段的偵測與定位能力相當穩健。在句級預測上，所有防偽偵測模型都維持良好水準，在測試集上最差也有 90% 的準確率。至於幀級預測，除模型一外，其他三個防偽偵測模型，甚至在訓練階段就難以收斂，測試階段亦呈現顯著的性能退步。

表 4.4: 語音防偽偵測模型在語音補全編輯資料集上的評估分數

模型	幀級 F1（%）		句級準確率（%）	
	訓練集	測試集	訓練集	測試集
模型一	98.19	89.74	100.00	99.69
模型二	51.40	24.20	99.99	96.24
模型三	42.47	23.45	95.35	90.03
模型四	67.53	37.83	99.98	98.48

我們認為模型一優異的能力主要得益於其在大規模語音資料上的自監督預訓練，使模型可以習得更具通用性的表徵；反觀其餘防偽偵測模型缺乏此階段預訓練，導致在幀級預測時未能捕捉到偽造片段的真正特徵。此外，這三個防偽偵測模型在從「半真實語音偵測資料集」切換成「語音補全編輯資料集」後，於幀級預測結果上的顯著退步，也反映了偽造片段邊界不再產生訊號不連續，定位防偽

doi:10.6342/NTU202501466

片段將成為一項更具挑戰性的任務。

4.6 本章總結

本章實驗首先在「半真實語音偵測資料集」上驗證了四種最先進防偽偵測模型的性能。結果顯示，所有模型無論在句級的準確率或幀級的 F1 分數都能輕鬆突破 95%，其中部分模型甚至達到近乎完美的 99–100%。這證明在此資料集中，偽造片段邊界產生的訊號不連續為偵測模型提供了強力的捷徑線索。接著，我們將相同模型應用於本研究提出的「語音補全編輯資料集」，觀察其能力。僅有在大規模自監督預訓練後微調的模型能同時維持句級高準確度與穩定的幀級定位表現，其餘模型在句級仍能保有一定水準，但在幀級多數無法收斂，表現大幅下滑。這些結果不僅凸顯了「語音補全編輯資料集」在移除偽造邊界不連續後對偵測任務的嚴苛挑戰，也驗證了該資料集對於推動部分偽造語音偵測技術進步的必要性。

doi:10.6342/NTU202501466

第五章　重合成語音的影響

5.1　本章簡介

在第四章，我們已經驗證「語音補全編輯資料集」對推動部分偽造語音偵測技術的貢獻。然而，若細究資料集中真實語音與偽造語音之間的差異，便會發現一處容易被忽略的不匹配：由於 VoiceNoNG 是用神經編解碼器 DAC 生成的量化前表徵當作輸入與輸出，偽造語音中「未編輯」的片段同樣經歷了編解碼器的重合成。這意味著，在幀層級預測中，標記為真實的幀，其實包含了兩種型態——一種是原始語音中從未經神經編解碼器處理的語音幀，另一種則是偽造語音中經神經編解碼器處理但未被編輯的語音幀。為了探討這一差異對防偽偵測模型性能的影響，本章將專門分析並比較這兩類「真實」語音幀在偵測結果中的表現差異。

5.2　實驗場景定義

為探討「真實」語音幀是否經過神經編解碼器重合成，對偵測結果的影響，我們設計了三種實驗場景。各場景中真實／重合／補全／剪貼語音的生成方式如圖 5.1 所示，藍色片段代表真實語音，紅色片段代表偽造語音，綠色片段則代表經神經編解碼器處理的語音。

doi:10.6342/NTU202501466

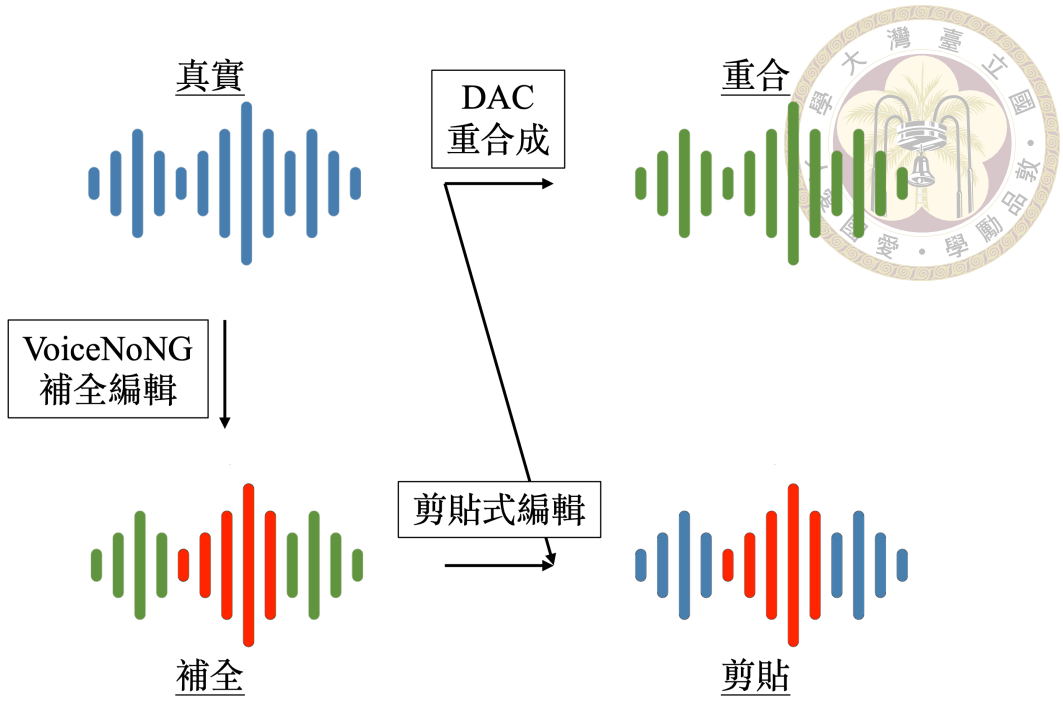


圖 5.1: 各場景中真實／重合／補全／剪貼語音的生成方式

1. **真實-補全**　真實語音採用原始錄音，而偽造語音則直接使用 VoiceNoNG 編輯後的輸出。此場景即是第四章實驗使用的場景。

2. **真實-剪貼**　在此場景中，我們不直接使用 VoiceNoNG 的輸出來作為偽造語音，而是將 VoiceNoNG 編輯後的偽造片段剪貼回原始錄音，確保所有標記為真實的語音幀均未經神經編解碼器處理。

3. **重合-補全**　先將整段真實語音通過神經編解碼器 DAC 的編碼器與解碼器進行重合成，而偽造語音則同樣直接使用 VoiceNoNG 編輯後的輸出，使所有標記為真實的語音幀均經歷一次編解碼過程。

透過比較這三種場景，我們能夠衡量神經編解碼器重合成對部分偽造語音偵測模型性能的具體影響。而每種語音幀是否經過神經編解碼器，整理在表 5.1。值得一提的是，以往反偽造研究通常將經過神經編解碼器重合成的語音視為偽造樣本 [52]。然而，本論文認為有必要在不同使用情境下將此類重合成語音獨立討論，我們將具體分析為何這些經重合成處理的語音不應一概歸為偽造語音。

doi:10.6342/NTU202501466

表 5.1: 各場景下不同類別語音幀是否經過神經編解碼器處理

場景	真實語音	偽造語音	
		未編輯	編輯
真實-補全	否	是	是
真實-剪貼	否	否	是
重合-補全	是	是	是

5.3 實驗結果

5.3.1 部分偽造語音偵測

實驗結果如圖 5.2 所示，模型一在三種場景的表現依然很好，都達到 88% 以上的表現，顯示自監督預訓練的通用特徵，對於「真實-剪貼」與「重合-補全」兩種新增場景同樣適用。而這也符合預期，因為新增的兩種場景相較於「真實-補全」場景，在所有標記為「真實」的語音幀上統一了是否經過神經編解碼器處理，任務難度因而降低。而模型三與模型四的實驗結果也證實了這個觀點（模型二的幀級預測在訓練集並沒有收斂，故不納入討論），兩者在「真實-剪貼」及「重合-補全」場景中，其幀級預測在訓練集有 3.79～32.52% 的 F1 分數進步。其中，「真實-剪貼」場境的進步幅度更勝「重合-補全」場景，其理由是在「真實-剪貼」場境下形成新的捷徑學習（參考表 5.1）：「真實-剪貼」場境下，真實語音幀都未經過編解碼，而偽造語音幀則都經過編解碼，原本的分類問題從語音幀是否經過編輯偽造，可以簡化為語音幀是否經過編解碼，形成新的捷徑學習路徑。

而值得注意的是，在「真實-剪貼」場境雖然對幀級預測提供了學習捷徑，但是在句級預測上則會增加難度，「真實-剪貼」場景在除了模型一的防偽偵測模型外，都出現了超過 25% 準確度大幅下降。這是因為偽造語音混合了經過編解碼與未經過編解碼的語音幀，使得對整句語音是否包含偽造片段的判斷變得更加複雜且容易過擬合。

doi:10.6342/NTU202501466

但幀級預測方面，除了模型一外，多數模型即使在訓練集上有可接受的表現，仍難以將該能力遷移至測試集，關於這一現象的深入分析將在後續實驗展開。

表 5.2: 語音防偽偵測模型在各場景下的評估分數

模型	場景	幀級 F1（%）		句級準確率（%）	
		訓練集	測試集	訓練集	測試集
模型一	真實-補全	98.19	89.74	100.00	99.69
	真實-剪貼	99.34	97.78	100.00	98.86
	重合-補全	96.59	88.05	100.00	98.39
模型二	真實-補全	51.40	24.20	99.99	96.24
	真實-剪貼	42.28	21.73	97.60	56.96
	重合-補全	39.60	18.56	99.26	78.74
模型三	真實-補全	42.47	23.45	95.35	90.03
	真實-剪貼	74.99	28.31	84.66	57.81
	重合-補全	61.12	18.61	96.86	73.02
模型四	真實-補全	67.53	37.83	99.98	98.48
	真實-剪貼	80.46	48.30	98.20	69.40
	重合-補全	71.32	30.24	99.38	76.86

表 5.3: 在「真實-剪貼」場景下對於不同資料集分割的幀級 F1 分數

模型	場景	不包含相同語者		包含相同語者	
		訓練集	測試集	訓練集	測試集
模型三	真實-剪貼	74.99	28.31	75.45	32.69
模型四	真實-剪貼	80.46	48.30	78.63	50.74

5.3.2 測試集消融實驗

為了排除語者分佈不一致導致的過擬合，我們以 6：2：2 的比例隨機切分資料集，不再依語者的語音數量排序，因而各子集中可能包含相同的語者（即訓練集的語者也出現在測試集中）。如表 5.3 所示，即便大多數語者在訓練階段已被模型見過，其幀級預測表現仍難以遷移至測試集，說明過擬合並非由語者差異所引起。我們推測，真正的原因在於神經編解碼器的向量量化機制：如第 3.4 章所證，只要模型的預測誤差未跨越量化決策邊界，向量量化便會將偏差糾正回正確

doi:10.6342/NTU202501466

符元，這在保證了語音品質的同時，也使得經 VoiceNoNG 編輯的語音分佈被拉回至真實語音分佈。結果，儘管模型在訓練集中記住了所有偽造片段，其對測試集中偽造片段的判斷仍易錯誤歸類為「真實」，最終導致預測錯誤。

5.3.3 跨場景泛化能力

從第 5.3.1 章的實驗結果我們可以發現，語音幀是否經過神經編解碼器處理對防偽偵測模型的性能有顯著影響。在「重合-補全」場景中，所有語音幀均經過神經編解碼器處理，是唯一所有語音幀型態都一致的場景。它消除了「偽造片段邊界不連續」與「編解碼處理差異」兩種本論文發現的模型學習捷徑，使模型只能依靠偽造片段的本質特徵進行學習。為驗證此假設，我們將探討模型一的跨場景泛化能力。如果「重合-補全」場景真正學習到了偽造片段的本質特徵，那其應該能藉由此能力泛化到另外兩個場景。實驗結果如表 5.4 所示，「重合-補全」場景訓練出的防偽偵測模型，展現出最佳的跨場景泛化能力。這證實當進一步去除了神經編解碼器處理所帶來的幀型態干擾後，模型確實得以聚焦於更普適的偽造片段特徵。

表 5.4: 語音防偽偵測模型在跨場景上的評估分數

訓練 \ 測試	真實-剪貼	真實-補全	重合-補全
幀級 F1（%）			
真實-剪貼	97.78	39.49	37.99
真實-補全	18.37	89.74	54.17
重合-補全	**76.13**	**80.70**	88.05
句級準確率（%）			
真實-剪貼	98.86	61.89	60.20
真實-補全	51.66	99.69	53.01
重合-補全	**83.19**	**93.25**	98.39

doi:10.6342/NTU202501466

5.3.4 跨編輯模型泛化能力

在確認「重合-補全」場景下訓練的防偽偵測模型具備最佳跨場景泛化能力後，我們進一步檢驗此泛化能力能否拓展至跨編輯模型的任務。為此，我們以 VoiceCraft 按照「語音補全編輯資料集」的生成流程，額外製作了一組「重合-補全」場景的測試集，用於跨編輯模型評估。由於 VoiceCraft 採用自回歸生成框架與神經編解碼器 EnCodec 處理，與 VoiceNoNG 的流匹配生成框架及神經編解碼器 DAC 存在顯著差異，可最大程度觀察模型習得的偽造片段特徵能否適用於不同編輯模型。所有防偽偵測模型均以 VoiceNoNG 的訓練集進行訓練，隨後分別在 VoiceNoNG 和 VoiceCraft 所生成的測試集上進行評估。實驗結果（表 5.5）顯示，在 VoiceNoNG 測試集上表現優異的防偽偵測模型，轉移到 VoiceCraft 測試集後，其幀級 F1 分數驟降至 1.59%，句級準確率也掉落至 52.81%。細究模型的預測結果可見，模型無論在幀級或句級上，幾乎都將 VoiceCraft 測試集所有的樣本判斷為「重合」，凸顯跨編輯模型偵測的極大難度，尤其在神經編解碼器存在差異時更為嚴峻。

針對這一挑戰，我們嘗試了兩種改進策略。其一，防偽偵測模型在「重合-補全」場景下，泛化到 VoiceCraft 測試集表現不佳的原因，我們推測可能在預測未見過的神經編解碼器處理的樣本時，模型未能區分「僅經神經編解碼器重合成」與「經語音編輯模型生成」之間的細微差異。換言之，如果模型學會辨識單純重合成後的語音特徵，便更有可能進一步偵測被編輯的語音。因此，我們在訓練集中加入了真實、剪貼兩種樣本，並改為「真實／重合／偽造（補全＋剪貼）」的三元標籤，使模型藉由區分「真實」與「重合」掌握編解碼流程帶來的特徵，同時從「偽造」類別學習檢測編輯偽造的痕跡，由此強化其在跨編輯模型下的泛化能力；其二，在此三元分類基礎上，對所有經過神經編解碼器處理的語音幀加入高

doi:10.6342/NTU202501466

斯雜訊（同時在波形及殘差向量量化的最終兩層碼字上施加噪聲），以增強模型對不同神經編解碼器的魯棒性。

儘管在 VoiceNoNG 測試集中，這兩種方法都能保持逾 95 % 的句級準確率，但幀級 F1 分數隨任務難度上升分別下降 5.30% 與 16.06%；在 VoiceCraft 測試集上，兩者的幀級 F1 分數均未突破 10%。但其模型輸出存在差異，三元分類是傾向於把 VoiceCraft 的樣本判段為「真實」，而三元分類＋高斯則傾向於判段為「重合」，這也是為什麼三元分類的句級準確率僅剩 4.12% 的原因。由此可見，即便進行了三元分類與高斯加噪優化，模型的跨編輯模型泛化能力仍未顯著提升，顯示學習跨編輯模型的偽造片段特徵仍是重點挑戰。

表 5.5: 語音防偽偵測模型在跨編輯模型上的評估分數

模型	VoiceNoNG 生成測試集		VoiceCraft 生成測試集	
	幀級 F1	句級準確率	幀級 F1	句級準確率
二元分類（重合／補全）	88.05	98.39	1.59	52.81
三元分類（真實／重合／偽造）	82.75	97.68	4.42	4.12
三元分類＋高斯（真實／重合／偽造）	71.99	95.68	8.83	50.35

5.4 本章總結

本章通過多場景、跨場景與跨編輯模型的實驗，系統地驗證了防偽偵測模型在移除學習捷徑後的真正泛化能力。首先，在「真實-補全」、「真實-剪貼」與「重合-補全」三種場景中，僅有經大規模自監督預訓練的模型能同時維持幀級與句級的高準確度。接著，證實當幀型態被統一的「重合-補全」場景下，模型才真正聚焦於偽造片段的本質特徵，而非「偽造片段邊界不連續」或「編解碼處理差異」造成的捷徑訊號。最後，跨編輯模型評測中，我們使用生成框架以及神經編

doi:10.6342/NTU202501466

解碼器與 VoiceNoNG 完全不匹配的 VoiceCraft 生成測試集，揭露防偽偵測模型在跨編輯模型任務上驟降的偵測效能——幀級 F1 分數甚至跌至不足 2%，句級準確率降至約 53%。對此，我們嘗試以三元分類（真實／重合／偽造）以及對神經編解碼器處理的語音幀加入高斯雜訊等策略強化模型，但仍無法改善防偽偵測模型的表現。

儘管三元分類的實驗不盡理想，我們認為將經神經編解碼器重合成的語音獨立分類，而不再與偽造語音劃為同類，仍是重要的研究方向：其一，隨著神經編解碼器技術成熟，大量合法語音將如 MP3 一般普及於壓縮／傳輸流程中；其二，來自「經編解碼」與「未編解碼」之間的聲學差異，已成為檢測捷徑，掩蓋真正的偽造特徵。未來防偽偵測模型應針對此新類別設計對應策略，以在更貼近實務的條件下，實現更為穩健的部分偽造語音偵測。

doi:10.6342/NTU202501466

第六章　結論與展望

本論文聚焦於「部分偽造語音」的合成與偵測，提出了一種基於流匹配的語音編輯模型 VoiceNoNG——它以神經編解碼器 DAC 輸出之量化前表徵作為符元，並同時條件於文字稿與語音上下文，實現了低延遲且高品質的語音生成，並依此構建了「語音補全編輯資料集」。在實驗中，我們首先驗證了 VoiceNoNG 相較於過去先進的語音編輯模型，在字詞錯誤率、訊噪失真比、語音品質感知評估以及主觀聆聽測試等多維度指標上都具有優勢。接著，我們以「語音補全編輯資料集」訓練四種防偽偵測模型，結果顯示在排除乏邊界不連續作為學習捷徑的的情況下，大多數偵測模型仍難以準確偵測與定位偽造片段。同時我們透過多場景、跨場景的實驗發現，編解碼前後的聲學差異亦成為模型的學習捷徑，只有在統一了編解碼處理的「重合-補全」場景訓練出的模型，才能真正聚焦於偽造片段的本質特徵。最後，在跨編輯模型（VoiceCraft）的泛化實驗中，現有的防偽偵測模型依然難以跨越生成框架與編解碼器的差異。

基於上述觀察，跨編輯模型的泛化能力應當是未來防偽研究的重要課題，而我們認為其關鍵應在於將神經編解碼器重合成的語音獨立於「偽造」或「真實」之外，形成新的評估類別，以進一步強化其對真實／重合／偽造片段的識別能力。隨著神經編解碼器在壓縮上的應用日益普及，未來合法錄音經過神經編解碼器處理已成常態，迫切需要開發對此不敏感、能聚焦於更普適偽造特徵的防偽技術，以應對日益複雜的真實應用場景。

doi:10.6342/NTU202501466

參考文獻

[1] Rohan Kumar Das, Xiaohai Tian, Tomi Kinnunen, and Haizhou Li. The attacker's perspective on automatic speaker verification: An overview. In Conference of the International Speech Communication Association, 2020.

[2] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970, 2023.

[3] Haibin Wu, Jiawen Kang, Lingwei Meng, Helen Meng, and Hung-yi Lee. The defender's perspective on automatic speaker verification: An overview. In Workshop on Deepfake Audio Detection and Analysis, 2023.

[4] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Conference of the International Speech Communication Association, 2017.

[5] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In Conference of the International Speech Communication Association, 2019.

doi:10.6342/NTU202501466

[6] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. Automatic Speaker Verification and Spoofing Countermeasures Workshop, 2021.

[7] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. In Automatic Speaker Verification and Spoofing Countermeasures Workshop, 2024.

[8] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In Automatic Speaker Verification and Spoofing Countermeasures Workshop, 2021.

[9] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Raw differentiable architecture search for speech deepfake and spoofing detection. In Automatic Speaker Verification and Spoofing Countermeasures Workshop, 2021.

[10] Nicolas M Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do asvspoof-trained models really learn? In Automatic Speaker Verification and Spoofing Countermeasures Workshop, 2021.

[11] Haibin Wu, Songxiang Liu, Helen Meng, and Hung-yi Lee. Defense against adversarial attacks on spoofing countermeasures of asv. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

doi:10.6342/NTU202501466

[12] Zhiyuan Peng, Xu Li, and Tan Lee. Pairing weak with strong: Twin models for defending against adversarial attack on speaker verification. In Conference of the International Speech Communication Association, 2021.

[13] Haibin Wu, Xu Li, Andy T. Liu, Zhiyong Wu, Helen Meng, and Hung-Yi Lee. Improving the adversarial robustness for speaker verification by self-supervised learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.

[14] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.

[15] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. Add 2023: the second audio deepfake detection challenge. In Workshop on Deepfake Audio Detection and Analysis, 2023.

[16] Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-truth: A partially fake audio detection dataset. In Conference of the International Speech Communication Association, 2021.

[17] Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Chao-Han Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung-yi Lee, and Szu-Wei Fu. Detecting the undetectable: Assessing the efficacy of current spoof detection methods against seamless speech edits. In IEEE Spoken Language Technology Workshop, 2024.

[18] Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Pin-Jui Ku,

doi:10.6342/NTU202501466

Ante Jukić, Chao-Han Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung-yi Lee, and Szu-Wei Fu. Voicenong: High-quality speech editing model without hallucinations. In Conference of the International Speech Communication Association, 2025.

[19] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. In Advances in neural information processing systems, 2023.

[20] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In Meeting of the Association for Computational Linguistics, 2024.

[21] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus. In Conference of the International Speech Communication Association, 2021.

[22] Junyi Sun. Jieba: Chinese text segmentation: built to be the best python chinese word segmentation module. https://github.com/fxsjy/jieba, 2019.

[23] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In International Conference on Machine Learning, 2018.

[24] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In International Conference on Machine Learning, 2018.

doi:10.6342/NTU202501466

[25] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[26] Cisco Systems. Global - 2021 forecast highlights. https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf, 2021.

[27] Manfred Schroeder and B Atal. Code-excited linear prediction (celp): High-quality speech at very low bit rates. In IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985.

[28] Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. IEEE Transactions on Audio, Speech and Language Processing, 2025.

[29] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. Transactions on Machine Learning Research, 2024.

[30] 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, 2017.

[31] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. Advances in Neural Information Processing Systems, 2023.

[32] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neu-

doi:10.6342/NTU202501466

ral ordinary differential equations. Advances in Neural Information Processing Systems, 2018.

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In International Conference on Learning Representations, 2023.

[34] 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 IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[35] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 2020.

[36] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In International Conference on Learning Representations, 2023.

[37] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. arXiv preprint arXiv:2312.15821, 2023.

[38] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and

doi:10.6342/NTU202501466

zero-shot speech and singing synthesizers. In International Conference on Learning Representations, 2024.

[39] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In Conference of the International Speech Communication Association, 2021.

[40] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In Conference of the International Speech Communication Association, 2019.

[41] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 podcasts: A spoken English document corpus. In International Conference on Computational Linguistics, 2020.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, 2023.

[43] Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2023.

[44] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan

doi:10.6342/NTU202501466

Habib, et al. Zephyr: Direct distillation of lm alignment. In Conference on Language Modeling, 2024.

[45] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. Fake audio detection based on unsupervised pretraining models. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.

[46] Haibin Wu, Heng-Cheng Kuo, Naijun Zheng, Kuo-Hsuan Hung, Hung-Yi Lee, Yu Tsao, Hsin-Min Wang, and Helen Meng. Partially fake audio detection by self-attention-based fake span discovery. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.

[47] Jie Liu, Zhiba Su, Hui Huang, Caiyan Wan, Quanxiu Wang, Jiangli Hong, Benlai Tang, and Fengjie Zhu. Transsionadd: A multi-frame reinforcement based sequence tagging model for audio deepfake detection. In Workshop on Deepfake Audio Detection and Analysis, 2023.

[48] Kang Li, Xiao-Min Zeng, Jian-Tao Zhang, and Yan Song. Convolutional recurrent neural network and multitask learning for manipulation region location. In Workshop on Deepfake Audio Detection and Analysis, 2023.

[49] 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, 2020.

[50] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In IEEE international conference on acoustics, speech and signal processing, 2015.

doi:10.6342/NTU202501466

[51] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. Assert: Anti-spoofing with squeeze-excitation and residual networks. In Conference of the International Speech Communication Association, 2019.

[52] Haibin Wu, Yuan Tseng, and Hung-yi Lee. Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. In Conference of the International Speech Communication Association, 2024.

doi:10.6342/NTU202501466