



DiffSinger 歌聲合成模型

原論文:《DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism》作者:劉靜林

學生:王竣揚 詹為翔 指導教授:葉經緯教授

1. 背景介紹

在當今蓬勃發展的人工智慧領域中，智慧語音處理正成為一個極具吸引力的研究方向。本專題的動機源於對語音合成技術的熱切興趣以及對「DiffSinger」這項尚未成熟但充滿潛力的技術吸引。我們希望深入探究語音合成領域，了解其原理和運作方式，並尋找可能的改進方法。

2. 摘要

本專題的目的在於探討歌唱語音合成系統(SVS)，且嘗試模擬相關環境、訓練神經網路模型，最終試圖提高模型的效能。首先，我們錄製語音，接者透過響度匹配、數據切分、標註及強制對齊等等步驟製作出用於訓練的數據集，並使用淺層擴散機制來訓練聲學模型，同時分析模型在訓練集與驗證集的損失值(Loss)，最終，通過調整訓練時的超參數，以確保模型在訓練過程中達到更高的性能水平。

3. 理論基礎

DiffSinger 以擴散機率模型為基礎並採用淺層擴散機制，優化了語音合成中的逆向轉換過程。傳統模型需要耗費大量計算，逐步從高斯分布的頻譜圖出發，DiffSinger 則在訓練階段使用 Aux Decoder 生成模糊的頻譜圖 M ，並隨著擴散步數增加，使其趨近於真實頻譜。這樣，逆向轉換的步數大幅減少，計算效率大幅提升，為語音合成技術帶來了更高的效能。

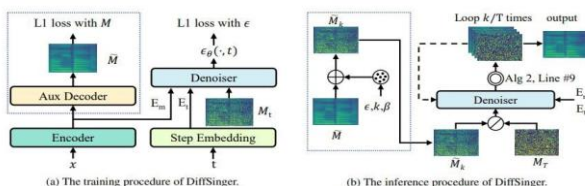


圖 1.diffsinger 模型架構

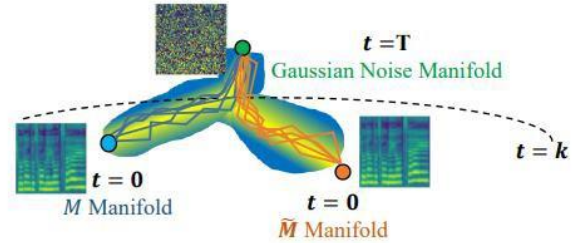
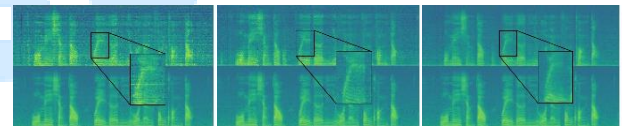


圖 2.淺層擴散機制

4. 實作方法&結果

本次專題研究旨在提高 DiffSinger 模型效能，其中調整 batch size 和調整 learning rate (學習率) 是有效的方法我們使用 Batch Size 24 和 48，以及 Learning Rate 0.0004 和 0.001 這兩種超參數做搭配，每個模型都進行了訓練，訓練步驟數為 9000，以下是不同參數下的 Loss 和訓練出的頻譜圖。



BS:48 LR:0.001 BS:24 LR:0.001 BS:24 LR:0.0004

超參數	訓練集	驗證集
BS:48 LR:0.001	0.7041	0.05437
BS:24 LR:0.001	0.07001	0.05303
BS:24 LR:0.0004	0.06881	0.02944

5. 討論與結論

這項研究針對 DiffSinger 的聲音合成技術進行了超參數優化，以提升合成效果和品質。經過實驗和調整，我們成功改進了模型性能，生成的聲音與原始聲音高度相似。最佳超參數配置為 Batch Size 24 和 Learning Rate 0.0004，有效提高了模型訓練效果。雖然仍有改進空間，但這項研究為聲音合成提供了寶貴洞見，未來應用潛力廣泛，包括音樂創作和語音助手等領域。同時，我們也探討了模型的結構優化可能性，為未來研究方向提供啟發。