基于扩散模型的钢琴音色转换系统研究

Chun-Chieh Hsu[1], Tsai-Ling Hsu[2], Chen-Chen Yeh[3], Shao-Chien Lu[4], Cheng-Han Wu[5], Bing-Ze Liu[6], Timothy K. Shih[7] 和 Yu-Cheng Lin[81] 元智大学，计算机科学与工程系，台湾 s1101545@mail.yzu.edu.tw[2] 元智大学，资讯国际学士学位学程，台湾 s1093530@mail.yzu.edu.tw[3] 元智大学，计算机科学与工程系，台湾 yehchenchen93@gmail.com[4] 元智大学，计算机科学与工程系，台湾 s1111534@mail.yzu.edu.tw[5] 国立中央大学，计算机科学与信息工程学系，台湾，whoshank@gmail.com[6] 国立中央大学，计算机科学与信息工程学系，台湾，iojkpghjkhik2@gmail.com[7] 国立中央大学，计算机科学与信息工程学系，台湾，timothykshih@gmail.com[8] 元智大学，计算机科学与工程系，台湾 linyu@saturn.yzu.edu.tw

摘要。我们提出了一种基于扩散架构的音色转换模型，旨在将各种乐器演奏的音乐精准转换为钢琴版本。该模型采用音高编码器和响度编码器提取音乐的音高特征与响度特征，并将其作为条件输入传递给扩散模型的解码器，以生成高质量的钢琴音色。案例分析结果表明，该模型在音高准确度和音色相似度方面表现出色，并且能够在不同音乐风格（古典音乐、爵士乐、流行音乐）和不同长度（从短片段到完整曲目）的音乐中保持稳定的转换效果。尤其是在处理音符快速变化和音乐结构复杂的内容时，该模型仍能保持较高的音质和准确度，展现出良好的泛化能力。此外，该模型具备实时音乐转换的潜力，适用于现场演出和数字音乐创作工具。未来研究将着重提升模型对响度动态的处理能力，并纳入更多音乐特征（如音色变化和节奏复杂度），以增强模型的适应性和表现力。我们还计划探索该模型在其他音色转换任务中的应用潜力，例如将人声转换为乐器音色，或与MIDI数码钢琴集成，进一步拓展基于扩散模型的音色转换模型在音乐生成领域的应用范围。

关键词：扩散架构、音色转换模型、钢琴版本、音高编码器、响度编码器、实时音乐转换、泛化能力。

1 引言

现代作曲家在音乐创作过程中，常常会在专辑中融入各种乐器。然而，不同乐器的录音过程复杂且耗时。使用软件音源库也需要较高的技能水平，因为其操作具有挑战性，最终成品往往在动态表现和音质方面不尽如人意。

钢琴音色的创作与使用同样具有挑战性。钢琴音色细腻多变，其多样的演奏技巧要求在钢琴声音的制作与使用过程中进行周密考量。在音色方面，钢琴家可以通过控制琴键敲击力度、踏板使用和速度变化来调节声音的丰富度与情感表达，使钢琴音色既能温柔细腻，也能强劲宏大，展现出高度的音乐表现力。在演奏技巧方面，钢琴包含多种富有表现力的方法，例如颤音、滑奏和和弦琶音。这些技巧使钢琴能够传达丰富的情感层次，并表现复杂的技术内容。例如，踏板技巧可以延长音符的持续时间，营造更加丰富的声音效果；而控制琴键敲击的力度与速度，则可以增强旋律线的层次感。然而，由于钢琴音色的细腻性以及演奏技巧的多样性，市面上的钢琴音源往往难以达到真实钢琴演奏的效果。

因此，为降低使用钢琴音色转换的门槛，并将目标音频转换为更接近真实钢琴的声音，我们旨在运用深度学习方法，将目标音乐转换为逼真的钢琴音色版本，从而提升音乐作品的质量。为此，我们设计了一种用于音色转换的自编码器架构。首先，将输入的目标音乐转换为与其特征相匹配的通用表示，同时降低其维度。然后，将这些低维特征用作无条件扩散模型的生成条件，由该模型将其转换为相应的钢琴声音。最后，用户可以使用训练好的模型输入任意 .wav 文件，并获得转换为钢琴声音的 .wav 文件，无需具备专业知识和技能，即可实现出色的钢琴音色表现。

2 相关工作

钢琴音色转换是一种将钢琴独特的声音特征转换为类似其他乐器声音的技术；随着深度学习的发展，这项技术在音乐技术领域受到广泛关注并得到应用。循环神经网络（RNN）和长短期记忆网络（LSTM）在钢琴音色转换中发挥着关键作用；这些模型擅长

处理钢琴音频的时间序列数据，捕捉钢琴声音随时间变化的复杂特征，并在转换过程中保留其细节和质感。例如，LSTM 模型擅长捕捉长期依赖关系，学习钢琴音色的时间动态特征，包括音符延续和力度变化；Qian Liu[1]（2023年）在一项关于基于深度学习的钢琴音乐审美情感表达的研究中展示了这一点。此外，门控循环单元（GRU）是 RNN 的一种变体，可进一步提高处理较长时间序列的效率和效果，使音色转换结果更加连贯自然。M. Sahaya Sheela、S. Surendran、P. Muthu Krishnam-mal、Vinothkumar K、Surendra Shetty、B. Venkataramanaiah [2]（2023年）在改进基于人工智能的钢琴音乐识别算法时提出了这一方法。自编码器，尤其是变分自编码器（VAE），能够将钢琴音频编码到低维潜在空间，再将其解码回音频数据，从而在实现音色转换的同时保留钢琴音色的核心特征。Shih-Lun Wu、Yi-Hsuan Yang [3]（2023年）在《MuseMorphose：使用 VAE 进行整曲及细粒度钢琴音乐风格转换》中提出了这一方法。

生成对抗网络（GAN）为实现高保真钢琴音色转换提供了一种创新方法。生成对抗网络通过生成器与判别器之间的对抗训练，使生成的钢琴音频样本更加逼真、自然，尤其适用于不需要配对训练数据的场景。正如 Lijun Zheng、Chenglong Li[4]（2024年）在基于生成对抗网络的实时情感钢琴音乐生成研究中提出的那样，这种技术能够在双向学习框架内实现钢琴音色与其他乐器音色之间的转换。此外，生成对抗网络还能通过多域转换技术同时处理多种乐器音色的转换，大幅提升模型的效率和可扩展性；Yigitcan Ö zer、Meinard Müller[5]（2024年）在利用音乐动机增强技术分离钢琴协奏曲声源的研究中，以及 Wenxuan Jiang、Yaqi Guo、Jiulian Li[6]（2023年）在基于卷积神经网络的数码钢琴音频信号特征识别方法研究中，都提出了相关应用。生成对抗网络的灵活性使其成为钢琴音色转换中的有力工具，并可与其他技术结合，进一步提升音色转换的保真度（High Fidelity，Hi-Fi）和细粒度。

未来的钢琴音色转换技术将致力于提升真实性、实现更平滑的转换过程，并拓展应用场景。研究人员正探索新的神经网络架构和深度学习算法，以提高音色转换的精度和效率。深度强化学习已应用于钢琴指法生成，例如 Ananda Phan Iman、Chang Wook Ahn[7]（2024年）提出了一种用于钢琴指法生成的无模型深度强化学习方法，并将其应用于复调音乐转录系统。Taehyeon Kim、Donghyeon Lee、Man-Je Kim、Chang Wook Ahn[8]（2024年）提出了一种利用音符间关系的复调钢琴音乐转录系统。这些应用展现了音色转换技术的潜力和发展方向。此外，研究人员正开发基于时频信息和卷积神经网络的技术，以改进钢琴音乐推荐算法。Sivagami. S、S. Ramani、Vishal Mehra、L. H. Jasim、M. Sahaya Sheela，

Kavitha S[9]（2023年）在一项关于基于时频的钢琴音乐推荐算法和特征识别方法的研究中提出了相关技术；Andres Fernandez[10]（2023年）则研究了使用卷积神经网络进行实时钢琴转录。随着技术不断发展，音色转换技术有望在音乐创作、教育和娱乐等领域发挥更大作用，为音乐表达和创作带来更多可能性。Ruimin Wu、Xianke Wang、Yuqing Li、Wei Xu、Wenqing Cheng[11]（2024年）提出了带谐波注意力的钢琴转录。

3 相关技术

我们在生成模型中使用的关键技术包括扩散模型、V-Diffusion、CREPE 和 U-Net，重点关注它们在数据生成与处理中的应用。

3.1 扩散模型

去噪扩散概率模型（DDPM）。DDPM 是一种通过逐步去除噪声来重建数据的生成模型。其核心思想是将数据生成视为一个反向过程，从随机噪声中逐步恢复原始数据。DDPM 的生成过程由一系列反向马尔可夫过程组成，模型在此过程中学习从高斯噪声生成逼真的数据样本。与其他生成模型相比，DDPM 在训练稳定性和生成样本质量方面具有显著优势。

去噪扩散隐式模型（DDIM）。DDIM 是对去噪扩散概率模型（DDPM）的改进，通过确定性生成过程降低计算负担。DDIM 减少了生成所需的步骤数并提高了效率，因此适用于需要快速生成高质量数据的场景。

3.2 V-Diffusion

V-Diffusion 是一种专注于音乐生成的扩散模型，通过逐步添加和移除噪声来模拟音乐分布，从而重建高质量的音乐片段。该模型的创新之处在于采用了可学习的噪声调度机制，增强了音乐细节和音质。

3.3 CREPE

CREPE 是一种用于音高估计的一维卷积神经网络。通过处理音频信号，该模型能够从原始音频数据中自动提取音高相关特征，在音乐分析中表现出色。

3.4 U-Net

U-Net 是一种用于图像分割的神经网络，通过收缩路径（编码器）和扩展路径（解码器）提取并还原图像特征。该模型的跳跃连接技术能够保留图像的空间信息，从而实现精确的图像分割。

4 方法架构

本文在下方的图1和图2中展示了模型架构，并将介绍下文提及的方法。

音高是音乐创作中的关键要素，确保音高随时间变化的准确性，对于音色转换的效果至关重要。为此，我们将音高信息作为输入条件，引导扩散模型的生成过程。音高编码器以需要转换的目标音频文件为输入。首先，我们使用 CREPE音高跟踪器确定输入音频波形的基频。接着，对估计出的基频进行词元化和离散化，帮助模型准确捕捉音频中对应的音高。将音频波形转换为对应的音高后，我们对结果进行独热编码，并通过全连接层将其转换为更复杂的向量表示。这一过程提升了模型在后续学习中的效果。

该方法确保音高信息的准确性和有效应用，在模型训练和转换过程中保留音乐的结构与音质。通过将音高作为关键控制因素，我们确保转换后的音频与原始版本音频之间的音高一致，从而保证转换结果的高质量与可靠性。

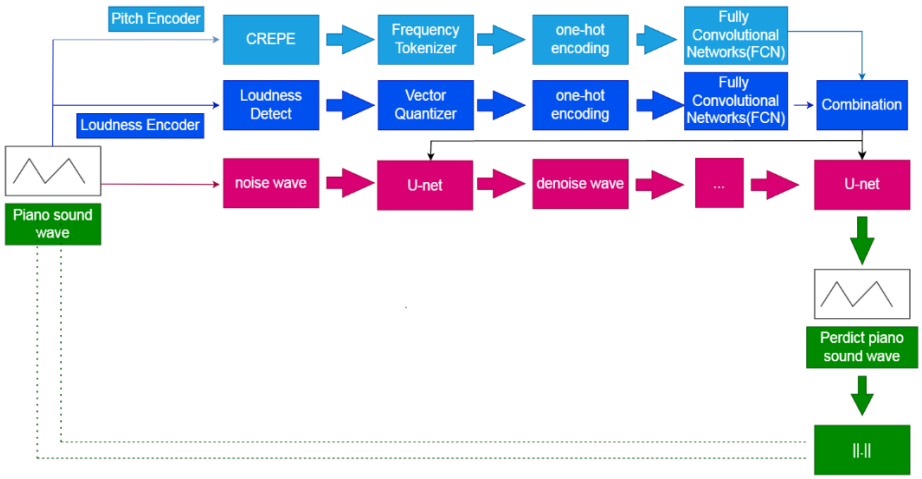


图 1. 模型训练流程图。

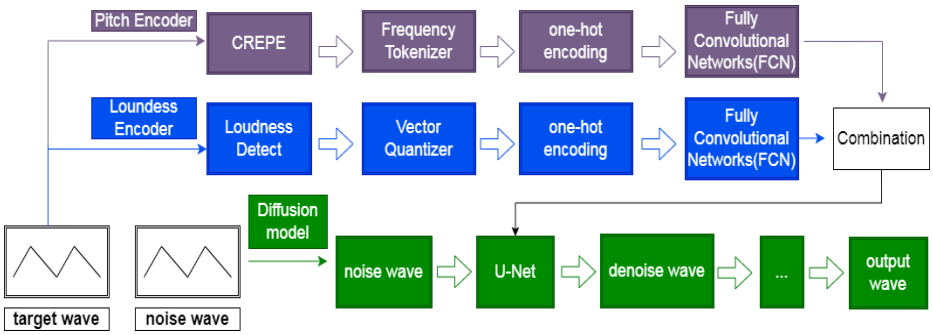


图2. 模型的完整架构。

嵌入是自然语言处理（NLP）中的一个基本概念。在自然语言处理中，嵌入指将词语或短语映射到低维向量空间中。这些向量能够捕捉特定的结构和语义关系，使模型能够识别词语之间的相似性与关联。借助嵌入，我们可以将文本转换为机器能够理解和处理的格式，并将这些向量表示应用于各种自然语言处理任务。嵌入的理论基础是分布假说，该假说指出，含义相似的词语往往出现在相似的语境中。基于这一假说，嵌入模型会学习将语义相关的词语映射到向量空间中彼此相邻的位置，从而有效地计算和比较它们之间的距离或相似度。

在我们的音色转换模型中，嵌入也发挥着至关重要的作用。我们将频率分词器得到的音高索引转换为相应的独热向量。这些独热向量具有高维、稀疏的特点，只有一个元素设为 1，其余元素均为 0。为了降低独热编码生成的这些高维稀疏向量的维度，我们使用全连接层将其转换为低维稠密向量，即学习得到的嵌入。这些稠密向量能够更有效地体现不同音高之间的相似性和差异，同时提高计算效率和可扩展性。

在音色转换过程中，音高索引的使用至关重要。我们从频率分词器中获取音高索引，并将其转换为相应的向量形式。表1展示了音高索引、对应的音高名称和基频（F0，单位为赫兹），提供了从 C4 到 B6 的音高信息，以及各自对应的索引和基频值。

表1. 基频区间参考表 le

索引	音高	F0（赫兹）	索引	音高	F0（赫兹）
0	None	0	19	F♯5/ G♭5	739.99
1	C4	261.63	20	G5	783.99
2	C♯4/ D♭4	277.18	21	G♯5/ A♭5	830.61
3	D4	293.66	22	A5	880.00

4	$D^{\#}4/ E^{\flat}4$	311.13	23	$A^{\#}5/ B^{\flat}5$	932.33
5	E4	329.63	24	B5	987.77
6	F4	349.23	25	C6	1046.50
7	$F^{\#}4/ G^{\flat}4$	369.99	26	$C^{\#}6/ D^{\flat}6$	1108.73
8	G4	392.00	27	D6	1174.66
9	$G^{\#}4/ A^{\flat}4$	415.30	28	$D^{\#}6/ E^{\flat}6$	1244.51
10	A4	440.00	29	E6	1318.51
11	$A^{\#}4/ B^{\flat}4$	466.16	30	F6	1396.91
12	B4	493.88	31	$F^{\#}6/ G^{\flat}6$	1479.98
13	C5	523.25	32	G6	1567.98
14	$C^{\#}5/ D^{\flat}5$	554.37	33	$G^{\#}6/ A^{\flat}6$	1661.22
15	D5	587.33	34	A6	1760.00
16	$D^{\#}5/ E^{\flat}5$	622.25	35	$A^{\#}6/ B^{\flat}6$	1864.66
17	E5	659.25	36	B6	1975.53
18	F5	698.46			

频率分词器在自然语言处理领域发挥着至关重要的作用，尤其是在深度学习领域，文本需要转换为数值形式，以便神经网络进行处理。频率分词器的主要用途是将输入文本拆分为更小的单元，例如词语、子词或字符，并将这些单元映射为特定的索引或词向量。这个过程会将文本字符串转换为模型能够理解的可变长度一维序列。

在我们的方法中，我们将 CREPE 预测的基频作为输入，并将其转换为对应的音高值。音高是一种感知属性，决定了人耳感知到的声音高低，这与声波的频率直接相关。通常，频率越高，音高越高；频率越低，音高越低。与基频相比，音高更直观地体现了人类对声音的感知，因此我们使用音高频率表将基频转换为对应的音高值。这减轻了模型的计算负担，使其能够专注于捕捉输入波形的节奏模式，并准确再现钢琴独特的音色。

考虑到钢琴音乐的典型音高范围跨越多个八度，我们聚焦于数据集中最常用的范围。我们根据预测的基频与该范围内音高频率的接近程度对其进行分类，再将这些分类组转换为索引，并按音高从低到高排列。这种分类方法使模型能够更有效地将音高信息融入转换过程，从而提升结果的整体质量和准确性。此外，我们将基频为 0 的情况设为索引 0，表示钢琴波形中的静音或非表现性片段，进一步完善了转换系统的设计。

除了音高，响度也是影响钢琴演奏的重要因素。响度描述人耳感知到的声音大小，是一种与声波声强相关的物理量。然而，由于人耳对不同频率的敏感度不同，即使声强相同，感知到的响度也会随频率而变化

。为了更准确地反映人类听觉感知，我们在模型中选择响度作为控制变量。

在响度编码器中，我们输入待转换的目标音频文件。为降低计算负载，我们首先将输入的钢琴声波划分为 16 个等长片段。这种分段方法不仅降低了计算复杂度，还能有效捕捉钢琴声波随时间变化的响度特征。通过这种方式，我们可以保留声波在各个时间点的细微变化，这对于后续的特征提取和分析至关重要。随后，我们根据 ITU-R BS.1770-4 标准测量每个片段的响度，并将最小响度值设为 -70 LKFS。与音高不同，响度无法像音符一样被直接表示或量化。因此，我们使用矢量量化模型识别音频中最常见的响度值，并将这些值转换为索引，从而实现进一步的数据压缩和特征简化。

为了有效整合音高特征，我们按比例缩放这些索引化的响度特征，使其在时间轴上的长度与音高编码器提取的音高特征相匹配，从而保持响度特征与音高特征之间的一致性。这增强了模型学习这些特征的效果。同样，这些经过压缩的索引化响度数据会被输入嵌入模型，以生成表现力更强的特征嵌入，帮助模型生成质量更高的钢琴音色。

在音色转换任务中，准确表示响度对于保持音乐的动态范围和情感表达至关重要。然而，连续的响度值会增加模型的计算负担，并对训练和收敛速度产生不利影响。为解决这些问题，我们采用了矢量量化（VQ），这是一种高效的数据压缩技术，可将连续响度数据向量转换为有限集合中的离散码，从而实现响度数据的压缩和编码。

矢量量化器的工作原理是对训练数据中的响度向量进行聚类，以生成具有代表性的码本。该码本包含各种响度特征的代表性信息，能够有效反映响度的分布特征和统计属性。在编码过程中，输入的响度向量会映射到码本中距离最近的码字，并以对应的码表示。这样，我们便可将连续响度数据转换为离散编码形式，不仅实现数据压缩和编码，还能有效降低存储和传输成本。

在从编码后的向量重建与原始输入相似的上采样结构时，我们借鉴了 Flavio Schneider 的架构设计[12] 。该设计提出使用扩散模型作为解码器，以生成高质量的钢琴音色。我们以 U-Net 作为扩散模型的骨干网络，自然而然地将自编码器功能与跳跃连接结合起来。编码器生成的潜向量被输入到 U-Net 的相应层级中，在这一框架中有效地充当“解码器”。与原始设计为适应不同波形长度而移除 U-Net 中自注意力机制的做法不同，我们选择保持输入和输出波形长度固定，同时保留 U-Net 中的自注意力机制。

这种方法使模型能够准确重建音高和响度编码器提取的特征之间的时间关系，让模型在减少对无关信息依赖的同时，专注于关键的音乐属性。自注意力机制使模型能够在生成过程中动态调整每个时间步的关注重点，从而有效捕捉乐曲中的关键细节。这种灵活性使模型能够适应不同的音乐风格和复杂程度，进而提升生成的钢琴音色的准确性与自然度。

音高编码器和响度编码器的输出经过合并后，作为条件输入传递给模型。合并后的条件输入被送入 U-Net 的下采样层，为生成音高和响度一致的钢琴音色提供必要的特征。

5 案例分析

5.1 数据集

案例数据集包含159首采用四四拍的江南风格中国传统音乐，总时长约为6919秒。这些曲目基于五声音阶，缺少 Fa 和 Si 两个音符。音乐曲目以44100赫兹的采样率录制，并分割为6秒的片段。下图和表格（见图3、表2、表3）分别展示了数据集中音高和节拍时长的信息。

在表3中，若拍号为四四拍且速度为每分钟60拍，则每小节时长为2秒，每个四分音符时长为0.5秒。不过，为了方便程序处理，时间通常以时钟刻度表示。本研究将一个四分音符的时长定义为固定的480个时钟刻度，因此每小节为1920个时钟刻度，每秒为960个时钟刻度。无论速度如何变化，该数值都保持不变。

在传统江南音乐中，节奏结构以四分音符和八分音符为主。鉴于这种对四分音符和八分音符的侧重，为了准确呈现音乐，1920 种可用的时钟刻度值并非必需。相反，我们只关注 10 种特定的时钟刻度值，它们最常出现在江南音乐的节奏模式中。通过聚焦这 10 种时钟刻度值，我们简化了分析，并确保模型能够准确反映江南音乐的节奏倾向，同时避免受到那些在其结构中并不起重要作用的多余时钟刻度值的干扰。

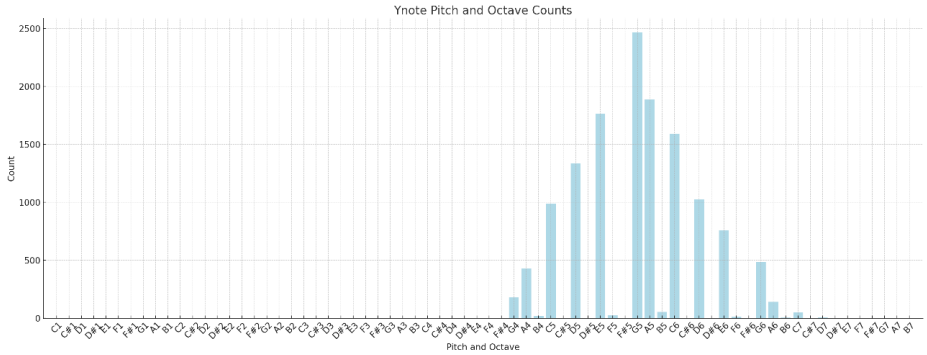


图3. 数据集的音高统计（C1 至 B7）

表2. 数据集的音高统计（C1 至 B7）

音高_0ct	计数	音高_0ct	计数	音高_0ct	计数
E4	2	F5	24	F6	10
G4	178	G5	2465	G6	485
A4	429	A5	1887	A6	141
B4	20	B5	54	B6	4
C5	987	C6	1591	C7	49
D5	1337	D6	1027	D7	5
E5	1763	E6	758		

表3. 数据集中音符时值的统计分析

Tick	音符时值	计数
180	附点十六分音符	75
240	八分音符	4813
360	附点八分音符	531
480	四分音符	1017
720	附点四分音符	209
960	二分音符	572
1440	附点二分音符	20
1920	全音符	16
3840	倍全音符	70
160	四分音符三连音	8

5.2 训练

我们从159个训练样本的原始版本波形中随机选取片段，每个片段由2^{18} 帧组成，作为模型的输入。训练在 NVIDIA RTX 4090 GPU 上进行，批量大小为8，共训练6000个周期，总耗时4天。

5.3 评估

下面的讨论分为四种情况：

1. 音高处于表2的范围内，且音符时值处于以下范围内
   表3
   T左侧图4左侧显示了小提琴演奏《一闪一闪小星星》的原始录音频谱，右侧显示了模型以小提琴演奏为输入生成的钢琴音版本的频谱。图5对比了可听频率范围内的两个频谱，结果显示它们的频谱特征高度相似。这表明，该模型成功生成了与真实钢琴声音非常相似的钢琴音色，同时保持了与原始小提琴版本相似的频率分布特征。这些结果表明，该模型在指定音高范围（见表2）和音符时长范围（见表3）内表现良好。

示例1：《一闪一闪小星星》小提琴演奏版。

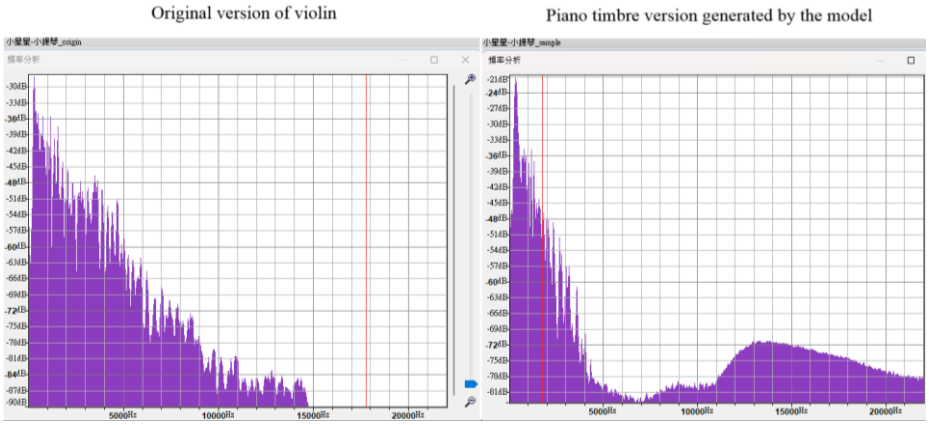


图4. 左侧为小提琴演奏《一闪一闪小星星》的响度分布，其中显示了基频和谐波峰；右侧为模型以小提琴演奏为输入生成的钢琴音色。

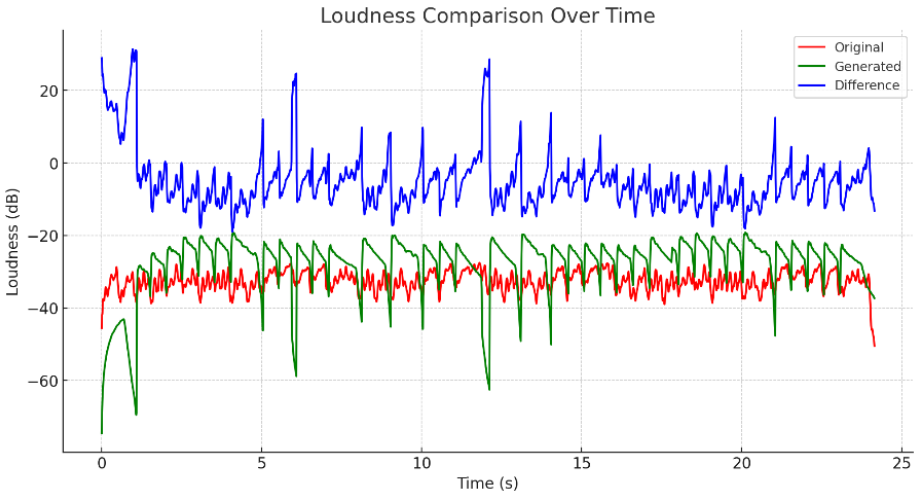


图5。原始小提琴音频（红线）、钢琴音色版本（绿线）及其差异（蓝线）随时间的响度变化，揭示了两个版本之间的时间响度差异。

我们探索了不同的音乐和乐器作为模型的输入，例如示例2中的 Taiho Boat 长笛演奏版，以及示例3中的 Jasmine 低音提琴演奏版。随后，模型生成了这些曲目的钢琴版本。如图7和图9所示，原始版本与生成版本在响度随时间的变化和频谱特征方面差异甚微。这表明，生成的钢琴版本的频谱属性与实际乐器的原始音频高度相似。这些结果进一步表明，模型成功合成了指定音高范围（见表2）和音符时长范围（见表3）内的钢琴声音；这些声音与真实钢琴音频十分相似，同时保持了与原始乐器版本相近的频率分布，从而证明了模型的有效性。

示例_2：Taiho Boat 由长笛演奏。

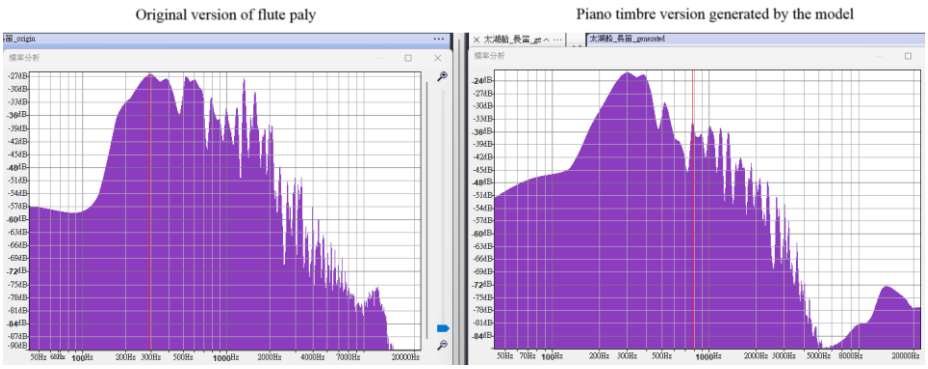


图6。右侧是长笛演奏的原始版本，左侧是以长笛演奏音频为输入、由模型生成的钢琴音色版本。

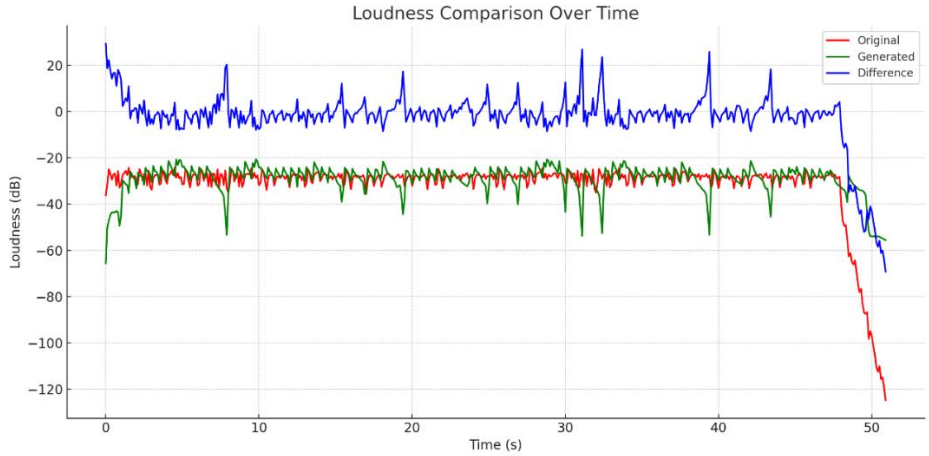


图7。原始长笛音频（红线）、生成的钢琴音色版本（绿线）及二者差值（蓝线）的响度变化随时间的变化表明，两个版本之间的时间响度差异相似。

示例_3：Jasmine 由低音提琴演奏。

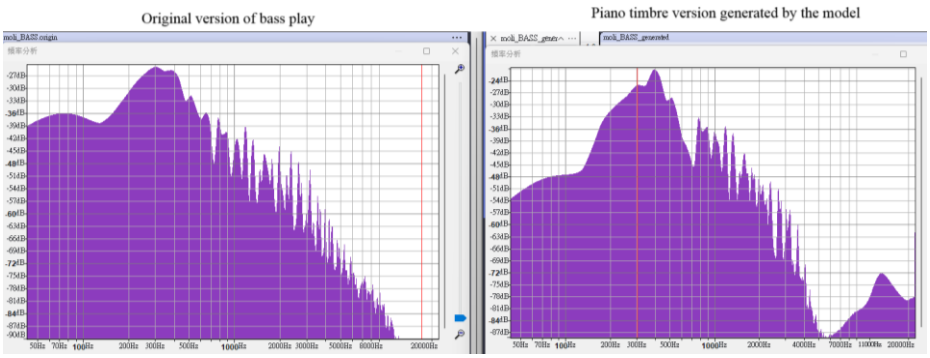


图8。右侧是低音提琴演奏的原始版本，左侧是以低音提琴演奏音频为输入、由模型生成的钢琴音色版本。

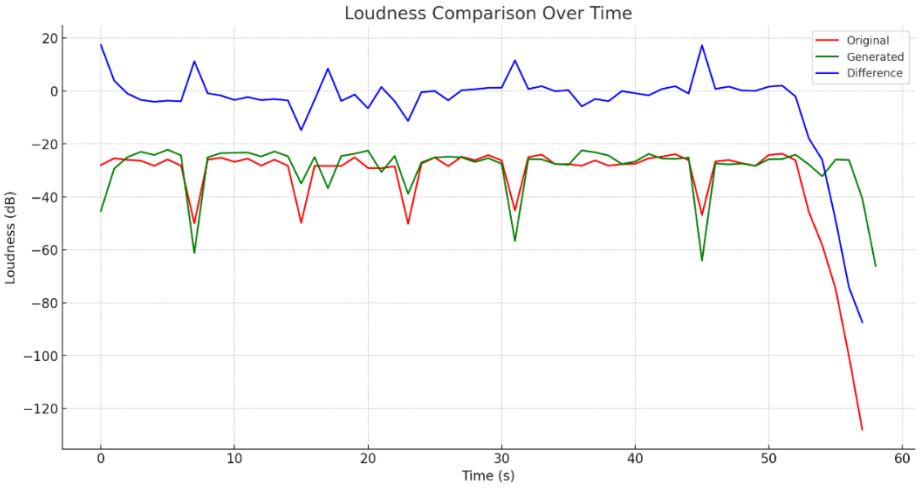


图9. 原始长笛音频（红线）、生成的钢琴音色版本（绿线）及其差值（蓝线）随时间的响度变化表明，两个版本的频谱特征高度相似。

2. 音高降低一个八度，且音符时值在表3的范围内：
图10左侧展示了小提琴低一个八度演奏的《一闪一闪小星星》原始版本，右侧则展示了模型以低八度版本为输入生成的钢琴音色版本。对比图11中的两组频谱可以发现，在可听频率范围内存在显著差异。尽管音高降低了一个八度，且音符时长保持不变，但与小提琴原版相比，模型生成的钢琴音色在低频范围内能量不足。这种差异可能是由于C1至C3八度范围内的训练数据相对缺乏所致。

根据图3和表2的数据分析，训练数据的分布显著影响了模型生成钢琴音色的能力，尤其是在低频区域。该频段数据的稀缺削弱了模型准确再现低音区域音色的能力。因此，这一差异凸显了训练数据分布对模型生成性能的关键影响，尤其是在再现低音音域音色时。

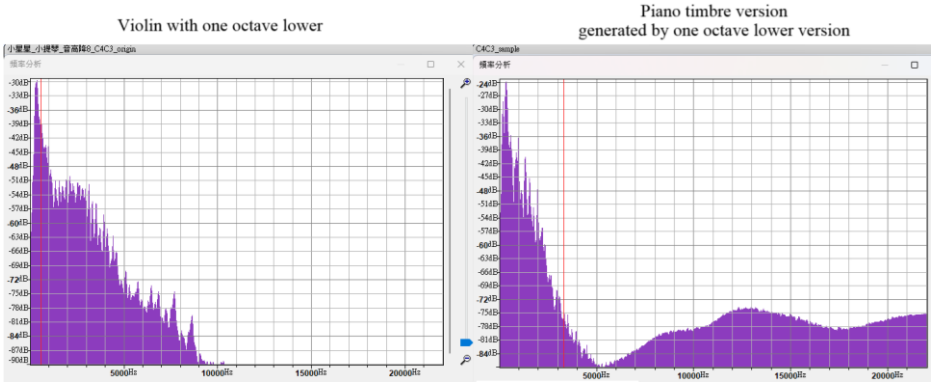


图10。左侧展示小提琴低八度版本的响度分布，右侧展示模型生成的钢琴音色的响度分布。

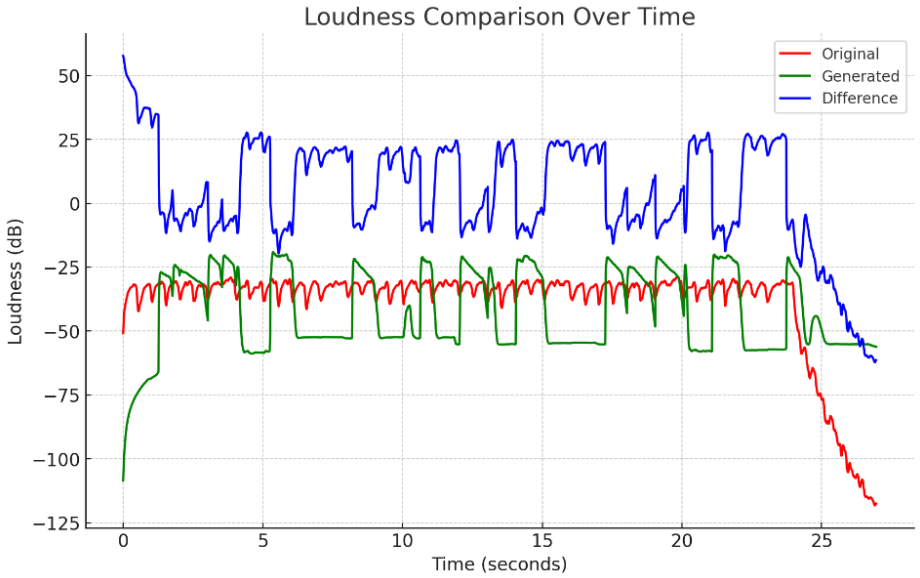


图11。比较原始版本与生成音频之间的响度变化，可见低频范围存在显著差异。这些图表凸显了训练数据的分布在音色生成过程中的重要性，尤其是在处理低频区域时。

3. 音高在表2规定的范围内，音符时值为八分音符时值——
tion:

图12分别展示了《一闪一闪小星星》中的小提琴演奏和钢琴版本，模型生成的《小星星》。两个版本的八分音符音高相同。从图13可以看出，在可听频率范围内，两个版本呈现出高度相似的分布。这表明

转换模型有效保留了原始音乐片段的音色特征，并成功模拟了钢琴音色。图13所示的相似性进一步证实了该模型在不同乐器之间进行音色转换时的可靠性和有效性。

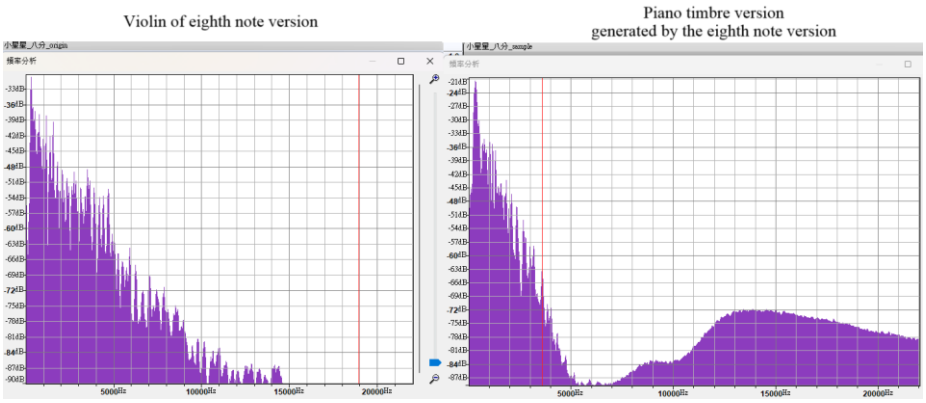


图12. 分别展示模型生成的《一闪一闪小星星》小提琴演奏和钢琴版本。

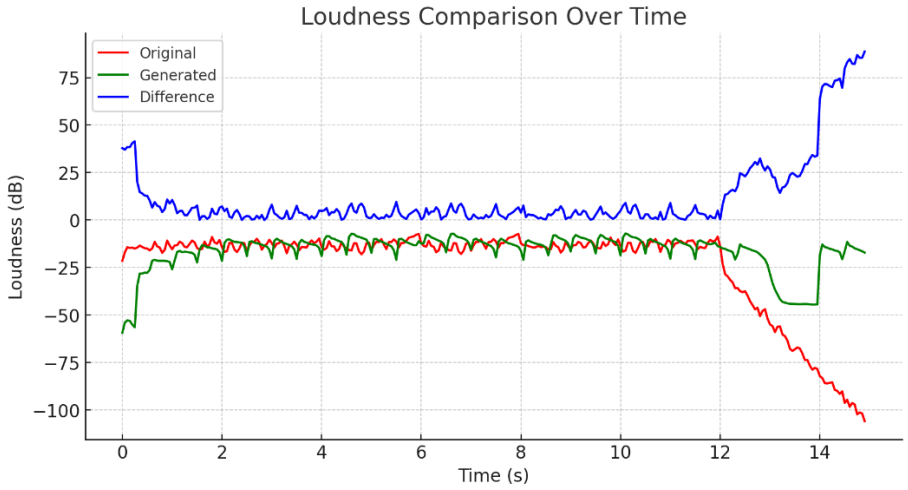


图13. 原始小提琴版本（红线）与生成的钢琴版本（绿线）的响度对比，蓝线表示两者之间的差异。x 轴表示时间（秒），y 轴表示响度（分贝）。两个版本的响度随时间变化高度相似，仅有细微偏差。

4. 音高降低一个八度，音符时值为三十二分音符：图14左侧展示了小提琴演奏的《一闪一闪小星星》原始版本，其音高降低了一个八度，音符时值为三十二分音符

节奏。频谱显示，能量主要集中在较低频率范围，尤其是低于 1000 赫兹的区域，呈现出显著的低频特征。图14右侧展示了模型生成的钢琴版本。频谱显示其频率分布更广，在较高频率范围内的能量响应相对稳定，表明生成的钢琴版本的能量分布在整个频谱范围内更为广泛。图15显示，响度差异凸显了两个版本之间的显著差异，尤其是动态范围方面。

这些差异可能归因于训练数据集存在不足。根据图3和表2中的分析，训练数据在C1至C3音域内缺少足够的样本，导致生成版本在低频段的表现较弱。此外，数据集中不包含三十二分音符节奏的示例，这进一步影响了生成音乐在节奏和时间特征方面的准确性。

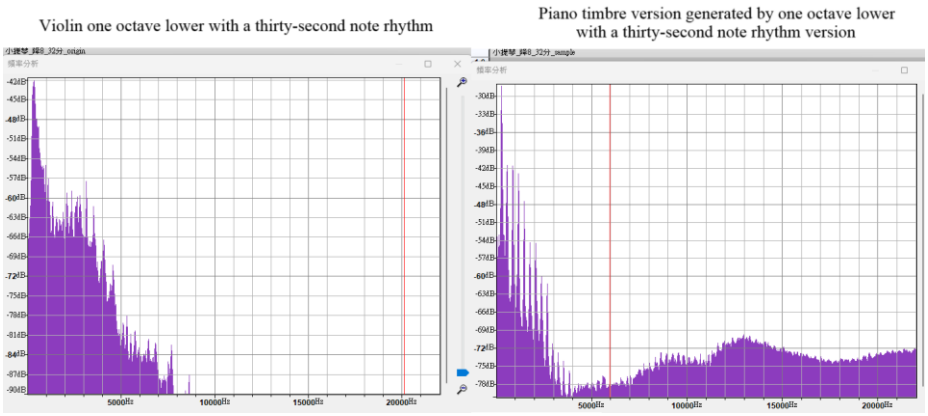


图14。左侧显示由小提琴演奏的原始版本，以低八度和三十二分音符节奏呈现；右侧展示模型生成的钢琴版本。

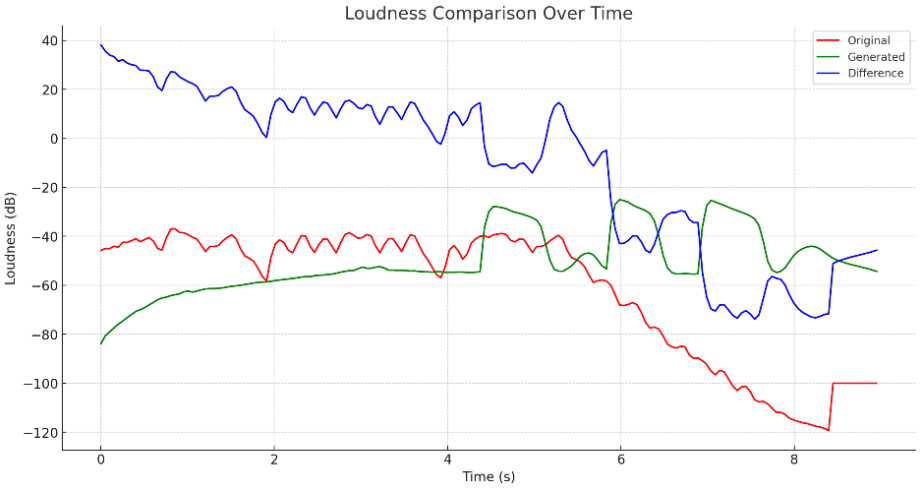


图15. 比较原始版本（红线）与生成版本（绿线）之间的响度差异（蓝线）。

6 结论

本文提出了一种基于扩散架构的音色转换模型，可将任意音乐转换为钢琴音色。该模型首先从音乐演奏中提取通用特征，然后将这些特征输入扩散模型解码器，以重建相应的钢琴演奏。在项目分析阶段，初始训练数据集由中国传统江南风格音乐组成，这类音乐主要基于五声音阶，不包含Fa和Si音高。因此，数据集并不完整，这可能影响了模型的表现。为解决这一局限，我们设计了多项测试来评估模型的表现，尤其关注音高准确度。结果表明，该模型在保持音高准确度方面表现良好，并能够生成与真实钢琴音色十分相似的声音。尽管数据集并不完整，该模型仍有效地捕捉了输入音乐的基本特征，并成功将其转换为连贯的钢琴输出。

参考文献

1. Qian Liu，《基于深度学习的钢琴音乐审美情感表达研究》，2023年第五届国际应用机器学习会议（ICAML），2023年。2. M. Sahaya Sheela、S. Surendran、P. Muthu Krishnammal、Vinothkumar K、Surendra Shetty 和 B. Venkataramanaiah，《基于人工智能的改进型钢琴音乐识别算法》，2023年第二届智能技术促进智慧国家国际会议（SmartTechCon），2023年。

3. Shih-Lun Wu 和 Yi-Hsuan Yang，《MuseMorphose：使用一个Transformer变分自编码器进行整曲与细粒度钢琴音乐风格迁移》，IEEE/ACM《音频、语音与语言处理汇刊》，第31卷，2023年。4. Lijun Zheng 和 Chenglong Li，《使用生成对抗网络（GAN）进行实时基于情感的钢琴音乐生成》，IEEE Access，第12卷，2024年。5. Yigitcan Ö zer 和 Meinard Müller，《使用基于音乐特征的数据增强技术对钢琴协奏曲进行源分离》，IEEE/ACM《音频、语音与语言处理汇刊》，第32卷，2024年。6. Wenxuan Jiang、Yaqi Guo 和 Jiulian Li，《MuseMorphose：使用一个Transformer变分自编码器进行整曲与细粒度钢琴音乐风格迁移》，IEEE/ACM《音频、语音与语言处理汇刊》，第31卷，2023年。7. Ananda Phan Iman 和 Chang Wook Ahn，《一种用于钢琴指法生成的无模型深度强化学习方法》，2024年 IEEE人工智能会议（CAI），2024年。8. Taehyeon Kim、Donghyeon Lee、Man-Je Kim 和 Chang Wook Ahn，《利用不同音符状态之间的相互关联性的复调钢琴音乐转录系统》，IEEE Access，第12卷，2024年。9. Sivagami. S、S. Ramani、Vishal Mehra、L. H. Jasim、M. Sahaya Sheela 和 Kavitha S，《基于时频信息的钢琴音乐推荐算法研究》，2023年第二届智能技术促进智慧国家国际会议（SmartTechCon），2023年。10. Andres Fernandez，《起音与力度：使用卷积神经网络实现经济实惠的实时钢琴转录》，2023年第31届欧洲信号处理会议（EUSIPCO），2023年。11. Ruimin Wu、Xianke Wang、Yuqing Li、Wei Xu 和 Wenqing Cheng，《带谐波注意力的钢琴转录》，ICASSP 2024——2024年 IEEE国际声学、语音与信号处理会议（ICASSP），2024年。12. Flavio Schneider，《Archisound：使用扩散模型进行音频生成》，arXiv预印本 arXiv:2301.13267，2023年。