用于文本引导音效生成特征增强扩散模型

苗向阳

湖南工商大学　湖南长沙

【摘要】音效在游戏、电影和虚拟现实等领域具有重要作用，它通过声音描述事件的发生，增强听众沉浸感。随着深度学习发展和大语言模型的出现，音效生成技术迎来革命性突破，特别是基于文本引导的音效生成技术，该技术通过文本描述就可以自动生成符合场景的音效。然而，现有生成模型和方法仍存在生成音频逼真度欠缺、文本音频相关度低等问题。本文针对这些问题提出了一种新型特征增强扩散模型（Feature Enhanced Diffusion Model, FEDM），（1）采用 Haar 小波变换进行下采样，有效保留高频特征信息；（2）设计多尺度特征提取模块，通过不同尺寸卷积核捕捉多层次特征。实验结果表明，所提方法在 AudioCaps 数据集上的 FAD 和 KL 指标上比基线模型提升了 33.3%和 18.1%。

【关键词】音效生成；文本引导；扩散模型；小波变换；多尺度提取

【基金项目】湖南省研究生科研创新项目（CX20231163）

【收稿日期】2025 年 3 月 15 日　【出刊日期】2025 年 4 月 16 日　【DOI】10.12208/j.aics.20250004

It is used to generate feature enhancement diffusion model for text-guided sound effects

Xiangyang Miao

Hunan University of Commerce, Changsha, Hunan

【**Abstract**】Sound effects play a crucial role in games, films, and virtual reality, enhancing the immersion of listeners by describing events through sound. With the development of deep learning and the emergence of large language models, sound effect generation technology has seen revolutionary advancements, particularly text-guided sound effect generation techniques, which can automatically generate sounds that match the scene based on textual descriptions. However, existing generation models and methods still suffer from issues such as insufficient audio realism and low relevance between text and audio. This paper proposes a novel feature-enhanced diffusion model (Feature Enhanced Diffusion Model, FEDM) to address these problems: (1) it uses Haar wavelet transform for downsampling, effectively retaining high-frequency feature information; (2) it designs a multi-scale feature extraction module to capture multi-level features through different-sized convolutional kernels. Experimental results show that the proposed method improves FAD and KL metrics by 33.3% and 18.1%, respectively, over the baseline model on the AudioCaps dataset.

【**Keywords**】Sound effect generation; Text guidance; Diffusion model; Wavelet transform; Multi-scale extraction

1 引言

1.1 研究背景

随着数字媒体技术的快速发展，音效在游戏、电影、虚拟现实等领域的应用日益广泛。传统音效制作主要依赖专业团队手工设计和录制，存在成本高、效率低、灵活性不足等问题[1]。近年来，深度学习技术在音频生成领域取得了显著进展，特别是基于文本引导的音效生成技术，能够根据自然语言描述自动生成符合场景需求的音效[2]，为音效制作带来了革命性的变革。

然而，现有的文本引导音效生成模型仍面临一些挑战。首先，在特征提取过程中，传统的卷积下采样操作会导致高频信息丢失，影响生成音效的清晰度和真实感。其次，多尺度特征融合机制不够完善，难以充分捕捉音频信号的多层次特征，导致生成音效与文本描述的匹配度不足。这些问题限制了文本引导音效生成技术的实际应用效果。

作者简介：苗向阳（1998-）男，汉族，山东滨州，硕士研究生，主要研究方向为智能语音处理。

1.2 研究现状

当前主流的音效生成方法主要基于生成对抗网络（GAN）、变分自编码器（VAE）和扩散模型。其中，扩散模型因其稳定的训练过程和高质量的生成效果，逐渐成为音频生成领域的主流方法。Yang 等人[3]提出的 Diffsound 模型首次将 VQ-VAE 与离散扩散过程相结合，实现了文本到音频的生成。Liu 等人[4]提出的 AudioLDM 模型则在潜在空间进行扩散，进一步提高了生成效率。

在特征提取方面，传统方法主要采用卷积神经网络进行下采样，但这种方法会丢失高频信息。小波变换因其良好的时频局部化特性，在图像处理领域已证明能有效保留高频细节，但在音频生成领域的应用研究相对较少。多尺度特征融合方面，注意力机制已被广泛应用于各种深度学习任务，但在音效生成中的特征融合优化仍有探索空间。

2 相关理论与技术

2.1 文本引导音效生成框架

当前主流的文本引导音效生成框架主要包括 Diffsound 和 AudioLDM 两种。Diffsound 框架由文本编码器、矢量量化变分自动编码器、解码器和声码器组成，其工作流程为：文本编码器提取文本特征，通过令牌解码器解码出码本序列，再经 VQ-VAE 转换为梅尔频谱图，最后使用声码器恢复波形。

AudioLDM 框架则基于去噪扩散隐式模型（LDMs）和对比学习，利用 CLAP 模型对齐文本和音频的嵌入空间，在潜在空间进行扩散过程。相比 Diffsound，AudioLDM 减少了对音频-文本数据对的依赖，且对计算资源要求更低。

2.2 扩散模型基本原理

扩散模型包括正向过程和逆向过程两个阶段[5]，正向过程通过马尔可夫链逐步向数据添加噪声，将原始数据分布转化为高斯分布，逆向过程则学习从噪声中逐步恢复数据。

2.2.1 正向过程

在正向过程中，生成扩散模型将原始数据逐步添加噪声，将数据从清晰的状态x_0转换成接近高斯噪声的状态x_T。该过程通常是一个马尔可夫链，在每个时间步t上会将一定量的高斯噪声加入到数据中。具体而言，对于时间步t，正向扩散过程可以表示为：

$$q\left(x_t \middle| x_{t-1}\right) = N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \quad (2\text{-}1)$$

其中β_t是在第t个时间步的噪声添加系数，一般设置为一个随t增加的逐渐上升的数值。$\sqrt{1-\beta_t}$用于保持数据的主要信息，并且对每一步的扰动较小。$\beta_t \mathbf{I}$是噪声项，其中I是单位矩阵，表示对所有维度均等添加的独立高斯噪声。

通过递归地应用上述公式，可以得到一个从x_0到x_T的联合分布：

$$q\left(x_{1:T} \middle| x_0\right) = \prod_{t=1}^{T} q\left(x_t \middle| x_{t-1}\right) \quad (2\text{-}2)$$

通过这一过程，初始数据x_0会逐渐变为高斯噪声x_T。在扩散过程中，随着时间步增加，数据中的原始信息逐渐丧失，而噪声强度逐渐增大。

2.2.2 逆向过程

逆向过程的目标是从纯噪声逐步去除噪声，恢复到原始数据的分布。这个过程与正向过程相反，将噪声中的信息逐步提取出来，因此也是一个马尔可夫链，每一步去除一定量的噪声。逆向过程可以表示为：

$$p_\theta\left(x_{t-1} \middle| x_t\right) = N\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \quad (2\text{-}3)$$

其中$\mu_\theta(x_t,\ t)$表示时间步t的均值参数，由神经网络学习得到。$\Sigma_\theta(x_t,\ t)$表示时间步t的方差参数，也可以是可学习参数。

为了训练模型，需要最大化逆向过程的似然函数，或者等价地最小化其负对数似然。在训练过程中可以通过变分推断方法将对数似然目标分解为每个时间步的 KL 散度之和：

$$L = \mathsf{E}_{q(x_0, x_{1:T})}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(q\left(x_{t-1} \middle| x_t, x_0\right) \middle\| p_0\left(x_{t-1} \middle| x_t\right)\right)\right] \quad (2\text{-}4)$$

其中，D_{KL}表示 KL 散度，$q(x_{t-1}|x_t,\ x_0)$是正向过程中的真实条件分布，$p_0(x_{t-1}|x_t)$是模型预测的分布。

在实际应用中，扩散模型的损失函数通常简化为预测噪声的均方误差（MSE）损失：

$$L = \mathsf{E}_{q(x_0, x_{1:T})}\left[\left\|\varepsilon - \varepsilon_\theta\left(x_t \middle| t\right)\right\|^2\right] \quad (2\text{-}5)$$

其中是正向过程中添加的噪声，$\varepsilon_\theta(x_t|t)$是神经网络预测的噪声。

3 特征增强扩散模型设计

3.1 基于 U-Net 的扩散模型网络框架

本文设计了一个小波变换下采样模块和多尺度注意力融合新的特征增强网络架构，应用于基于 U-Net 的潜在扩散模型，将其命名为特征增强扩散模型（Feature Enhanced Diffusion Model，简称 FEDM），总体框架简图如图 3-1 所示。

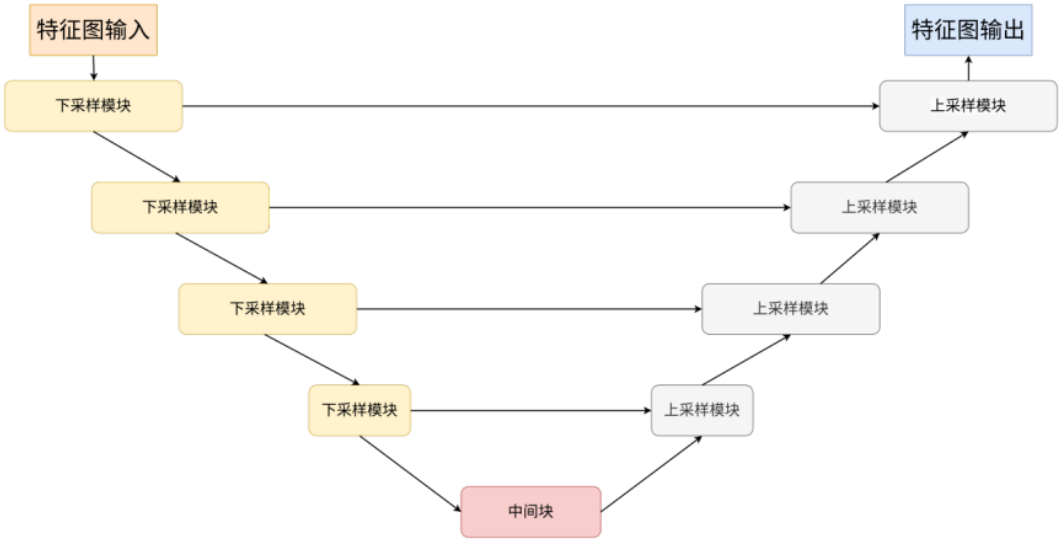

图 3-1 FEDM 总体框架简图

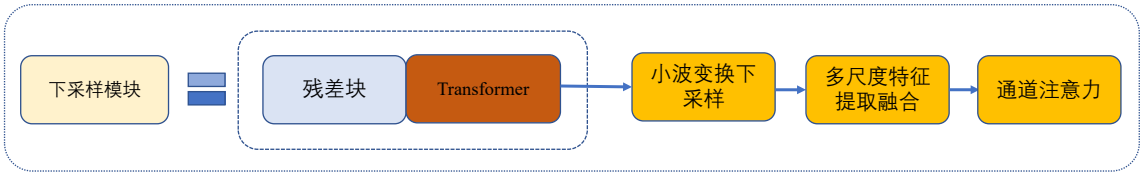

图 3-2 下采样模块具体结构

图 3-1 中每层下采样模块都包括 5 个部分，具体结构如 3-2 所示。

其中残差块原本文本引导音效生成模型中的结构，主要利用跳跃连接结构解决解决深层网络训练时的梯度消失和爆炸问题，Transformer 也是模型中的主要结构，其中的交叉注意力是文本和音效特征的交互，目的是实现文本引导的条件生成。

本文方法是受到视觉处理领域中小波变换和注意力机制的启发，借鉴图像处理中的成功应用案例，设计了一种新性特征增强网络架构应用于扩散模型。新架构采用 Haar 小波变换进行下采样，降低特征图空间分辨率的同时，尽可能保留原始音频信息的细节，该策略源于小波变换在图像处理中的优势，即能够在下采样过程中有效地保留高频信息，从而减少音效生成过程中的信息丢失，避免噪声的产生。

3.2 小波变换下采样

图 3-2 中的小波变换下采样主要过程包括对输入特征图进行 Haar 小波变换降低特征图的空间分辨率，同时原始特征图被分解为四个通道，分别是低频子带（LL）、水平高频子带（LH）、垂直高频子带（HL）、对角线高频子带（HH）。然后将低频子带和三个高频子带 HL、LH、HH 沿着通道维度拼接起来，将不同频率和方向的信息整合到一起形成一个新的特征图。最后通过一个 1x1 的卷积层，减少特征图的通道数，并进行特征的重新组合，完成下采样操作。

首先对输入特征图进行离散小波变换，将特征图分解为低频和高频部分，公式如下：

$$f(t)=\sum_{k}\psi_{k}(t)\cdot f(t) \qquad (3\text{-}1)$$

其中，$\psi_k(t)$是小波函数，$f(t)$是输入特征图。

然后将低频子带和三个高频子带沿着通道维度拼接，将不同频率和方向的信息整合成一个新的特征图。然后通过一个卷积层，减少特征图的通道数，再进行归一化：

$$\hat{x}=\frac{x-\mu}{\sigma}\ \gamma+\ \beta \qquad (3\text{-}2)$$

其中，μ是输入特征的均值，σ是标准差，γ和 β是可学习的缩放和平移参数，$\hat{x}$是归一化后的输出。最后利用 ReLU 激活函数增强网络的表达能力。

3.3 损失函数设置

在深度学习模型中，损失函数（Loss Function）是

用于衡量模型预测与实际标签之间差距的函数，优化模型时损失最小化，对模型的训练和性能有着重要作用，并且不同任务和模型类型会使用不同的损失函数。本文在广义的音频生成模型损失函数的基础上进行了优化，用以下损失函数进行训练：

（1）对比损失（Contrastive Loss），用于 CLAP 预训练模型，目的是使得匹配的文本和音频的表示更接近，而不匹配的文本和音频的表示更远离，使用 InfoNCE 损失函数[6]，公式为：

$$L_{contrast} = -\log \frac{\exp\left(\frac{\mathrm{sim}\left(a^{+}, t^{+}\right)}{\tau}\right)}{\sum_{i=1}^{N} \exp\left(\frac{\mathrm{sim}\left(a^{+}, t_i^{-}\right)}{\tau}\right)} \quad (3\text{-}3)$$

其中a^{+}和t^{+}是正样本对，t_i^{-}是负样本，sim是相似度函数（如余弦相似度），τ是温度参数。

（2）扩散过程损失（Diffusion Loss），在潜在扩散模型的训练中，最关键的部分是通过逆向扩散过程来生成音频特征，目标是最大化训练过程中的数据重构，同时最小化生成潜在表示的差异。扩散损失通常是通过优化生成的潜在表示与真实潜在表示之间的差异来实现的[7]。

$$L_{diffusion} = E_{q(z_t|z_0)}\left[\| z_0 - p_\theta(z_0 \mid z_t)\|_2^2\right] \quad (3\text{-}4)$$

（3）直构损失（Reconstruction Loss），用于计算输入样本和重构的梅尔频谱图之间的差异，使用平均绝对误差（Mean Absolute Error， MAE）或均方误差（Mean Squared Error，MSE）来计算[8]。公式可以表示为：

$$L_{recon} = \frac{1}{N}\sum_{i=1}^{N} | x_i - \hat{x}_i | \quad (3\text{-}5)$$

其中x_i是原始梅尔频谱图的第i个元素，$\hat{x}_i$是重构的梅尔频谱图的第i个元素，N是元素总数。

4 实验设计

4.1 实验设置

（1）实验平台

实验采用 PyTorch 的深度学习框架，以及 Ubuntu 18.04 LTS 操作系统。硬件使用含 15 个虚拟 CPU 的 Intel（R） Xeon（R） Platinum 8358P 处理器和 Nvidia A40 显卡，分别具有 80GB 的内存和 48GB 的显存。

（2）数据集

实验使用 AudioCaps 数据集[9]，包含约 50，000 个音频剪辑及其文本描述。数据集划分为训练集、验证集和测试集，比例分别为 80%、10%和 10%。

（3）评估指标

遵循 Audio Gen 的评估协议[10]，该协议计算两个客观指标，包括 Frechet Audio Distance （FAD）、Kullback-Leibler Divergence（KL），其中 FAD 是一种无参考音频质量度量，是根据从 VGGish 模型中提取的目标特征与生成音频之间的分布距离来计算的，KL 散度通过音频标记模型 Patch-out Transformer 计算的标签来衡量生成的音频和目标音频之间的相似性，其方式与 AudioGen 相同，指标都是越小越好。

4.2 对比实验结果与分析

为了验证本研究提出的基于小波变换和多尺度特征提取注意力融合的增强扩散模型的有效性，本文通过客观评价进行评估。

表 3-1 本文模型和其他对比结果

Model	Params	训练时间（h）	FAD	KL
DiffSound	400M	5420	7.75	2.52
Audio Gen	285M	8067	3.13	2.09
Make-an-Audio	453M	3000	2.66	1.61
AudioLDM-L	739M	145	2.08	1.86
AudioLDM-S	181M	145	2.43	1.97
FEDM	296M	145	1.6214	1.6136

在实验中，音频生成模型的性能从多个方面进行了对比分析，其中，模型的参数数量、训练时长以及 FAD 和 KL 指标是评估模型性能的重要因素。本文模型（FEDM）在这几个因素方面都表现出色，其中参数数量 296M，训练时间 145 小时，在较低的参数和训练时间下依旧具有最低的 KL 值最低的 FAD，这两个指标是衡量两个分布之间的相似度，越低表示模型性能更好。

5 结论与展望

本文提出了一种基于特征增强的文本引导音效生成扩散模型 FEDM，通过小波变换下采样和多尺度特征提取，有效解决了传统方法中高频信息丢失和特征表达能力不足的问题。实验证明，该方法显著提升了音效生成质量和文本相关性。本研究为文本引导音效生成提供了新的技术思路，有望推动智能音效生成技术的进一步发展与应用。

参考文献

[1] 刘天羽. 物理建模合成在游戏音效制作中的应用研究——以水流声合成为例[J]. 电声技术, 2022, 46 (11): 45-48.

[2] 王珏,李洽楠.AI 音频技术在电影对白和音效制作中的应用探究[J].现代电影技术,2024,(12):13-21.

[3] Yang D, Yu J, Wang H, et al. Diffsound: Discrete diffusion model for te-xt-to-sound generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1720-1733

[4] Liu H, Chen Z, Yuan Y, et al. Audioldm: Text-to-audio generation with latent diffusion models[J]. arXiv preprint arXiv:2301.12503, 2023.

[5] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International conference on machine learning. pmlr, 2015: 2256-2265.

[6] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv: 1807.03748, 2018.

[7] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[8] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[9] Kim C D, Kim B, Lee H, et al. Audiocaps: Generating captions for audios in the wild[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 119-132.

[10] Kreuk F, Synnaeve G, Polyak A, et al. Audiogen: Textually guided audio generation[J]. arXiv preprint arXiv: 2209.15352, 2022.