语音与音频领域中的人工智能生成与检测技术综述

金天昊 [1]

(1 中国科学技术大学 少年班学院，安徽 合肥，中国)

摘 要：近年来，人工智能生成内容 (AI Generated Content, AIGC) 技术在语音与音频领域取得了显著发展，并在多个领域（如人机交互、内容创作、娱乐产业等）展现出巨大应用潜力。然而，随着 AIGC 技术发展而来的潜在风险和挑战也日益凸显。尤其是语音深度伪造 (Voice Deepfake) 技术可能会被用于诈骗活动等违法行为，对个人隐私、社会信任和网络安全造成严重威胁。AI 生成内容的版权和原创性也容易引发争议。因此，研究和发展 AI 生成音频检测技术来应对这些挑战，具有相当的重要性和紧迫性。本文旨在对语音与音频中的 AIGC 及检测技术进行系统性、深入性的综述。本文将内容根据历史发展顺序划分为机器学习时期、早期深度学习时期，以及现代深度学习时期。内容涵盖核心模型的具体原理，以及在文本到语音合成 (Text-to-Speech, TTS)、声音转换 (Voice Conversion, VC)、通用音频与音乐生成等方面的代表性应用与前沿进展。特别地，本文还包含了作者的实践探索，详细报告了对声音转换、AI 音乐生成、语音反欺骗领域的较新项目的复现实验过程、结果与分析。本文提供了一个涵盖了理论、应用和实践的综合性框架，全面总结了语音与音频领域 AIGC 及检测技术的现状，讨论了当前面临的关键技术挑战，并对未来的研究方向和伦理考量进行了前瞻性思考，以推动该领域的发展和应用创新。

关键字：人工智能生成内容；语音与音频生成；语音合成；声音转换；AI 生成音频检测；综述

Speech and Audio AI Generation and Detection: A Review

Tianhao Jin[1]

(1 School of the Gifted Young, University of Science and Technology of China, Hefei, Anhui, China)

Abstract: In recent years, AI Generated Content (AIGC) technology has achieved significant development in the field of speech and audio, demonstrating immense application potential in various domains (e.g., human-computer interaction, content creation, entertainment industry, etc.). However, potential risks and challenges associated with the advancement of AIGC technology are also increasingly emerging. In particular, Voice Deepfake technology may be utilized for illegal activities such as fraud, posing serious threats to personal privacy, social trust, and cybersecurity. Copyright and originality of AI-generated content are also prone to controversy. Therefore, researching and developing AI-generated audio detection technology to address these challenges is of considerable importance and urgency. This paper aims to provide a systematic and in-depth review of AIGC and detection technologies in speech and audio. The content is structured according to historical development periods: the machine learning era, the early deep learning era, and the modern deep learning era. The review covers the specific principles of core models, as well as representative applications and frontier advancements in areas such as Text-to-Speech (TTS), Voice Conversion (VC), and general audio and music generation. Notably, this paper also incorporates the authors' practical exploration, detailing the experimental process, results, and analysis from reproducing recent projects in the fields of voice conversion, AI music generation, and speech anti-spoofing. This paper offers a comprehensive framework encompassing theory, applications, and practice. It provides a thorough summary of the current state of AIGC and detection technologies in speech and audio, discusses the critical technical challenges presently encountered, and offers forward-looking perspectives on future research directions and ethical considerations, with the aim of promoting the field's development and application innovation.

Keywords: AI Generated Content; Speech and Audio Generation; Text-to-Speech; Voice Conversion; AI-Generated Audio Detection; Review

1 引言

1.1 研究背景与动机

人工智能生成内容 (AI Generated Content, AIGC) 在近些年各个领域呈现出发展态势，在文本、图像、视频等模态有了突破以后，AIGC 技术快速朝着语音和音频这个更具感知沉浸感以及交互自然性的领域延伸，语音和音频是人类交流与信息传递的基础载体，还是情感表达和艺术创作的关键媒介。

AIGC 通过模拟、合成、转换及增强等手段，在语音与音频领域进行了广泛的应用。例如，它可以创造出个性化、富有情感的智能语音助手，提升人机交互的自然度和用户体验；能够为语言障碍人士提供辅助发声工具，帮助他们更流畅地与外界交流；可以自动化播客、有声读物等内容的制作流程，降低生产成本并提高效率；在音乐创作领域，AI 作曲工具使得不具备专业音乐技能的普通用户也能参与到音乐创作中，拓展了创意表达的方式；同时，在游戏音效设计、影视配音、虚拟偶像塑造以及元宇宙等新兴数字娱乐场景中，AIGC 技术也为创造沉浸式、高保真的听觉体验提供了技术支撑。

然而 AIGC 技术在给人们带来便利以及创新的情形下，它的负面效应渐渐变得明显起来，引起了社会的广泛关注，语音深度伪造 (Voice Deepfake) 技术，它可以依靠学习少量目标人物的语音样本，合成出几乎可达到以假乱真程度的虚假语音，AIGC 技术被滥用存在着风险，这是需要关注的：比如利用伪造语音实施电信诈骗、敲诈勒索，进行恶意诽谤、损害个人名誉，以及传播虚假新闻、操纵社会舆论等威胁场景。另外 AI 音乐生成等技术尽管在辅助创作方面有一定潜力呈现，可也引发了一系列复杂的法律和伦理问题，像版权归属、风格抄袭以及对人类艺术家原创积极性造成的冲击等。

面对 AIGC 技术潜在的滥用风险和带来的多重挑战，研究和发展高效、鲁棒的 AI 生成音频检测技术，构建可信的 AI 音频环境，已成为学术界和产业界共同关注的焦点。有效的检测不仅是技术层面应对风险的必要手段，更是保障 AIGC 技术能够朝着健康、可控方向发展，维护个人权益、社会秩序和网络安全的关键环节。这要求我们不仅要深入理解 AIGC 的生成机理，也要不断探索和创新检测与防御的策略。

1.2 本文主要内容与结构

本文聚焦于语音与音频领域的人工智能生成及检测技术进行综述，其主要覆盖了对该领域中关键技术发展历程给予回顾，对核心模型以及算法原理展开解析，对典型应用场景加以介绍，还包括对当前所面临的主要挑战以及未来发展趋势展开探讨，特别值得一提的是，本文会详细阐述笔者在声音转换、AI 音乐生成以及语音反欺骗检测方面的实践探索情况与实验结果。

本文的组织结构如下：第二章介绍了语音信号处理的基础知识，包括数字化处理方法和声学特征提取技术。第三章回顾了早期基于统计参数的语音合成方法，为后续理解深度学习模型奠定基础。第四章回顾了深度学习在语音音频 AIGC 领域的早期探索，重点介绍了生成对抗网络 (Generative Adversarial Networks, GAN)、变分自编码器 (Variational Autoencoders, VAE)、流模型 (Flow-based Models) 等核心生成模型的基本原理及其在声码器、文本到语音合成 (Text-to-Speech, TTS)、声音转换 (Voice Conversion, VC) 等任务上的初步应用。第五章深入阐述了现代深度学习技术（如 Transformer、预训练范式、Diffusion 模型）如何驱动 AIGC 取得突破性进展，并详细介绍了作者在声音转换和 AI 音乐生成方面的实践工作。第六章全面介绍了 AI 生成音频的检测技术，包括 ASVspoof 挑战赛、主流的语音 Deepfake 检测模型，以及作者在 AASIST 模型复现方面的实践。第七章对全文进行总结，系统分析当前 AIGC 及检测技术面临的核心挑战，并对未来的研究方向（包括新型网络架构和主动防御技术）和伦理问题进行展望。

2 语音信号处理基础

语音信号作为 AIGC 的基础载体，其理解和处理是后续各类 AIGC 技术的前提。本章首先介绍语音信号从模拟到数字的转换过程和基本处理方法，然后重点阐述传统声学特征提取技术，为理解后续章节中的深度学习方法提供必要的理论基础。

2.1 语音信号数字化处理

语音信号作为一种承载丰富信息的声学波，其物理本质是空气或其他介质的振动。从发声机理来看，人类语音主要由声带振动产生的准周期激励（浊音）或口腔、鼻腔等声道中气流受阻产生的湍流噪声（清音）经过声道（包括咽腔、口腔、鼻腔）的谐振调制而形成。为了能够利用计算机对语音信号进行分析、处理和合成，首先需要将其从连续的模拟信号转换为离散的数字信号。这一过程主要包括采样和量化两个步骤。采样是指按照一定的时间间隔（采样周期）对模拟信号的瞬时幅度进行提取，根据奈奎斯特采样定理，采样频率至少应为信号最高频率的两倍才能无失真地恢复原始信号。量化则是将采样得到的幅度值用有限个离散的电平值来表示，量化的精度通常用比特深度来衡量。经过数字化后，语音信号可以表示为一系列离散的采样点，即时域波形。此外，为了分析语音信号的频率特性，通常采用短时傅里叶变换 (Short-Time Fourier Transform, STFT) 将其转换到频域，得到随时间变化的频谱或语谱图，这为后续的特征提取和建模提供了基础。

2.2 声学特征提取方法

在数字化的语音信号基础上，提取能够有效表征语音内容、说话人特性或情感状态的声学特征至关重要。梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCCs)[1] 是早期最为经典且广泛应用的声学特征之一。MFCCs 的提取过程模拟了人类听觉系统对频率的感知特性（即对低频更敏感，对高频的敏感度则按对数关系降低）。具体步骤通常包括：1) 预加重，以提升高频分量；2) 分帧，将语音信号分割成短时帧；3) 加窗，对每帧信号加窗以减少频谱泄露；4) 快速傅里叶变换 (Fast Fourier Transform, FFT)，计算每帧的功率谱；5) 通过梅尔滤波器组，将线性频率谱映射到梅尔频率谱；6) 取对数，得到对数梅尔频谱；7) 离散余弦变换 (Discrete Cosine Transform, DCT)，将对数梅尔频谱解相关并降维，得到 MFCCs。由于 MFCCs 能够较好地表征语音的短时谱包络信息，且对加性噪声和信道失真具有一定的鲁棒性，因此在早期的语音识别、说话人识别等任务中取得了显著成功。

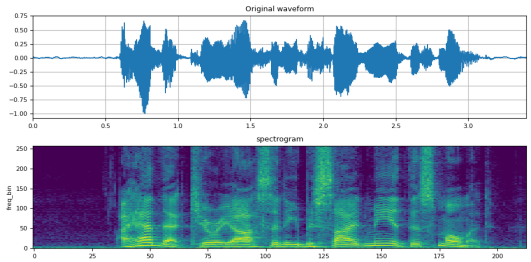

图 1: 原声波及其对应 MFCCs 图 [2]

除了 MFCCs，其他常用的声学特征还包括基频 (Fundamental Frequency, F0)，它反映了声带振动的频率，是构成语音语调和声调的主要物理量；语音能量 (Energy)，表征语音信号的强度；过零率 (Zero-Crossing Rate, ZCR)，常用于区分清浊音；线性预测系数 (Linear Prediction Coefficients, LPCs)，通过对语音信号进行线性预测建模得到的一组参数，能够有效地表示声道特性；以及共振峰 (Formants)，即声道谐振频率，对元音的区分至关重要。在深度学习方法广泛应用之前，基于这些手工设计的声学特征，结合隐马尔可夫模型 (Hidden Markov Models, HMM)、矢量量化 (Vector Quantization, VQ)、混合高斯模型 (Gaussian Mixture Models, GMM) 等统计建模技术 [3]，构成了传统语音识别、说话人识别、语音编码和早期语音合成系统的核心。

3 统计参数语音合成方法

统计参数语音合成 (Statistical parametric speech synthesis, SPSS) 方法作为连接传统语音处理技术与现代深度学习的重要桥梁，奠定了参数化语音建模的理论基础。本章首先阐述 SPSS 的核心技术原理，重点分析以 HMM 为代表的统计建模方法；随后详细介绍 SPSS 的完整合成流程，并深入分析其技术优势与固有局限性，为理解后续深度学习方法的突破性进展提供必要的技术背景。

3.1 SPSS 技术原理

在神经网络语音合成方法兴起之前，SPSS[4] 是主流的合成技术之一。SPSS 的核心思想是利用统计模型来描述输入文本的语言学特征与输出语音的声学参数（如频谱包络、基频、时长等）之间的映射关系。其中，基于 HMM 的语音合成（也称 HTS, HMM-based Text-to-Speech）是最具代表性的 SPSS 方法。在 HTS 的训练阶段，首先从文本中提取上下文相关的语言学特征（如音素、音调、词性、位置信息等），同时从对应的录音中提取声学参数。然后，利用 HMM 来对每个语言学单元（如上下文相关的音素）的声学参数序列进行建模，学习 HMM 的状态转移概率以及每个状态下声学参数（通常假设服从高斯分布或高斯混合分布）的概率密度函数。

3.2 SPSS 合成流程与局限性分析

在 SPSS 的合成阶段，当输入一段新的文本时，系统首先将其转换为语言学特征序列。然后，根据训练好的 HMM 模型，利用最大似然参数生成算法（如 Viterbi 算法确定最优状态序列，再根据状态序列生成声学参数）来预测与该语言学特征序列最匹配的声学参数序列（包括频谱参数、F0 序列和音素时长）。最后，使用一个独立的声码器 (Vocoder)，将生成的声学参数序列转换为可听的语音波形。

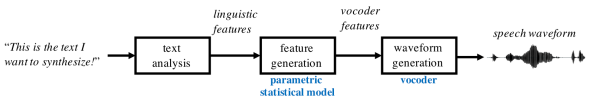

图 2: SPSS 合成流程图 [5]

SPSS 方法和早期的拼接合成 (Concatenative Synthesis) 相比有一些优势，像是模型尺寸较小，可便捷地对合成语音的特征（如语速、平均 F0、频谱倾斜等）做平滑控制与修改，达成一定程度的风格转换或者情感表达，不过 SPSS 也有较大的局限性，统计模型（特别是 HMM）对声学参数做了平均化以及过度平滑处理，使得合成语音的自然度和逼真度一般较低，听起来大多时候带有十分突出的非人感或者噪声，难以呈现细腻的情感和丰富的韵律细节。即便如此，SPSS 在建模语音的统计特性、参数化表示以及可控合成方面的探索，为后续基于深度学习的语音合成方法，端到端模型的兴起，提供了关键的技术积累和研究思路。

4 早期深度学习时期

早期深度学习在语音与音频 AIGC 领域的探索始于 2010 年代中期，标志着从传统统计方法向神经网络方法的转变。本章首先介绍核心模型架构，包括循环神经网络 (Recurrent Neural Networks, RNN)/长短期记

忆网络 (Long Short-Term Memory, LSTM)/门控循环单元 (Gated recurrent units, GRU) 的序列建模、GAN 的对抗生成、VAE 的概率生成以及流模型的可逆变换。随后按照技术发展脉络展开，从神经声码器（以 WaveNet 为代表）到端到端文本到语音合成（如 Tacotron 系列），最后到基于深度学习的声音转换方法。

4.1 核心模型架构原理

随着计算能力不断提升以及大规模数据集的涌现，深度学习于语音与音频处理领域开始彰显出巨大潜力，渐渐取代了传统的统计建模方法，在 Transformer 架构问世以前，一些关键的深度学习架构为后续 AIGC 技术奠定了基础。

首先，RNN[6] 因其能够处理序列数据的特性，在语音信号这种典型的时序数据建模中得到了广泛应用。RNN 通过在隐藏层引入循环连接，使得网络能够利用前一时刻的输出作为当前时刻的输入的一部分，从而捕捉时间序列中的依赖关系。然而，传统的 RNN 在处理长序列时容易出现梯度消失或梯度爆炸问题，导致难以学习到长距离的依赖。为了解决这一问题，LSTM[7] 和 GRU[8] 被相继提出。LSTM 通过引入输入门、遗忘门和输出门三个门控单元以及一个细胞状态来精细地控制信息的流动和保留，有效地缓解了长期依赖问题。GRU 则对 LSTM 的门控结构进行了简化，参数量更少，在某些任务上也能达到与 LSTM 相当的性能。这些改进的 RNN 结构在早期的语音识别、语音合成和声音转换等任务中扮演了重要角色。

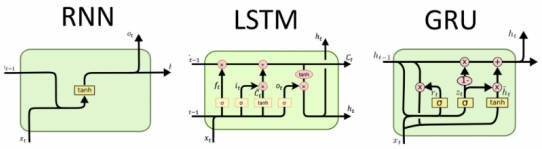

图 3: RNN/LSTM/GRU 结构图 [9]

其次，GAN[10] 的提出为无监督生成模型带来了革命性的突破。GAN 的核心思想是通过构建一个生成器 (Generator, G) 和一个判别器 (Discriminator, D) 之间的对抗博弈来学习数据的真实分布。生成器 G 的目标是学习从一个简单的先验分布（如高斯噪声）到目标数据分布的映射，生成尽可能逼真的“假”数据；而判别器 D 的目标则是尽可能准确地区分输入数据是来自真实数据集还是由生成器 G 生成的“假”数据。两者通过一个极小极大博弈 (minimax game) 的目标函数进行交替优化：G 试图最小化 D 正确判断的概率，而 D 试图最大化正确判断的概率。理想情况下，当达到纳什均衡时，G 能够生成与真实数据无法区分的样本，而 D 对任何输入的判断概率都接近 0.5。GAN 的优势在于能够生成非常清晰和锐利的样本，尤其在图像生成领域取得了巨大成功。然而，GAN 的训练过程通常不稳定，容易出现模式崩溃（mode collapse，即生成器只能生成非常有限种类的样本）和梯度消失等问题，对超参数和网络结构设计较为敏感。

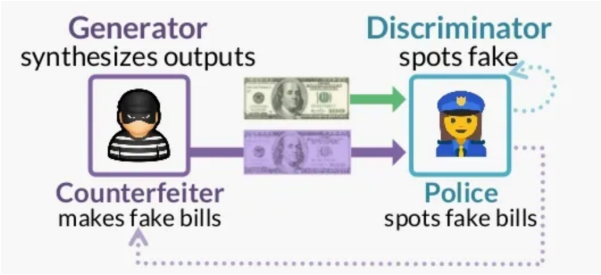

图 4: GAN 原理图 [11]

再次，VAE[12] 是另一种重要的深度生成模型，它将自编码器 (Autoencoder) 的思想与概率图模型和变分推断相结合。VAE 由一个编码器 (Encoder) 和一个解码器 (Decoder) 组成。编码器网络学习将输入数据 x 映射到潜在空间 (latent space) 中的一个概率分布（通常为高斯分布），即学习潜在变量 z 的后验分布 $q_\phi(z|x)$ 的参数（如均值和方差）。然后从该分布中采样得到潜在向量 z，再由解码器网络学习从潜在向量 z 重构回原始数据空间，即学习生成数据的似然分布 $p_\theta(x|z)$。VAE 的优化目标是最大化数据的边际似然 $p(x)$ 的变分下界 (Evidence Lower Bound, ELBO)，该下界由两部分组成：重构损失（期望对数似然 $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$），衡量模型重构输入数据的能力；以及一个正则化项，通常是潜在分布 $q_\phi(z|x)$ 与先验分布 $p(z)$（通常为标准正态分布）之间的 KL 散度，用于约束潜在空间的结构，使其更加平滑和连续。VAE 的优点在于训练过程相对稳定，并且由于潜在空间的良好结构，非常适合进行数据插值和有意义的属性控制。然而，VAE 生成的样本通常不如 GAN 生成的样本清晰，可能会显得较为模糊。

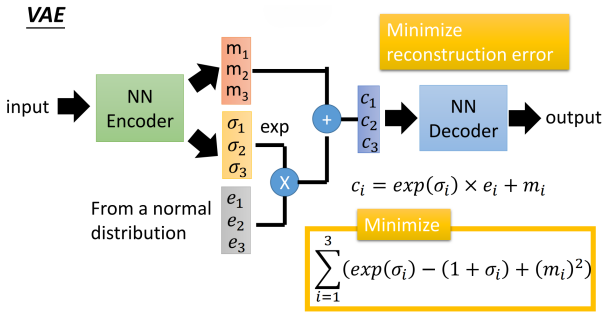

图 5: VAE 原理图 [13]

最后，流模型 [14] 是另一类强大的深度生成模型，其核心思想是构建一系列可逆的、且其雅可比行列式 (Jacobian determinant) 容易计算的非线性变换函数。这些变换函数将一个简单的、已知的先验概率分布（如标准高斯分布）逐步映射到一个复杂的、目标数据所在的概率分布。由于变换是可逆的，因此既可以从先验分布中采样，然后通过正向变换生成新的数据样本，也可以将真实数据通过逆向变换映射回潜在空间。更重要的是，

根据概率密度变换公式，可以直接计算出目标数据在模型下的精确对数似然度 (exact log-likelihood)，这使得流模型可以直接通过最大似然估计进行优化，训练过程通常较为稳定。NICE (Non-linear Independent Components Estimation)[15] 是早期流模型的代表作，它通过设计特殊的耦合层来实现可逆且雅可比行列式易于计算的变换。后续的 RealNVP[16]、Glow[17] 等模型在此基础上进行了改进和扩展。流模型在密度估计和高质量样本生成方面都展现出优秀的能力。

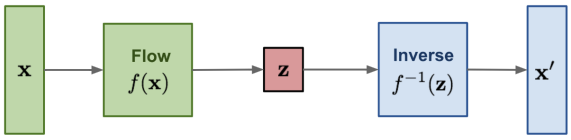

图 6: 流模型原理图 [18]

4.2 早期神经声码器

神经声码器 (Neural Vocoder) 是将声学特征（如梅尔频谱图）转换为高质量、可听的音频波形的关键组件，广泛应用于 TTS 和 VC 系统中。传统的基于信号处理的声码器虽然计算效率较高，但在合成语音的自然度和真实感方面往往存在不足，容易产生非人感或相位失真。深度学习的引入为声码器技术带来了革命性的突破，使得直接从声学特征生成高质量波形成为可能。

WaveNet[19] 是神经声码器领域的里程碑式工作。它是一个基于卷积神经网络 (Convolutional Neural Networks, CNN) 的自回归生成模型，直接对原始音频波形的每个采样点进行建模。WaveNet 的核心结构采用了扩张因果卷积 (Dilated Causal Convolutions)，使得网络在保持因果性的同时（即当前采样点的预测只依赖于过去的采样点），能够拥有非常大的感受野，从而捕捉音频信号中的长距离依赖关系。此外，WaveNet 还会引入全局或局部条件（如语言学特征、梅尔频谱或说话人 ID）来指导波形的生成。通过逐点预测下一个采样点的概率分布并进行采样，WaveNet 能够生成与真实录音质量相当的自然语音。然而，其主要的缺点在于自回归的生成方式导致推理速度极慢，每个采样点的生成都需要等待前一个采样点完成，这使得 WaveNet 难以直接应用于实时或大规模的语音合成任务。

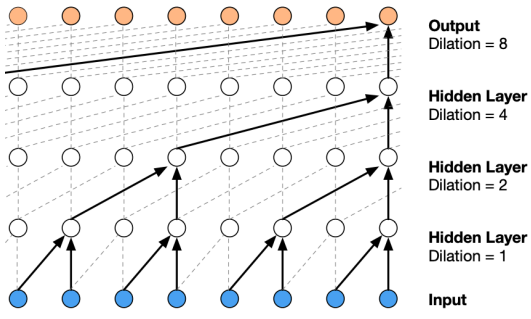

图 7: WaveNet 结构图 [19]

为了克服 WaveNet 推理速度慢的问题，研究者们提出了一系列改进方案。WaveRNN[20] 是其中一个重要的尝试，它将 WaveNet 中的卷积层替换为更紧凑的 RNN 结构，并通过一些优化技巧（如子尺度生成、权重剪枝）来加速采样过程，能够在图形处理器 (Graphics Processing Unit, GPU) 上实现更快的生成速度，同时保持了较高的合成质量。另一个重要的方向是流模型声码器，例如 WaveGlow[21]。WaveGlow 借鉴了 Glow [17] 的思想，构建了一个从简单高斯分布到目标梅尔频谱条件下的音频波形分布的可逆映射。由于流模型的变换是可逆且并行的，WaveGlow 能够实现非常快速的非自回归波形生成，其推理速度远超 WaveNet 和 WaveRNN，同时合成质量也与 WaveNet 相当。这些早期的神经声码器极大地推动了语音合成自然度的提升，并为后续更高效、更高质量的声码器研究奠定了基础。

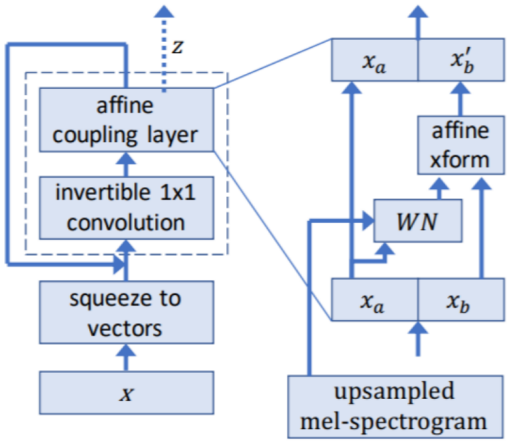

图 8: WaveGlow 结构图 [21]

4.3 早期文本到语音合成

传统的 TTS 系统一般采用复杂的流水线式处理流程，其中包含文本分析前端（涉及文本正则化、分词、词性标注以及韵律预测等），以及声学模型（把语言学特征映射为声学参数）和声码器（将声学参数合成为语音波形）等多个独立优化的模块。这种分阶段的处理方式设计较为复杂，容易出现人工设计带来的误差，而且各模块间的错误会不断累积，对最终合成语音的质量造成了限制。端到端 (End-to-End) TTS 模型出现后，借助一个或少数几个联合优化的神经网络模型，直接依据输入的文本（一般是字符或音素序列）生成声学特征或者原始音频波形，以此简化系统结构并提高合成效果。

Tacotron[22] 是早期具有代表性的端到端 TTS 模型之一。它采用了一个基于注意力机制的序列到序列 (Sequence-to-Sequence, Seq2Seq) 架构。其 Encoder 部分使用卷积层和循环层来提取输入文本序列的上下文表示。Decoder 部分则是一个自回归的 RNN，它在每个时间步利用注意力机制 (Attention Mechanism) 来关注编码器输出的相关部分，并预测声学特征序列（通常是梅尔

频谱帧）。注意力机制的引入使得模型能够自动学习文本与语音之间的对齐关系，解决了传统 TTS 中对齐模块难以设计的问题。Tacotron 能够直接从字符级输入生成梅尔频谱，避免了复杂的语言学特征工程。

在 Tacotron 的基础上，Tacotron 2[23] 进一步提升了端到端 TTS 的性能，达到了当时业界领先的水平。Tacotron 2 的声学模型部分同样采用了基于注意力机制的 Seq2Seq 架构，但在编码器和解码器的具体结构上使用了更简单的 LSTM 层和卷积层。更重要的是，Tacotron 2 将生成的梅尔频谱图作为条件输入，送入一个经过修改的 WaveNet 声码器 [19] 来直接合成高质量的时域波形。通过将高质量的声学模型与强大的神经声码器相结合，Tacotron 2 生成的语音在自然度和清晰度方面都非常接近人类录音的水平，其平均意见分 (Mean Opinion Score, MOS) 显著优于当时的参数合成和拼接合成系统。这些早期的端到端 TTS 模型确立了基于 Seq2Seq 和注意力机制的 TTS 基本框架，为后续的非自回归模型和更高质量的语音合成技术铺平了道路，但也普遍存在由于自回归解码导致的推理速度较慢的问题。

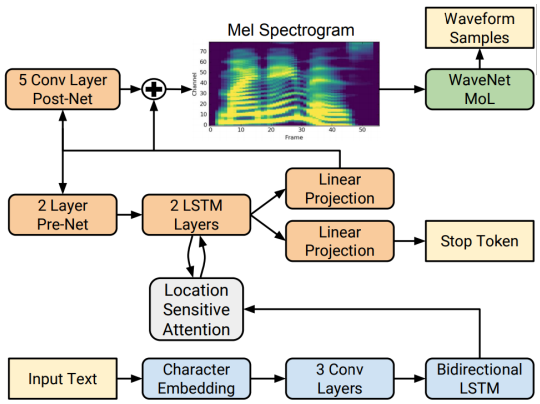

图 9: Tacotron 2 结构图 [23]

4.4 早期声音转换

VC 的目标是在保留原始语音内容（即说了什么）的前提下，将其中的说话人身份特征（即是谁说的，以及怎么说的风格）转换为目标说话人的特征。VC 技术在个性化语音合成、语音模仿、电影配音、隐私保护等领域具有广泛的应用前景。传统的 VC 方法通常依赖于源说话人和目标说话人说相同内容的平行语音语料库，通过对齐技术建立帧级别的对应关系，然后学习声学特征（如 MFCCs、LPCs、F0 等）的转换函数。然而，获取大规模、高质量的平行语料在实际中往往非常困难，这极大地限制了传统 VC 方法的应用范围和效果。

为了克服对平行数据的依赖，研究者们开始探索基于深度学习的非平行数据声音转换方法。基于 VAE 的 VC 方法 [24] 是其中的一个重要方向。这类方法的核心思想是尝试将语音信号分解为与说话人身份无关的内容表示 (content representation) 和与说话人身份相关的风格/音色表示 (style representation)。例如，可以训练一个编码器将输入语音映射到一个潜在空间，并假设该潜在空间中内容信息和说话人信息是可分离的。在转换阶段，提取源语音的内容表示，并结合目标说话人的嵌入向量，然后通过解码器重构出目标说话人风格的语音。VAE 的正则化项有助于学习到平滑的潜在空间，从而支持内容和风格的解耦。

另一类重要的非平行数据 VC 方法是基于 GAN 的。CycleGAN-VC[25] 借鉴了图像风格迁移中 CycleGAN[26] 的思想，通过构建两个方向的转换生成器（源到目标，目标到源）和对应的判别器，并引入循环一致性损失 (cycle-consistency loss) 来约束转换过程。循环一致性损失确保了将源语音转换为目标风格后再转换回源风格时，其内容应与原始源语音尽可能一致，从而在没有平行数据的情况下学习到有效的转换映射。对抗损失则有助于生成更逼真、更少平滑的声学特征。StarGAN-VC[27] 进一步将 StarGAN[28] 架构应用于 VC 任务，通过引入域标签 (domain labels) 代表不同的说话人，使得单个生成器和判别器能够学习多个说话人之间的转换，实现了多对多的非平行声音转换。CycleGAN-VC2[29] 则在 CycleGAN-VC[25] 的基础上，对生成器和判别器的网络结构以及目标函数进行了改进，进一步提升了转换后语音的自然度和与目标说话人的相似度。这些基于深度学习的非平行数据 VC 方法极大地拓展了声音转换技术的适用范围，并为后续更高质量、更灵活的 VC 系统奠定了基础。

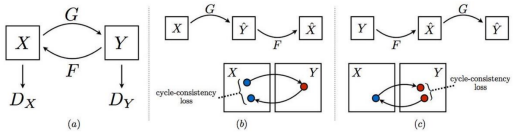

图 10: CycleGAN 原理图 [26]

5 现代深度学习时期

步入 21 世纪 20 年代，语音与音频领域的 AIGC 技术迎来了前所未有的重大突破。本章首先介绍推动现代 AIGC 技术发展的核心架构创新，包括 Transformer 及其变体 Conformer 的序列建模革命、大规模预训练范式带来的表示学习突破，以及扩散模型在生成质量上的新高度。随后，本章从技术发展顺序展开，从现代神经声码器（以 HiFi-GAN、DiffWave 为代表）的效率与质量提升谈起，接着探讨基于 Transformer 的现代 TTS 技术（如 FastSpeech 系列）如何实现高质量非自回归合成，然后深入分析现代声音转换技术的发展并详述作者在 RVC 和 GPT-SoVITS 方面的实践探索，最后介绍通用音频与 AI 音乐生成的前沿进展及 QA-MDT 的复现实验。

5.1 核心模型架构原理

现代语音音频 AIGC 技术的飞速发展，离不开几个关键基础模型和训练范式的支撑。

Transformer 架构 [30] 凭借其独特的自注意力 (Self-Attention) 机制，彻底改变了序列建模领域。传统的 RNN 在处理长序列时存在梯度传播和并行计算的瓶颈，而 Transformer 通过允许模型在处理序列中的每个元素时，同时关注到序列中所有其他元素，并根据相关性动态分配权重，从而能够高效地捕捉长距离依赖关系。其核心组件多头注意力 (Multi-Head Attention) 通过在不同的线性投影子空间中并行计算自注意力，并拼接结果，进一步增强了模型从不同角度理解和表示序列信息的能力。典型的 Transformer 通常采用编码器-解码器结构，编码器将输入序列映射为一系列上下文相关的表示，解码器则基于这些表示自回归或非自回归地生成目标序列。位置编码 (Positional Encoding) 的引入解决了 Transformer 本身缺乏序列顺序信息的问题。残差连接 (Residual Connections) 和层归一化 (Layer Normalization) 则有助于训练更深、更稳定的 Transformer 网络。由于其强大的并行计算能力和对长序列的建模优势，Transformer 及其变体已成为现代语音 AIGC 领域的主流骨干网络。

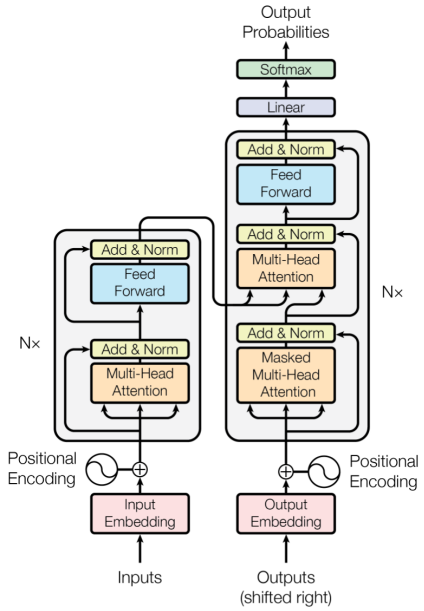

图 11: Transformer 结构图 [30]

针对语音信号的特性，Conformer[31] 模型巧妙地将 Transformer 的全局上下文建模能力与 CNN 在捕捉局部声学特征方面的优势相结合。Conformer 的核心模块通常采用一种类似“三明治”的结构，即一个前馈网络层 (Feed-Forward Network, FFN) 之后接一个多头自注意力模块，然后是一个卷积模块，最后再接一个 FFN 层。卷积模块能够有效地提取语音信号中的局部时频模式，而自注意力模块则负责建模全局的上下文依赖。这种结合使得 Conformer 在语音识别 (Automatic Speech Recognition, ASR) 任务上取得了超越纯 Transformer 或纯 CNN 模型的性能，并逐渐被应用于语音合成和声音转换等 AIGC 任务中。

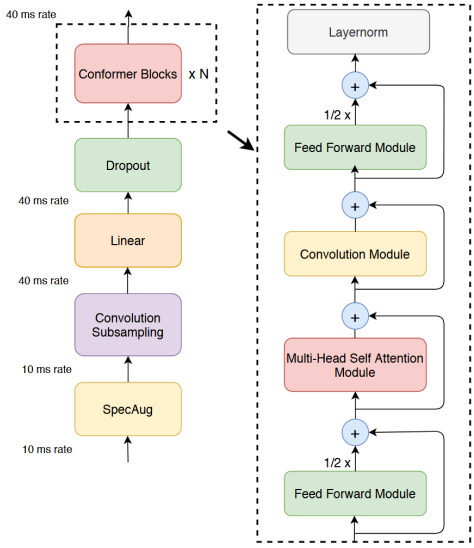

图 12: Conformer 结构图 [31]

大规模预训练 (Pre-training) 范式，借鉴了自然语言处理领域 BERT[32] 等模型的成功经验，为语音音频 AIGC 带来了新的突破。其核心思想是“在大规模无标注或弱标注数据上进行自监督学习 (Self-Supervised Learning) 以获得通用的、高质量的特征表示，然后在特定的下游任务上使用少量有标注数据进行微调 (Fine-tuning)”。对于语音领域，自监督学习任务通常围绕着如何从原始音频中学习有意义的声学和语言学信息展开，例如 HuBERT[33] 通过预测被掩码音频片段对应的离散聚类单元 (acoustic units) 来学习表示；WavLM[34] 则进一步结合了去噪任务，使其学习到的表示对噪声更鲁棒，并能更好地适应多种语音处理任务。BEATs[35] 则通过迭代音频训练和掩码预测任务，面向更加泛用的通用音频事件。预训练模型能够从海量数据中习得丰富的先验知识，显著提升了模型在数据稀疏的下游 AIGC 任务上的性能和泛化能力。

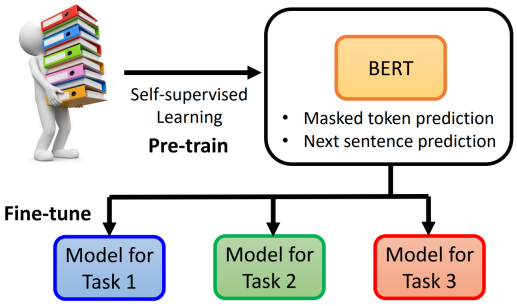

图 13: 预训练-微调流程图 [13]

去噪扩散概率模型 (Denoising Diffusion Probabilistic Models, DDPM)[36] 作为一种新兴的强大生成模型，近年来在图像、音频等多个领域展现出惊人的生成质量。其核心原理模拟了一个物理扩散过程：首先定义一个前向扩散过程 (forward process)，在该过程中，原始数据通过逐步、迭代地添加高斯噪声，直至完全变为服从简单先验分布（如标准正态分布）的纯噪声。然后，模型的

核心任务是学习一个反向去噪过程 (reverse process)，即训练一个神经网络来预测并移除每一步添加的噪声，从而从纯噪声逐步恢复出清晰的原始数据。在生成新样本时，从先验分布中采样一个随机噪声，然后迭代地应用学习到的反向去噪网络。Diffusion 模型的主要优势在于能够生成非常高质量和多样性的样本，并且训练过程相对稳定，不易出现模式崩溃。然而，其主要的计算挑战在于生成过程通常需要较多的迭代步骤（采样步数），导致推理速度相对较慢，但后续研究如去噪扩散隐式模型 (Denoising Diffusion Implicit Models, DDIM)[37] 等也在不断改进其采样效率。

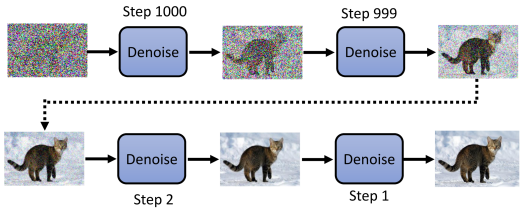

图 14: Diffusion 模型流程图 [13]

5.2 现代声码器

现代声码器的发展目的在于，在保证高保真度 (high-fidelity) 以及自然度的情况下，尽量去提升波形生成的效率，当 WaveNet[19] 证明了神经声码器在质量方面有巨大潜力后，研究的重点便转向了非自回归或者高效并行生成方案。

基于 GAN 的声码器是当前最主流和成功的高效高保真方案之一。HiFi-GAN[38] 是其中的杰出代表。HiFi-GAN 的生成器采用轻量级的 CNN 结构，直接从梅尔频谱图上采样生成原始波形。其成功的关键在于判别器的设计：它使用了多个并行的判别器，如多尺度判别器 (multi-scale discriminator, MSD) 和多周期判别器 (multi-period discriminator, MPD) 上对生成的波形进行判别。MSD 关注波形的整体结构和平滑性，而 MPD 则专注于捕捉音频中不同周期的谐波结构和相位关系，这对于生成具有丰富细节和自然音质的语音至关重要。通过这种多判别器联合训练的方式，HiFi-GAN 能够在保持极高合成质量的同时，实现非常快的推理速度，使其成为众多现代 TTS 和 VC 系统的首选声码器。

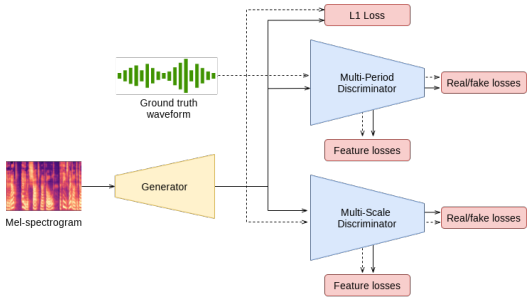

图 15: HiFi-GAN 结构图 [39]

基于扩散模型的声码器也展现出强大的竞争力。DiffWave[40] 是较早将扩散模型应用于音频波形生成的声码器之一。它将梅尔频谱图作为条件输入，指导从高斯噪声到目标语音波形的反向去噪过程。DiffWave 能够生成与 WaveNet 质量相当甚至更好的音频，并且其非自回归的特性使其推理速度远快于 WaveNet。尽管扩散模型的采样过程通常需要较多迭代步骤，但通过一些加速采样技术（如 DDIM 采样）[37]，其效率可以得到一定提升。后续也出现了更多针对音频特性优化的扩散声码器。

5.3 现代文本到语音合成

现代 TTS 系统，在 Transformer[30] 架构以及高效声码器的推动作用下，有了较大进展，其在自然度方面逐渐靠近人类水准，合成速度与可控性这两方面也有了较为十分突出的提升。

基于 Transformer 的 TTS 模型 [41] 利用自注意力机制并行处理输入文本序列，并生成声学特征（如梅尔频谱）。相比于早期基于 RNN 的 Seq2Seq 模型，Transformer 能够更好地捕捉长距离的文本依赖关系，并且训练效率更高。然而，早期的 Transformer TTS 模型（如 Transformer-TTS）的解码器部分仍采用自回归方式生成梅尔频谱，限制了推理速度。

为了解决自回归 TTS 模型的速度瓶颈，非自回归 TTS 模型应运而生。FastSpeech[42] 是其中的开创性工作，它引入了一个独立的时长预测器 (duration predictor)，该预测器从一个预训练的自回归 Teacher 模型中提取音素-梅尔频谱的对齐信息，并学习预测每个输入音素对应的梅尔频谱帧数。在推理时，首先预测音素时长，然后根据时长扩展音素序列，最后通过一个基于 Transformer 的前馈网络并行地生成整个梅尔频谱序列。FastSpeech 2[43] 则进一步改进了这一框架，它不再依赖 Teacher 模型的对齐信息和知识蒸馏，而是直接从真实语音中提取时长、基频和能量等韵律信息作为模型的条件输入进行训练，并在推理时由模型内部的预测器进行预测。这种方式不仅简化了训练流程，还使得模型能够更好地控制合成语音的韵律（如语速、音高变化、能量起伏），从而生成更自然、更富表现力的语音。

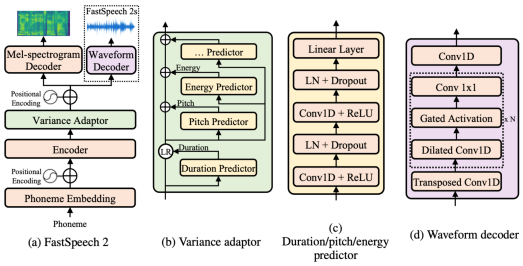

图 16: FastSpeech 2 结构图 [43]

VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech)[44] 是近年来另一个备受关注的端到端 TTS 模型，它将声学模型和声码器完全整合到一个统一的框架中，并结合了变分推断、标准化流

(Normalizing Flows) 和对抗训练。VITS 首先通过一个文本编码器将输入音素序列转换为潜在表示，然后利用一个随机时长预测器（引入了语音韵律的多样性）和一个标准化流将该表示映射到梅尔频谱的潜在空间。最后，一个基于 GAN 的解码器（类似于 HiFi-GAN 的生成器）直接从这个潜在表示生成高质量的音频波形。VITS 通过端到端的优化，实现了非常高的合成质量和较快的推理速度，并且能够通过对潜在变量的控制实现一定的风格迁移。

5.4 现代声音转换与实践探索

5.4.1 现代 VC 技术与开源项目概述

现代 VC 技术在深度学习，特别是预训练模型和先进生成模型的推动下，取得了显著的进步，尤其是在 Few-shot 甚至 Zero-shot 场景下的性能得到了大幅提升。与早期依赖平行语料或复杂解耦机制的方法不同，现代 VC 系统更倾向于利用强大的预训练声学模型 [33] 中蕴含的与说话人无关的内容特征，然后结合目标说话人的音色嵌入 (speaker embedding)，通过一个高效的解码器生成目标语音。

目前，学术界和开源社区涌现了许多优秀的 VC 项目，极大地推动了该技术的普及和发展。RVC (Retrieval-based Voice Conversion)[45] 是一个广受欢迎的开源项目，它巧妙地将基于 VITS 的合成框架与特征检索机制相结合。在训练阶段，它学习从 HuBERT 等模型提取的内容特征到目标说话人音色的映射。在推理阶段，除了直接使用模型预测，它还会从目标说话人的训练数据中检索与当前输入内容特征最相似的声学特征片段，并将其融入到最终的合成过程中，这种检索机制有助于生成更逼真、更符合目标说话人细节特征的语音，尤其在歌声转换任务中表现出色。so-vits-svc (SoftVC VITS Singing Voice Conversion)[46] 及其后续的 GPT-SoVITS[47] 是另一系列备受关注的开源项目。它们以 VITS[44] 架构为基础，so-vits-svc 侧重于高质量的歌声转换，而 GPT-SoVITS 则进一步引入了大型语言模型 GPT(Generative Pre-trained Transformer) 的能力来提升文本理解、韵律控制和少样本学习能力，使其不仅能进行高质量的声音转换，还能实现 Few-shot 甚至 Zero-shot 的 TTS。DDSP-SVC[48] 则将可微分数字信号处理 (Differentiable Digital Signal Processing, DDSP)[49] 的思想应用于歌声转换。DDSP 通过将传统的信号处理模块（如正弦振荡器、噪声滤波器）参数化并用神经网络进行预测，使得模型具有更好的可解释性和对音高、音色等属性的精细控制能力，非常适合于歌声的音高修正和音色转换。这些开源项目的共同特点是大大降低了高质量声音转换的技术门槛，使得研究者和爱好者能够基于少量数据快速训练出效果不错的个性化声音模型。

5.4.2 实践案例 1：RVC 微调与歌声转换

为能更透彻地明白现代声音转换技术实际应用所呈现的效果，此项研究针对开源的 RVC 模型展开了微调以及测试工作，着重剖析其在仅有少量特定说话人数据的场景里歌声转换的能力，同时也格外留意了其在跨语种（日语到英语）歌声转换方面的表现。

实验所使用的数据集来源于游戏《明日方舟》中角色” 蓝毒” 的日语配音。我收集了该角色约 12 分钟时长的清晰语音片段，所有音频均为 44.1kHz 采样率、16 位深度的 WAV 格式。在训练前，对原始音频数据进行了预处理，包括使用 MSST (Music Source Separation Training)[50] 等工具进行背景噪音消除和人声分离，以提高训练数据的纯净度。实验的硬件环境为配备 NVIDIA RTX 3060 Laptop GPU 的计算机，操作系统环境下配置了 CUDA 12.8。软件环境通过 Anaconda 进行管理，主要使用了 Python 3.10.16 和 PyTorch 2.1.0。在 RVC 模型微调过程中，设置 Batch Size 为 4，总共训练了 50 个 Epochs。

在模型训练的过程中，我利用 TensorBoard 对关键指标进行了监控，包括特征匹配损失 (Feature Matching Loss)、KL 散度损失 (KL Loss)、梅尔频谱损失 (Mel Loss) 以及总损失 (Total Loss) 等。通过观察这些损失函数随训练轮数的变化趋势，评估得到模型的收敛情况和训练稳定性良好。具体的训练监控结果如图 17-20 所示。

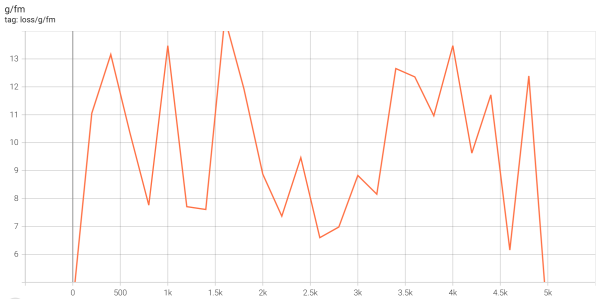

图 17: 特征匹配损失变化曲线

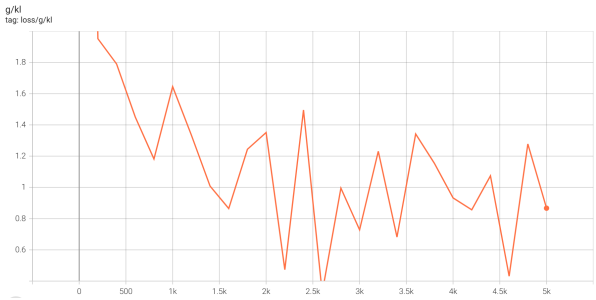

图 18: KL 散度损失变化曲线

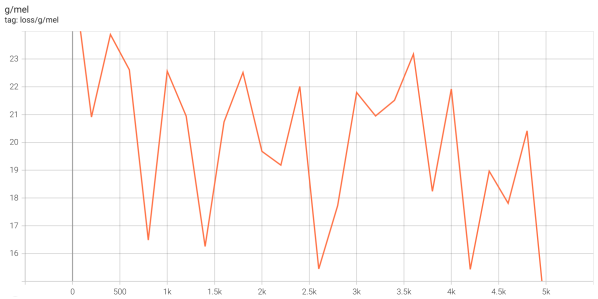

图 19: 梅尔频谱损失变化曲线

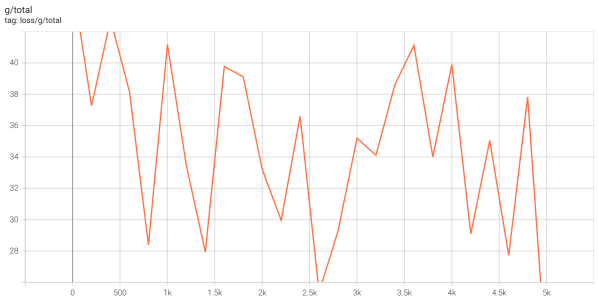

图 20: 总损失变化曲线

在对微调后的 RVC 模型进行效果评估时，我主要进行了主观听感分析。对于日语歌声的转换，转换后的歌声在音色上与” 蓝毒” 的原声线具有较高的相似度，能够基本还原其音色特点。歌声的自然度和流畅度也相当优秀。音准和节奏感方面，在大部分情况下能够保持与原唱一致，但对于音域跨度较大的部分，出现了一定的偏差。注意到，使用 RVC 提供的 RMVPE[51] 音高提取与修复算法，对于改善高音部分常见的” 哑音” 现象具有明显的修复作用。

在跨语种歌声转换的探索性测试中，我发现 RVC 模型在 Few-shot 条件下也展现出了一定的可行性。转换后的英文歌声能够保留” 蓝毒” 的基本音色特征，但发音清晰度和地道性方面有较大提升空间，带有较为明显的日语口音痕迹，且在某些英文音素的发音上不太自然。这表明在跨语种场景下，仅仅依赖声学特征的转换可能不足以完全克服语言模型层面的差异，需要更针对性的跨语言内容与风格解耦技术。同时，在源音频的和声未完全去除部分，转换后的歌声会出现跑调的现象。

总结本次 RVC 实践，其主要局限性体现在以下几个方面：首先，源音频的质量对最终转换效果影响显著，如果输入音频包含较多背景噪声、混响或未被彻底分离的多人声，转换结果很容易出现跑调、发音混乱等问题。其次，在 Few-shot 数据条件下，模型对于目标说话人音色的细微特征以及歌曲中复杂的高低音细节和发音清晰度的捕捉仍有待提升。这些问题可能源于训练数据的数量和质量限制、模型本身对复杂声学环境的鲁棒性不足，以及当前 VC 技术在内容与风格的彻底解耦方面仍面临挑战。相关的音频演示样本请参见附录 A 中的音频示例列表。

5.4.3 实践案例 2：GPT-SoVITS 微调与多语种语音克隆

为了能更深入地剖析现代声音转换与合成技术所有的能力，此项研究针对另一个受到广泛关注的开源项目 GPT-SoVITS[47] 展开了微调以及测试工作，GPT-SoVITS 的核心特性在于将大型语言模型 GPT 的能力和 SoVITS[46] 的优势相互融合，以期借助少量数据达成高质量的语音克隆，同时还支持多语种的 TTS 以及声音转换 VC。此次实践的目的在于验证 GPT-SoVITS 在 Few-shot 场景下的语音克隆效果，并且着重关注其在多语种混合推理以及学习并复现说话人独特发音风格方面的表现情况。

实验所使用配置均与 RVC 实践一致。GPT-SoVITS 的训练过程相对复杂，需要分别训练 SoVITS 模型（负责声学建模和波形生成）和 GPT 模型（负责从文本生成中间表示）。在本次实践中，我使用了项目的整合包，具体训练参数为：Python 3.9.13, PyTorch 2.0.0；SoVITS 模型训练设置 Batch Size 为 3，训练 3 个 Epochs；GPT 模型训练设置 Batch Size 为 3，训练 15 个 Epochs。

在模型训练过程中，同样利用 TensorBoard 对关键指标进行了监控，包括损失函数以及 Top-3 准确率等。通过分析这些指标的变化趋势，评估得到各模块的训练情况和整体模型的收敛性良好。具体的训练监控结果如图 21-22 所示。

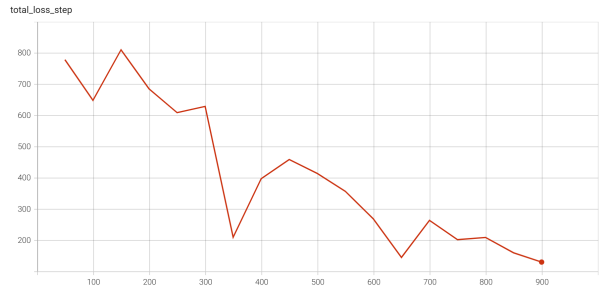

图 21: 损失函数变化曲线

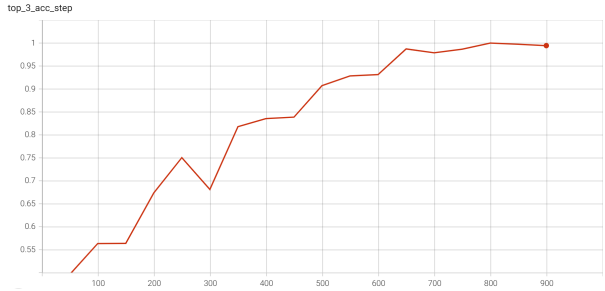

图 22: Top-3 准确率变化曲线

对微调后的 GPT-SoVITS 模型进行效果评估，我重点关注了以下几个方面。首先，在多语种推理能力方面，GPT-SoVITS 展现出令人印象深刻的表现。给定中文、日文或英文文本，模型均能合成出与目标说话人（蓝毒）音色相似的语音，且在单一语言内的发音清晰度和自然度都达到了较高水平。在进行中日英混合文本输入时，模型也能较好地在不同语言间进行切换。

其次，一个显著的亮点是 GPT-SoVITS 在学习并复现说话人独特发音习惯方面的出色能力。在我的测试中，对于” 蓝毒” 这一角色在原生日语配音中存在的，将” わたし” 中的” た” 后多发出”k” 的音这一习惯，GPT-SoVITS 在合成对应文本时能够较好地模仿出来，使得克隆出的语音更具” 灵魂” 和辨识度。这表明 GPT-SoVITS 模型在捕捉和生成细微的语言和声学风格特征方面具有优势。

然而，GPT-SoVITS 在实践中也暴露出一些局限性。在处理较长的文本段落时，模型自动进行的断句、停顿以及整体韵律的控制有时显得不够智能或自然，会出现不合逻辑的停顿或者加快的情况。此外，对于包含特殊符号、非标准缩写或者特定领域术语的文本，模型的理解和合成效果不佳，出现读错或跳过等现象。这些问题可能与 GPT 模型本身的文本理解能力、韵律建模的复杂度以及在 Few-shot 场景下训练数据对复杂语言现象覆盖不足有关。相关的音频演示样本也请参见附录 A 中的音频示例列表。

5.5 通用音频与音乐生成

5.5.1 通用音频生成技术

通用音频生成 (Text-to-Audio, TTA) 要依据文本描述去合成涉及语音与音乐等各类声音，像环境噪声（如下雨声、街道嘈杂声）以及特定音效（如物体破碎声、动物叫声）等。TTA 技术在电影音效制作、游戏开发、虚拟现实环境构建以及给视障人士提供更丰富音频信息等方面有着关键应用价值，和 TTS 以及音乐生成相比，通用音频的声学特性更为多样复杂，并且一般缺少明确结构性，这给 TTA 模型给予了独特挑战。

近年来，随着深度生成模型的发展，TTA 技术也取得了显著进展。AudioGen[52] 是一个基于自回归 Transformer 的模型，它首先将音频编码为离散的声学 token 序列，然后训练一个 Transformer 模型根据文本描述自回归地生成这些 token 序列，最后通过一个声码器将 token 序列转换回音频波形。AudioGen 通过多流建模等技术来提升生成质量和对文本的依从性。AudioLDM[53] 则将潜在扩散模型 (Latent Diffusion Models, LDM) 的思想应用于 TTA。它首先利用一个预训练的对比语言-音频预训练 (Contrastive Language-Audio Pretraining, CLAP)[54] 模型将文本和音频分别映射到共享的潜在空间，然后在该潜在空间中训练一个扩散模型来学习从文本嵌入生成音频嵌入，最后通过一个声码器将生成的音频嵌入转换为音频波形。通过在潜在空间进行操作，AudioLDM 在保持较高生成质量的同时，也提高了计算效率。Bark[55] 是 Suno AI 开发的一个基于 Transformer 的文本提示生成音频模型，它不仅能够生成语音（包括多语种和不同说话人风格），还能生成音乐、背景噪声和简单的音效，展现了通用音频生成模型向多功能、多模态方向发展的趋势。这些模型通常依赖大规模的文本-音频对数据进行训练，并面临着如何更好地理解文本描述中的复杂语义、生成具有长时一致性和高保真度的音频，以及实现对生成音频更精细控制等挑战。

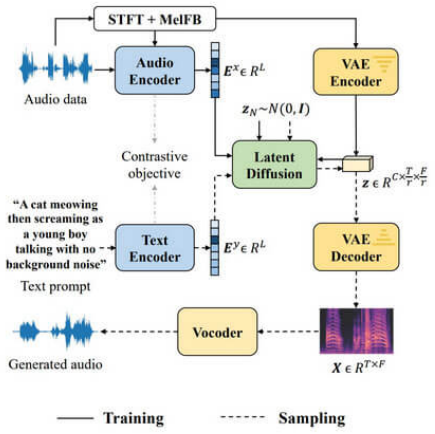

图 23: AudioLDM 结构图 [53]

5.5.2 AI 音乐生成技术

AI 音乐生成 (AI Music Generation) 的目标是根据文本描述、情感标签、风格指定、甚至简单的哼唱旋律等输入，自动创作出具有一定结构和美感的音乐作品。AI 音乐生成技术在辅助专业音乐人创作、为普通用户提供个性化音乐体验、自动化影视游戏配乐等方面具有巨大潜力，是 AIGC 领域最具创造性和挑战性的方向之一。

早期的 AI 音乐生成方法多基于规则或统计模型，生成的音乐往往较为简单或缺乏新意。现代 AI 音乐生成模型则广泛采用深度学习技术。MusicLM[56] 是 Google 提出的一个能够从文本描述生成高保真音乐的模型。它将音乐生成视为一个分层的序列到序列建模任务，能够生成长达数分钟且在 24kHz 采样率下保持一致性的音乐。MusicLM 不仅能根据文本生成音乐，还能根据图片并结合标题进行相应风格音乐的生成。MusicGen[57] 是 Meta AI 提出的一个单阶段 Transformer 语言模型，它直接在压缩的离散音乐表示上进行操作，能够根据文本描述或旋律特征生成高质量的单声道或立体声音乐，并强调了其模型的简洁性和可控性。这些模型通常需要在包含音乐音频及其对应文本描述（如风格、流派、乐器、情绪等标签）的大规模数据集上进行训练。

5.5.3 实践案例 3：QA-MDT 复现与基于 Prompt 的音乐片段生成

为探寻 AI 音乐生成技术的实际能力，本研究对中国科学技术大学所提出的 QA-MDT (Quality-Aware Masked Diffusion Transformer)[58] 模型进行了复现，还尝试借助不同的文本提示 (Prompt) 去引导模型生成不同风格的音乐片段，QA-MDT 的核心思想是引入质量感知机制，让模型于训练过程中可辨别音乐波形的质量，并且结合掩码扩散 Transformer 架构进行音乐生成，以此提升生成音乐的整体质量以及与文本描述的一致性。

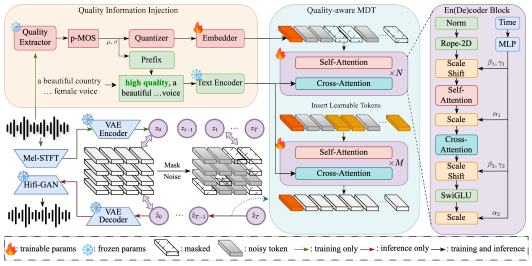

图 24: QA-MDT 结构图 [58]

复现 QA-MDT 模型代码的过程主要依赖其官方提供的开源实现。实验的硬件环境为配备 NVIDIA A40 GPU 的服务器，操作系统环境下配置了 CUDA 12.2。软件环境通过 Anaconda 进行管理，主要使用了 Python 3.10.13 和 PyTorch 2.3.0。由于 QA-MDT 通常需要在大规模音乐数据集上进行预训练，本次实践我主要利用了其提供的预训练模型。

我设计了两组对比鲜明的文本 Prompt 来测试 QA-MDT 的风格控制能力。Prompt 1 为：“soft synth chords, gentle pads, peaceful soundscape, slow rhythm, ethereal textures, warm ambiance, deep relaxation, sleep”，其意图是引导模型生成一段节奏舒缓、氛围宁静、织体空灵、适合放松或助眠的环境氛围音乐 (Ambient Music)。Prompt 2 为：“powerful breakbeats, rapid tempo, deep basslines, energetic synths, high-energy, dynamic intensity, explosive drops, futuristic”，其意图则是引导模型生成一段节奏强劲、充满能量、具有未来感的 Drum & Bass 风格的音乐。

由于 QA-MDT 的能力限制，我生成的音频片段长度均为 10 秒。对生成的两段音频样本进行主观听感分析，可以观察到模型在一定程度上响应了不同 Prompt 的引导。对于 Prompt 1，生成的音频片段具有较慢的节奏，使用了柔和的合成器音色，整体听感较为平和、舒缓，符合“peaceful”、“relaxation”等关键词的描述。对于 Prompt 2，生成的音频片段则具有更快的节奏，包含更具冲击力的鼓点、更深沉的贝斯线以及更具能量感的合成器主音或效果音，整体风格更偏向于“energetic”、“high-energy”。

然而，在 10 秒的短片段内，音乐的结构发展和复杂性有限。尽管风格有所区分，但生成的片段在旋律性、和声进行或织体层次感方面略显单薄或重复。这可能与模型的生成长度限制、预训练数据的多样性以及 Prompt 描述的精细程度有关。本次实践主要验证了 QA-MDT 根据文本提示生成不同风格音乐片段的基本能力。由于时间和计算资源的限制，我未能对 QA-MDT 模型在特定音乐风格数据集（如更长的 Ambient Music）上进行深入的微调。未来的工作可以探索通过微调来增强模型对特定风格的表达能力，并研究如何生成更长、结构更完整的音乐作品。相关的音频演示样本同样请参见附录 A 中的音频示例列表。

6 AI 生成音频的检测技术

随着 AIGC 技术的快速发展和广泛应用，其潜在的滥用风险也引起了社会各界的高度关注。特别是语音深度伪造技术，因其能够生成高度逼真、难以识别的特定人物语音，已被视为一种严重的安全威胁，可能被用于身份冒用、电信诈骗、传播虚假信息、甚至干预政治活动。此外，AI 生成的音乐和通用音频也可能涉及版权侵权、虚假内容传播等问题。因此，研究和开发可靠、高效的 AI 生成音频检测技术 [59]，对于维护信息安全、保护个人权益、规范 AIGC 技术的健康发展至关重要。

本章全面介绍了 AI 生成音频检测技术的发展现状和关键方法。首先，从语音反欺骗的角度回顾了 ASVspoof 挑战赛 [60][61] 这一重要评测平台在推动相关技术发展中的作用，分析了从传统统计方法到现代深度学习检测模型的技术发展历程。接着，深入分析了基于深度学习的语音 Deepfake 检测模型的原理和结构，重点介绍了 RawNet2[62]、RawGAT-ST[63]、AASIST[64] 等代表性模型及其技术特点。随后，探讨了更具挑战性的 AI 生成音乐与通用音频检测问题，分析了这一新兴领域面临的独特困难和技术瓶颈。最后，详细描述了作者在 AASIST 模型复现方面的实践工作，通过具体的实验验证了先进检测模型的性能表现，并总结了实践过程中的关键发现和技术见解。

6.1 语音反欺骗与 ASVspoof 挑战赛

语音反欺骗技术的核心目的在于让自动说话人验证 (Automatic Speaker Verification, ASV) 系统有辨别真实 (bona fide) 语音与各类伪造 (spoofed) 语音的能力。伪造语音有多种类型，比如 TTS、VC、重放攻击 (Replay Attack) 等。随着 AIGC 技术不断发展，基于深度学习的 TTS 和 VC 系统生成的语音变得日益逼真，对 ASV 系统的安全性造成了严峻挑战。

ASVspoof 系列挑战赛 [60][61] 在国际上是推动语音反欺骗技术发展的极具权威性的评测平台，从 2015 年首次举办起，ASVspoof 就会定期发布由当前先进伪造技术生成的大规模语音数据集，还会制定统一的评估协议以及性能指标（如 EER (Equal Error Rate) 和 min-tDCF (minimum Tandem Detection Cost Function)[65]），为全球研究者打造了一个能公平比较与验证反欺骗算法性能的基准。每一届 ASVspoof 挑战赛都会引入新的攻击类型以及更具挑战性的评估条件，持续引领该领域的研究方向。

例如，ASVspoof 2019[60] 首次将逻辑访问 (Logical Access, LA) 场景下的 TTS 和 VC 攻击与物理访问 (Physical Access, PA) 场景下的重放攻击纳入统一评估框架，其 LA 场景的数据集包含了当时各种主流和先进的语音合成与转换技术生成的样本，对检测算法提出了很高

的要求。而最新的 ASVspoof 5[61] 则进一步提升了挑战难度，其特点包括：1) 引入了更大规模、更多样化的众包 (crowdsourced) 语音数据，覆盖了更广泛的声学环境和说话人；2) 关注了针对检测系统本身的对抗攻击 (Adversarial Attacks)，要求检测算法具备更强的鲁棒性；3) 引入了欺骗鲁棒的说话人验证 (Spoofing-Aware Speaker Verification, SASV) 任务，旨在评估检测模块与 ASV 系统的联合性能。ASVspoof 的持续举办极大地促进了学术界和产业界在语音反欺骗技术上的投入和创新。

6.2 语音 Deepfake 检测模型

语音 Deepfake 检测模型的技术发展路径大致经历了从传统的基于手工提取声学特征结合机器学习分类器的方法，到现代的基于端到端深度学习模型自动学习判别特征的演进。

早期的检测方法通常依赖于分析真实语音和伪造语音在声学特性上可能存在的差异。研究者们尝试利用各种声学特征，如 MFCCs[1]、线性预测倒谱系数 (Linear Prediction Cepstral Coefficients, LPCCs)、常 Q 倒谱系数 (Constant Q Cepstral Coefficients, CQCCs)、线性频率倒谱系数 (Linear Frequency Cepstral Coefficients, LFCCs) 等，来捕捉语音信号的谱包络、细节以及可能由合成过程引入的失真或伪影。然后，将这些提取到的特征输入到传统的机器学习分类器中，如 GMM、支持向量机 (Support Vector Machines, SVM) 或逻辑回归，进行真伪分类。这类方法的优点是可解释性相对较好，但其性能高度依赖于所选特征的有效性和分类器的能力，对于日益逼真的深度伪造语音，其检测能力往往有限。

随着深度学习的兴起，基于神经网络的端到端检测模型逐渐成为主流。这类模型能够直接从原始音频波形或其频谱表示（如语谱图）中自动学习区分真伪语音的关键特征，避免了复杂且可能次优的手工特征工程。例如，RawNet2[62] 是一个基于 CNN 和 GRU 的端到端模型，它直接将原始音频波形作为输入，通过多层卷积和循环连接来学习语音的底层声学表示，并进行分类。这种直接处理原始波形的方式有助于捕捉可能在传统特征提取过程中丢失的细微伪造痕迹。

另一类表现出色的模型是基于图神经网络 (Graph Neural Networks, GNN) 的方法。例如，RawGAT-ST[63] 通过将语音的语谱图视为一个图结构，其中图节点代表时频谱单元，然后利用图注意力网络 (Graph Attention Networks, GAT) 来建模这些单元之间的时域和频域相关性，从而更有效地捕获伪造语音中可能存在的局部和全局不一致性或伪影。AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks)[64] 及其后续改进版本 AASIST3(引入了 KAN (Kolmogorov-Arnold Networks)[66] 以提高特定网络层的功能)[67] 拓展了这一思路，借助更精巧的图结构设计与注意力机制，达成了在 ASVspoof 等评测任务上的领先表现，比如 AASIST 运用异构注意力机制和堆栈节点来学习跨异构时间和频谱间隔的欺骗伪影特征，捕捉更全面的欺骗线索。像 USTC-KXDIGIT 系统 [68] 等在 ASVspoof 竞赛中取得优异成绩的系统，一般还会采用多模型融合、数据增强、对抗训练等策略提高检测系统的整体性能与鲁棒性，Transformer 和 Conformer 等强大序列建模架构也越来越多地被用于语音反欺骗任务中，依靠其自注意力机制捕捉伪造语音中的长程依赖与细微不一致性。

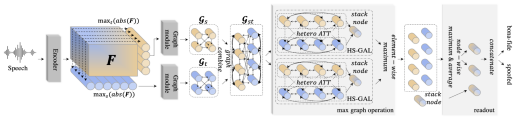

图 25: AASIST 结构图 [64]

6.3 AI 生成音乐与通用音频检测

与发展相对成熟的语音 Deepfake 检测相比，针对 AI 生成音乐和通用音频的检测技术研究尚处于起步阶段，面临着其独有的挑战。首先，音乐和通用音效的内容、风格、结构和声学特性远比人类语音更加多样化和复杂，这使得定义” 真实” 与” 伪造” 的边界更加模糊。例如，AI 辅助音乐创作工具生成的音乐片段，与完全由 AI 独立生成的音乐，其” 伪造” 程度和检测难度可能存在显著差异。其次，音乐创作本身就充满了模仿、借鉴和风格融合，如何区分 AI 的” 创造性模仿” 与对现有版权作品的” 侵权性复制” 是一个难题。再者，目前国际上缺乏像 ASVspoof 那样针对 AI 生成音乐和通用音频的大规模、标准化、多样化的评测数据集和统一的评估指标 [69]，这在一定程度上阻碍了相关检测算法的快速发展和公平比较。

即便如此，研究者们已然着手探索适用于 AI 生成音乐以及通用音频的检测方法 [59]，其中一种思路是参照语音或图像 Deepfake 检测的技术路径，尝试把现有的基于深度学习的分类模型迁移至新的音频类型上，不过这一般需要依据音乐和通用音频的特性来进行模型调整以及特征选择。另外一种思路是着重分析音乐自身的特性，比如，检测 AI 生成音乐在旋律连贯性、和声进行合理性、节奏稳定性、乐器音色真实性、音乐结构完整性等方面是否会存在与人类创作习惯不符的统计规律或者异常模式，研究特定生成模型 [56][57][58] 可能在生成结果中留下的固有特征或者特定类型的伪影，同样是一个潜在的检测方向。构建涉及多种 AI 生成算法、多种风格以及高质量真实样本的专用检测数据集，对推动这一领域的研究十分重要。

6.4 实践案例 4：AASIST 模型复现与性能评估

为能更透彻地明白当下先进语音反欺骗检测模型的工作原理以及实际性能状况，本研究针对在 ASVspoof 挑战赛里表现出色的 AASIST[64] 系列模型展开了复现与评估工作，并且把该系列模型的性能和基线模型做了对比。

本次实验所使用的数据集为 ASVspoof 2019 LA track[60] 的官方划分，该数据集包含了多种类型的真实语音和由 19 种不同 TTS 及 VC 算法生成的伪造语音，是语音反欺骗领域广泛使用的标准评测基准。实验的硬件环境和基本软件环境与 QA-MDT 复现实验相同，主要使用了 Python 3.8.20 和 PyTorch 2.4.1。我首先评估了 AASIST 官方提供的预训练模型，包括标准版 AASIST 和轻量版 AASIST-L。随后，我选择了两个在 AASIST 文献中被用作基线的端到端模型进行对比：RawNet2[62] 和 RawGAT-ST[63]。最后，我基于 AASIST 的开源代码，自行对 AASIST-L 模型在 ASVspoof 2019 LA 训练集上进行了重新训练，以验证复现效果并进一步熟悉模型的训练过程和参数设置。

本实践采用语音反欺骗领域常用的两个主要评估指标：EER 和 min-tDCF[65]。EER 是指在检测系统的判决阈值设定在使得虚警率 (False Alarm Rate, FAR) 等于漏检率 (False Rejection Rate, FRR) 时的错误率，该值越低表示检测性能越好。min-tDCF 则是一个更侧重于实际应用场景的指标，它考虑了 ASV 系统本身的性能以及反欺骗模块的引入对整体系统性能的影响，同样是值越低越好。

在自行训练 AASIST-L 模型的过程中，我利用 TensorBoard 对训练过程进行了监控，主要观察训练集损失函数 (Training Loss) 的下降情况以及在开发集 (Development Set) 上的 EER 和 min-tDCF 指标的变化。训练过程的可视化结果如图 26-28 所示：

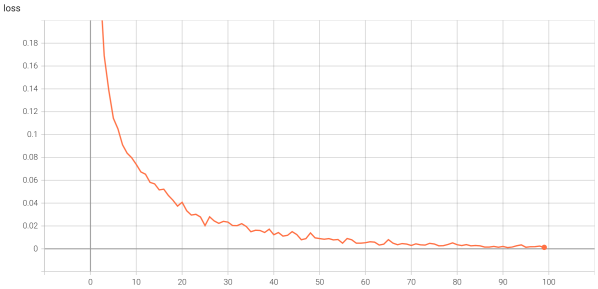

图 26: 训练集损失函数变化曲线

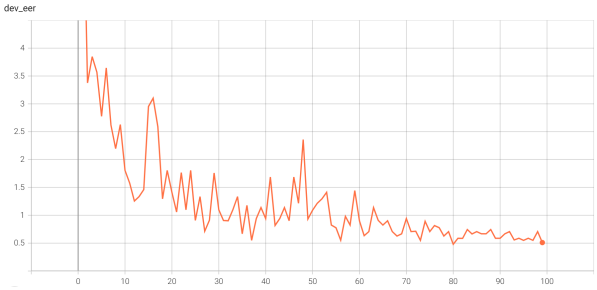

图 27: 开发集 EER 变化曲线

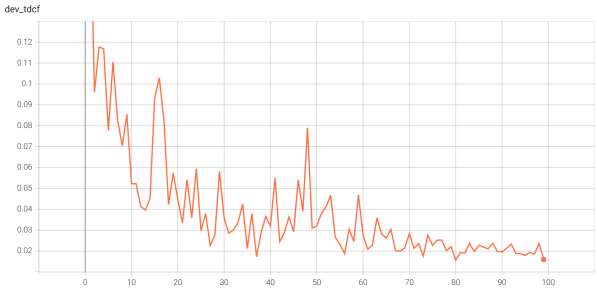

图 28: 开发集 min-tDCF 变化曲线

表 1: 各模型在 ASVspoof 2019 LA 评估集上的性能对比

模型	EER (%)	min-tDCF
RawNet2	4.48	0.1209
RawGAT-ST	1.42	0.0427
AASIST	0.83	0.0275
AASIST-L	0.99	0.0309
AASIST-L (自训练)	1.37	0.0422

实验结果显示，官方提供的 AASIST 和 AASIST-L 模型在 ASVspoof 2019 LA 评估集上都取得了与原论文报告结果相当的出色性能，证实了它们强大的伪造语音检测能力。我自行训练的 AASIST-L 模型，虽然由于训练环境、随机种子等因素与官方预训练模型在具体数值上存在可接受的偏差，但其性能仍优于所选的两个基线模型。如表 1 所示，在 EER 指标上，自训练 AASIST-L 达到了 1.37%，显著优于 RawNet2 的 4.48%，与 RawGAT-ST 的 1.42% 相当；在 min-tDCF 指标上，自训练 AASIST-L 达到了 0.0422，同样优于 RawNet2 的 0.1209 和 RawGAT-ST 的 0.0427。

通过本次 AASIST 模型的复现与评估实践，我不仅成功验证了该模型在标准语音反欺骗任务上的有效性，也深化了对当前先进检测模型的网络结构、训练方法以及影响模型性能关键因素的理解。例如，AASIST 通过图注意力网络有效融合时频谱信息和时域信息的策略，对于捕获细微的伪造痕迹展现出良好效果。同时也认识到，即使是相同的模型和数据集，不同的训练环境也可能导致最终性能的差异，这对于后续进行更深入的研究和模型优化具有指导意义。此外，GitHub 等平台上也存在许多关于语音 Deepfake 检测的汇总性资源，如 media-sec-lab 的仓库 Audio-Deepfake-Detection[70]，为研究者提供了便利。

7 总结与展望

凭借对语音与音频领域人工智能生成以及检测技术展开系统性回顾，对核心模型原理给予深入剖析，介绍代表性应用，呈现作者相关实践探索，本文针对该快速发展的交叉领域进行了阶段性总结，同时对未来面临的挑战以及潜在研究方向作出展望。

7.1 技术发展总结

在 AIGC 生成技术方面，语音与音频领域经历了从早期基于统计参数的 SPSS 方法 [4] 到现代深度学习模型的深刻变革。深度学习的引入，首先是 RNN/LSTM/GRU[6][7][8] 在序列建模上的应用，随后 GAN[10]、VAE[12]、流模型 [14][15][17] 等生成框架为语音合成和转换带来了新的思路。而 Transformer 架构 [30] 的出现，凭借其强大的并行处理能力和对长距离依赖的捕捉，彻底革新了序列到序列任务，并迅速成为现代 TTS[41][42][43]、VC 以及音乐/音频生成 [57][52] 的核心。大规模自监督预训练范式 [33][34] 的成功，使得模型能够从未标注数据中学习到丰富的声学和语义表示，显著提升了下游 AIGC 任务的性能和对少样本数据的适应性。最新的 Diffusion Models[36][37] 则在生成音频的质量和多样性方面达到了新的高度 [40][58]。总而言之，语音音频 AIGC 技术正朝着更高保真度、更强表现力、更优可控性和更广泛应用场景的方向快速演进。作者的实践也验证了 RVC[45]、GPT-SoVITS[47] 等开源模型在 Few-shot 声音转换和风格学习上的潜力，以及 QA-MDT[58] 在基于 Prompt 生成不同风格音乐片段方面的能力。

在 AI 生成音频的检测技术方面，其发展始终与生成技术的进步相伴相生。早期针对传统合成方法（如拼接或参数合成）的检测主要依赖于分析声学特征的统计差异或特定伪影。随着深度伪造语音的出现，检测技术也迅速转向基于深度学习的方法。ASVspoof 系列挑战赛 [60][61] 极大地推动了该领域的发展，催生了如 RawNet2[62]、RawGAT-ST[63]、AASIST 及其后续版本 [64][67] 等先进的检测模型。这些模型通过直接处理原始波形、利用图神经网络捕捉时频谱关系等方式，不断提升对各种伪造攻击的检测精度。作者对 AASIST 的复现实践也验证了其相较于基线模型的优越性能。然而，随着生成技术的不断迭代，检测技术依然面临着对未知攻击泛化能力不足、对真实世界干扰鲁棒性不强以及易受对抗攻击等严峻挑战。针对 AI 生成音乐和通用音频的检测研究 [69][59] 虽然尚在起步，但也开始受到关注。

7.2 当前挑战

虽然语音与音频 AIGC 和它的检测技术已经有了一定成果，然而当下还是存在不少急需解决的难题。

在 AIGC 生成技术方面的主要挑战包括：

1) **语义理解与生成对齐**：当前模型虽然能根据文本生成音频，但对文本中复杂语义、细微情感、隐含意图的理解仍有不足，导致生成的音频内容有时与用户期望存在偏差。如何实现更深层次的语义理解和更精准的跨模态对齐是一个核心难题。

2) **长时结构与内容连贯性**：尤其在生成较长篇幅的语音（如演讲、有声读物）或音乐作品时，保持风格、主题、情感和逻辑的全局一致性与连贯性，避免重复、突兀转变或结构涣散，仍然是一个巨大的挑战。

3) **可控性与可编辑性的精细化**：用户往往期望对生成音频的各种属性（如 TTS 的语速、语调、情感强度、口音风格；VC 的音色混合度、转换强度；音乐生成的乐器组合、节奏型、和声进行等）进行细粒度的、可预测的、解耦的控制。同时，对已生成音频进行便捷、无痕的后期编辑也是重要的需求，而当前多数模型仍接近“黑箱”操作。

4) **数据瓶颈与效率问题**：高质量 AIGC 模型的训练通常依赖大规模、多样化、高保真且带有精细标注（如文本-音频对齐、音乐描述、情感标签等）的数据集，这类数据的获取成本高昂。如何在小样本、零样本甚至无监督的条件下实现高质量的 AIGC 是重要的研究方向。此外，许多先进生成模型（如大型 Transformer、Diffusion 模型）的训练和推理计算开销巨大，限制了其在资源受限设备上的部署和应用。

5) **客观评估与主观感知**：如何客观、全面、且与人类主观感知一致地评估生成音频的质量（如自然度、真实感、情感表达准确性、音乐性等）仍然是一个开放性问题。现有的客观指标（如 MOS）往往难以完全捕捉人类听觉的复杂感受。

在 AI 生成音频的检测技术方面的主要挑战包括：

1) **泛化能力不足**：当前检测模型往往在训练时见过或相似的攻击类型上表现良好，但当遇到由全新、未知的生成算法产生的伪造音频，或来自不同声学环境、不同语言、不同数据分布的样本时，其性能往往会显著下降。

2) **鲁棒性差**：真实世界中的音频信号在传播和存储过程中，不可避免地会受到各种失真（如背景噪声、信道传输效应、有损压缩、重采样、混响等）的影响。许多检测模型对这些常见的真实世界干扰非常敏感，导致其在实验室环境下的高性能难以在实际应用中复现。

3) **对抗攻击的脆弱性**：如同其他深度学习模型一样，AI 音频检测模型易于遭受经过精心设计的对抗样本的攻击，攻击者可凭借在伪造音频里添加微小扰动，这种扰动是人耳难以察觉到的，轻易地欺骗检测系统，致使检测系统把伪造音频错误地判断为真实音频。

4) **可解释性与可信度缺失**：多数深度学习检测模型

类似“黑箱”，其作出判断所依据的内容以及内部的决策过程大多时候难以得到解释，这种情况妨碍了人们对于模型出现错误原因的理解，也不利于进行针对性的改进，同时还降低了用户对检测结果的信任程度，对构建可信的 AI 系统产生不利影响。

5) **数据不平衡与多样性匮乏**：高质量的伪造音频样本（特别是针对最新生成技术的）获取难度大，且真实语音与各类伪造语音之间的数据量往往存在严重不平衡，这给模型的训练带来了困难。同时，训练数据对各种可能的生成技术、说话人、语言、环境覆盖不足，也会限制模型的泛化能力。

7.3 未来研究方向

面对上述挑战，并结合当前技术发展趋势，未来语音与音频 AIGC 及其检测技术的研究可以从以下几个方面重点突破：

在基础模型与算法创新层面：

1) **探索新型网络架构**：持续关注和探索超越传统 Transformer 的新型序列建模架构。例如，状态空间模型 (State Space Models, SSM) 如 Mamba[71]，通过引入选择性状态压缩机制，在保持对长序列建模能力的同时，实现了线性计算复杂度，为处理超长音频序列提供了新的思路。RWKV[72] 模型尝试结合 RNN 的线性复杂度推理与 Transformer 的并行训练优势。而 KAN[66] 则提出了一种与传统多层感知机 (Multi-Layer Perceptron, MLP) 不同的参数化方式（将激活函数置于边上并使其可学习），能在提升模型可解释性和逼近复杂函数方面带来新的视角。将这些新兴架构应用于语音音频领域，有望在效率、性能和可解释性上取得新的突破。

2) **深度多模态融合与协同生成**：未来的 AIGC 将更强调多模态信息的深度融合。例如，在 TTS/VC 中，结合面部表情、口型运动等视觉信息，生成更具表现力和真实感的视听同步语音；在音乐生成中，融合歌词文本、情感标签、甚至相关的图像或视频场景，创作出更符合整体意境的音乐。反过来，音频信息也可以驱动其他模态内容的生成。这需要研究更有效的跨模态特征表示、对齐和融合机制。

3) **强化学习与人类对齐**：将强化学习 (Reinforcement Learning, RL)，特别是基于人类反馈的强化学习 (Reinforcement Learning from Human Feedback, RLHF) 应用于 AIGC 模型的优化，有望更好地将模型的生成结果与人类的复杂偏好、主观评价标准以及特定任务的细微要求对齐，从而提升生成内容的用户满意度和实用性，解决客观评估指标与主观感受的偏差问题。

在 AIGC 生成技术前沿探索层面：

1) **逼真与富有表现力的语音合成/转换**：追求超越当前水平的语音自然度、情感表现细腻度、风格多样性以及对细微韵律特征的精准控制。探索在极低资源（如一次性发声 few-shot/zero-shot ）条件下实现高质量、个性化的声音克隆与转换。研究超低延迟的实时交互式语音生成技术，以支持更自然的对话系统和虚拟人交互。

2) **结构化与可控的 AI 音乐创作**：发展能够生成具有复杂音乐结构（如乐句、乐段、多声部织体）、符合音乐理论（如和声、调性、配器）、并具有长时一致性的 AI 音乐创作系统。增强模型对音乐风格、情感、乐器组合、节奏型等高级属性的细粒度控制能力。探索交互式音乐生成，允许用户参与到创作过程中，通过迭代修改和引导，共同完成音乐作品。

3) **高效轻量化与端侧部署**：针对移动设备、物联网终端等资源受限场景，研究模型压缩、知识蒸馏、量化、剪枝等技术，开发高效、轻量化的 AIGC 模型，以实现低功耗、低延迟的端侧智能音频生成与处理。

在 AI 音频检测技术与主动防御层面：

1) **提升泛化性与鲁棒性的检测模型**：研究能够更好泛化到未知攻击类型和真实世界复杂声学环境的检测方法。这可能包括利用元学习 (Meta-learning)、领域自适应 (Domain Adaptation)、测试时训练 (Test-Time Training, TTT)[73] 等技术提高模型对分布外数据的适应能力。探索对各种音频失真（压缩、噪声、混响）更不敏感的鲁棒声学特征和网络结构。

2) **对抗攻击的防御与检测**：深入研究针对 AI 音频检测模型的对抗攻击机制，并开发相应的防御策略，如对抗训练、梯度掩蔽、输入变换等，以增强检测系统的安全性。同时，研究专门检测对抗样本的技术。

3) **可解释的检测与伪影分析**：开发可解释的人工智能检测模型，该模型要可判定真伪，还要可指出伪造音频里或许存在的具体伪影或者不一致之处，以此为人工核查提供相应依据，指导生成模型进行改进。

4) **主动防御技术与数字音频水印**：大力发展和应用数字音频水印 (Audio Watermarking) 技术 [74]。研究如何在 AIGC 模型生成音频的阶段，嵌入对人类听觉透明（不可感知或不影响听感）但能被机器检测的“水印”信息。这种水印可以用于标识音频的 AI 生成来源、原始创作者、允许的使用范围等，从而为后续的版权保护、内容溯源和真实性验证提供可靠的技术手段，实现对 AIGC 内容的主动、可信管理。

7.4 伦理考量与社会影响

语音与音频 AIGC 技术的飞速发展在带来巨大机遇的同时，也引发了一系列深刻的伦理、法律和社会问题，需要我们进行前瞻性的思考和负责任的应对。

首先，**语音深度伪造的治理**是一个极其复杂且紧迫的问题。这不仅需要持续投入研发更先进的检测和溯源技术，更需要法律法规的完善、平台主体责任的落实、以及公众媒介素养和辨别能力的提升。如何界定 Deepfake 的恶意使用与合理应用（如影视制作中的修

复），如何追究恶意制作者和传播者的法律责任，如何平衡言论自由与信息安全，都是亟待解决的难题。

其次，**AI 生成内容的版权与原创性**问题日益突出。就那些由 AI 独立生成或者是在人类指导之下生成的音乐、有声读物等音频内容而言，其版权究竟应该归属于何处呢？AI 到底能不能被看作是主体享有著作权呢？要是 AI 生成的音乐和现有的作品在旋律、和声或者是风格等方面极为相似，这是否会构成侵权行为呢？这些问题给现有的版权法律体系给予了十分严峻的挑战，并且很可能会对人类艺术家的创作积极性以及整个创意产业的生态造成较为深远的影响。有必要去探索全新的版权认定规则以及利益分配机制，以此来适应 AIGC 时代内容创作的新型模式。

再次，**用户数据的隐私与安全**保护面临新的考验。许多 AIGC 应用（尤其是个性化声音克隆、语音助手等）需要收集和处理用户的语音数据，这些数据属于高度敏感的生物特征信息。如何确保这些数据在采集、存储、使用过程中的安全合规，防止数据泄露、滥用或被用于未经授权的身份模拟，是技术开发者和应用提供商必须承担的重要责任。

此外，还需要关注**技术偏见与社会公平性**问题。要是训练 AIGC 模型所使用的数据存在偏差（例如，在语种、口音、性别、年龄、族裔等方面分布不均），那么模型或许会习得并放大这些偏见，使得其在某些群体上的性能表现欠佳（如特定口音的语音识别率低，或生成的语音带有刻板印象），甚至产生歧视性结果。这可能加剧数字鸿沟，造成新的社会不公。因此，在模型设计、数据收集和算法评估等环节，都需要充分考虑公平性和包容性，努力消除或减轻潜在的技术偏见。

面对语音与音频 AIGC 技术带来的深刻变革，应秉持**负责任人工智能 (Responsible AI)** 的原则，并将该原则贯穿于技术研发、产品设计、应用部署以及法规制定的整个过程，此过程需要学术界、产业界、政府部门、法律界以及社会公众共同付出努力，去建立完善的伦理规范指引、行业自律标准、技术安全评估体系以及有效的法律监管框架，以此保证 AIGC 技术的发展可切实服务于人类福祉，推动社会进步，同时最大程度地规避其潜在风险，达成科技向善的目标。

参考文献

[1] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE transactions on acoustics, speech, and signal processing, 1980, 28(4): 357 – 366.

[2] KUMAR A, NI Z. Torchaudio-Squim: Non-intrusive Speech Assessment in TorchAudio[EB/OL]. 2024. https://docs.pytorch.org/audio/2.6.0/tutorials/squim_tutorial.html.

[3] PICONE J W. Signal modeling techniques in speech recognition[J]. Proceedings of the IEEE, 1993, 81(9): 1215 – 1247.

[4] ZEN H, TOKUDA K, BLACK A W. Statistical parametric speech synthesis[J]. speech communication, 2009, 51(11): 1039 – 1064.

[5] BÄCKSTRÖM T, RÄSÄNEN O, ZEWOUDIE A, et al. Introduction to Speech Processing[M/OL]. 2nd ed. 2022. https://speechprocessingbook.aalto.fi.

[6] ELMAN J L. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179 – 211.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735 – 1780.

[8] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[9] OLAH C. Understanding LSTM Networks[EB/OL]. 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[10] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[11] BONDARENKO K. Using GAN for fashion[EB/OL]. 2020. https://medium.com/analytics-vidhya/using-gan-for-fashion-6dc11012ac6a.

[12] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[13] LEE H. Machine Learning (2021 Spring)[EB/OL]. 2021. https://speech.ee.ntu.edu.tw/~hylee/ml/2021-spring.php.

[14] REZENDE D, MOHAMED S. Variational inference with normalizing flows[J]. International conference on machine learning, 2015: 1530 – 1538.

[15] DINH L, KRUEGER D, BENGIO Y. Nice: Non-linear

independent components estimation[J]. arXiv preprint arXiv:1410.8516, 2014.

[16] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real nvp[J]. arXiv preprint arXiv:1605.08803, 2016.

[17] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1x1 convolutions[J]. Advances in neural information processing systems, 2018, 31.

[18] WENG L. Flow-based Deep Generative Models[EB/OL]. 2018.
https://lilianweng.github.io/posts/2018-10-13-flow-models/.

[19] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016, 12.

[20] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[J]. International Conference on Machine Learning, 2018 : 2410 – 2419.

[21] PRENGER R, VALLE R, CATANZARO B. Waveglow: A flow-based generative network for speech synthesis[J]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019 : 3617 – 3621.

[22] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.

[23] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[J]. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018 : 4779 – 4783.

[24] HSU C-C, HWANG H-T, WU Y-C, et al. Voice conversion from non-parallel corpora using variational auto-encoder[J]. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 : 1 – 6.

[25] KANEKO T, KAMEOKA H. Parallel-data-free voice conversion using cycle-consistent adversarial networks[J]. arXiv preprint arXiv:1711.11293, 2017.

[26] ZHU J-Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. Proceedings of the IEEE international conference on computer vision, 2017 : 2223 – 2232.

[27] KAMEOKA H, KANEKO T, TANAKA K, et al. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks[J]. 2018 IEEE Spoken Language Technology Workshop (SLT), 2018 : 266 – 273.

[28] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018 : 8789 – 8797.

[29] KANEKO T, KAMEOKA H, TANAKA K, et al. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion[J]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019 : 6820 – 6824.

[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[31] GULATI A, QIN J, CHIU C-C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.

[32] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. 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 : 4171 – 4186.

[33] HSU W-N, BOLTE B, TSAI Y-H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29 : 3451 – 3460.

[34] CHEN S, WANG C, CHEN Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022,

16(6) : 1505 – 1518.

[35] CHEN S, WU Y, WANG C, et al. Beats: Audio pre-training with acoustic tokenizers[J]. arXiv preprint arXiv:2212.09058, 2022.

[36] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33 : 6840 – 6851.

[37] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

[38] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in neural information processing systems, 2020, 33 : 17022 – 17033.

[39] TEAM P. HiFi GAN - PyTorch Hub[EB/OL]. 2025. https://pytorch.org/hub/nvidia_deeplearningexamples_hifigan/.

[40] KONG Z, PING W, HUANG J, et al. Diffwave: A versatile diffusion model for audio synthesis[J]. arXiv preprint arXiv:2009.09761, 2020.

[41] LI N, LIU S, LIU Y, et al. Neural speech synthesis with transformer network[J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(01) : 6706 – 6713.

[42] REN Y, RUAN Y, TAN X, et al. Fastspeech: Fast, robust and controllable text to speech[J]. Advances in neural information processing systems, 2019, 32.

[43] REN Y, HU C, TAN X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv preprint arXiv:2006.04558, 2020.

[44] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[J]. International Conference on Machine Learning, 2021 : 5530 – 5540.

[45] RVC-PROJECT. Retrieval-based Voice Conversion WebUI[EB/OL]. 2025. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI.

[46] svc-develop TEAM. so-vits-svc[EB/OL]. 2025. https://github.com/svc-develop-team/so-vits-svc.

[47] RVC-BOSS. GPT-SoVITS[EB/OL]. 2025. https://github.com/RVC-Boss/GPT-SoVITS.

[48] YXLLLC. DDSP-SVC[EB/OL]. 2025. https://github.com/yxlllc/DDSP-SVC.

[49] ENGEL J, HANTRAKUL L, GU C, et al. DDSP: Differentiable digital signal processing[J]. arXiv preprint arXiv:2001.04643, 2020.

[50] ZFTURBO. Music-Source-Separation-Training[EB/OL]. 2025. https://github.com/ZFTurbo/Music-Source-Separation-Training.

[51] WEI H, CAO X, DAN T, et al. RMVPE: A robust model for vocal pitch estimation in polyphonic music[J]. arXiv preprint arXiv:2306.15412, 2023.

[52] KREUK F, SYNNAEVE G, POLYAK A, et al. Audiogen: Textually guided audio generation[J]. arXiv preprint arXiv:2209.15352, 2022.

[53] LIU H, CHEN Z, YUAN Y, et al. Audioldm: Text-to-audio generation with latent diffusion models[J]. arXiv preprint arXiv:2301.12503, 2023.

[54] ELIZALDE B, DESHMUKH S, AL ISMAIL M, et al. Clap learning audio concepts from natural language supervision[J]. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023 : 1 – 5.

[55] suno AI. bark[EB/OL]. 2025. https://github.com/suno-ai/bark.

[56] AGOSTINELLI A, DENK T I, BORSOS Z, et al. Musiclm: Generating music from text[J]. arXiv preprint arXiv:2301.11325, 2023.

[57] COPET J, KREUK F, GAT I, et al. Simple and controllable music generation[J]. Advances in Neural Information Processing Systems, 2023, 36 : 47704 – 47720.

[58] LI C, WANG R, LIU L, et al. QA-MDT: Quality-aware Masked Diffusion Transformer for Enhanced Music Generation[J]. arXiv preprint arXiv:2405.15863, 2024.

[59] LI Y, MILLING M, SPECIA L, et al. From Audio Deepfake Detection to AI-Generated Music Detection–A Pathway and Overview[J]. arXiv preprint arXiv:2412.00571, 2024.

[60] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech & Language, 2020, 64 : 101114.

[61] WANG X, DELGADO H, TAK H, et al. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale[J]. arXiv preprint arXiv:2408.08739, 2024.

[62] TAK H, PATINO J, TODISCO M, et al. End-to-end anti-spoofing with rawnet2[J]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021 : 6369 – 6373.

[63] TAK H, JUNG J-W, PATINO J, et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection[J]. arXiv preprint arXiv:2107.12710, 2021.

[64] JUNG J-W, HEO H-S, TAK H, et al. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks[J]. ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2022 : 6367 – 6371.

[65] KINNUNEN T, LEE K A, DELGADO H, et al. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification[J]. arXiv preprint arXiv:1804.09618, 2018.

[66] LIU Z, WANG Y, VAIDYA S, et al. Kan: Kolmogorov-arnold networks[J]. arXiv preprint arXiv:2404.19756, 2024.

[67] BORODIN K, KUDRYAVTSEV V, KORZH D, et al. AASIST3: KAN-Enhanced AASIST Speech Deepfake Detection using SSL Features and Additional Regularization for the ASVspoof 2024 Challenge[J]. arXiv preprint arXiv:2408.17352, 2024.

[68] CHEN Y, WU H, JIANG N, et al. USTC-KXDIGIT System Description for ASVspoof5 Challenge[J]. arXiv preprint arXiv:2409.01695, 2024.

[69] AFCHAR D, MESEGUER-BROCAL G, HENNEQUIN R. AI-Generated Music Detection and its Challenges[J]. arXiv preprint arXiv:2501.10111, 2025.

[70] media-sec LAB. Audio-Deepfake-Detection[EB/OL]. 2025. `https://github.com/media-sec-lab/Audio-Deepfake-Detection`.

[71] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.

[72] PENG B, ALCAIDE E, ANTHONY Q, et al. Rwkv: Reinventing rnns for the transformer era[J]. arXiv preprint arXiv:2305.13048, 2023.

[73] SUN Y, WANG X, LIU Z, et al. Test-time training with self-supervision for generalization under distribution shifts[J]. International conference on machine learning, 2020 : 9229 – 9248.

[74] ROMAN R S, FERNANDEZ P, DÉFOSSEZ A, et al. Proactive detection of voice cloning with localized watermarking[J]. arXiv preprint arXiv:2401.17264, 2024.

A 音频示例列表

本文档中提及的音频示例均可在作者的个人网站上收听与体验：

https://bluerevo.top/audio/

以下是音频示例的清单，方便读者对照查阅。

A.1 RVC 项目的歌声转换音频

以下是使用 RVC 项目进行的歌声转换音频示例，包括一首日语歌曲和一首英语歌曲。每首歌曲都提供了带背景音乐（BGM）和干声两个版本。

日语歌曲（カナデトモスソラ）

- **带 BGM 版本：** RVC 歌声转换 (日语, 带 BGM)
- **干声版本：** RVC 歌声转换 (日语, 干声)

英语歌曲（Beyond Your Words）

- **带 BGM 版本：** RVC 歌声转换 (英语, 带 BGM)
- **干声版本：** RVC 歌声转换 (英语, 干声)

A.2 GPT-SoVITS 项目的生成语音音频

以下是使用 GPT-SoVITS 项目生成的两个语音音频示例。

音频 1 (多语种)

语音内容：这里是明日方舟的 Blue Poison です。私は Artificial Intelligence 由金天昊训练。

- GPT-SoVITS 语音生成 (多语种)

音频 2 (长难句)

语音内容：Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

- GPT-SoVITS 语音生成 (长难句)

A.3 QA-MDT 项目复现的音频

以下是复现 QA-MDT 项目生成的音频示例，基于不同的文本 Prompt。

音频 1 (Prompt 1)

Prompt 内容：soft synth chords, gentle pads, peaceful soundscape, slow rhythm, ethereal textures, warm ambiance, deep relaxation, sleep

- QA-MDT 音频复现 (Prompt 1)

音频 2 (Prompt 2)

Prompt 内容：powerful breakbeats, rapid tempo, deep basslines, energetic synths, high-energy, dynamic intensity, explosive drops, futuristic

- QA-MDT 音频复现 (Prompt 2)

致　　谢

感谢我的科技写作与表达老师俞能海教授和储琪老师，在研究过程中给予的悉心指导和宝贵建议。感谢凌震华老师和杜俊老师在我对语音与音频方向感兴趣时提供交流机会。同时也感谢国立台湾大学李宏毅老师的机器学习课程，为我的深度学习理论打下了坚实的基础。特别感谢孙悦淳学长在算力方面提供的大力支持，使得相关实验得以顺利进行。感谢李畅学长为我复现他的工作 QA-MDT 提供指导。感谢《明日方舟》的制作公司鹰角网络和“蓝毒”的配音演员种崎敦美提供训练数据与精神支持。最后，感谢所有在我学习和论文撰写过程中给予过帮助的老师、同学、家人和朋友们。