智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于分层联邦框架的音频模型生成技术研究

王健宗, 张旭龙, 姜桂林, 程宁, 肖京

引用本文:
王健宗, 张旭龙, 姜桂林, 程宁, 肖京. 基于分层联邦框架的音频模型生成技术研究[J]. 智能系统学报, 2024, 19(5): 1331-1339.
WANG Jianzong, ZHANG Xulong, JIANG Guilin, et al. Research on audio model generation technology based on a hierarchical federated framework[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(5): 1331-1339.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202306054

DOI: 10.11992/tis.202306054
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20240412.1559.002

基于分层联邦框架的音频模型生成技术研究

王健宗[1], 张旭龙[1], 姜桂林[2], 程宁[1], 肖京[1]
(1. 平安科技 (深圳) 有限公司, 广东 深圳 518046; 2. 湖南财信金融控股集团有限公司, 湖南 长沙 410035)

摘　要: 针对音频模型, 围绕下一代音频生成技术研究, 构建联邦音频模型训练框架, 面向超大规模音频数据进行音频表征学习, 为音频下游任务提供高效鲁棒的解决方案。提出一种适用于音频模型的联邦学习框架, 解决数据异构性、通信效率、隐私保护等问题; 提出一种基于对比学习的音频模型的预训练方法, 利用<音频, 文本描述>数据学习语义特征, 提高模型的泛化能力和多样化能力; 提出一种基于提示学习的音频生成微调方法, 利用少量标注数据提高模型的适应能力和定制化能力; 提出一种音频模型分布式优化算法进行模型压缩, 降低模型的复杂度和资源消耗, 提高模型的部署效率和运行效率。通过在下游任务音效转换上的实验, 提出的方法在语音质量平均意见得分可以达到 3.81。实验结果表明, 该方法在音效转换任务上取得了良好的效果。

关键词: 音频模型; 联邦学习框架; 音频表征学习; 数据异构性; 隐私保护; 对比学习; 提示学习; 模型压缩
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)05−1331−09

中文引用格式: 王健宗, 张旭龙, 姜桂林, 等. 基于分层联邦框架的音频模型生成技术研究 [J]. 智能系统学报, 2024, 19(5): 1331–1339.
英文引用格式: WANG Jianzong, ZHANG Xulong, JIANG Guilin, et al. Research on audio model generation technology based on a hierarchical federated framework[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1331–1339.

Research on audio model generation technology based on a hierarchical federated framework

WANG Jianzong[1], ZHANG Xulong[1], JIANG Guilin[2], CHENG Ning[1], XIAO Jing[1]
(1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518046, China; 2. Hunan Chasing Financial Holdings Co., Ltd., Changsha 410035, China)

Abstract: This study focuses on the development of next-generation audio generation techniques, specifically through the construction of a federated audio model training framework. The goal is to enable efficient and robust audio representation learning on data massive scale, providing high-performance solutions for various downstream audio tasks. The key scientific challenges addressed in this research and their corresponding methods include the following: 1) Proposing a federated learning framework suitable for audio models to address issues such as data heterogeneity, communication efficiency, and privacy protection. 2) Introducing a pretraining method based on contrastive learning, utilizing <audio, text description> data pairs to learn semantic features and enhance the model's generalization and diversification capabilities. 3) Presenting a fine-tuning method grounded in prompt learning, utilizing a small amount of annotated data to improve the model's adaptability and customization capabilities. 4) Developing a distributed optimization algorithm to compress audio models so as to reduce model complexity and resource consumption, thereby improving deployment and operational efficiency. Through experimental evaluation in the downstream task of sound effect conversion, the proposed method achieved a score of 3.81 in terms of mean opinion score. The experimental results show that the proposed method achieves good performance in sound effect conversion tasks.

Keywords: audio model; federated learning framework; audio representation learning; data heterogeneity; privacy protection; contrastive learning; prompt learning; model compression

近年来, 随着计算资源、算法技术和数据规模的不断提升, 自然语言处理 (natural language processing, NLP) 模型在自然语言处理领域取得了

收稿日期: 2023−06−30.　**网络出版日期:** 2024−04−16.
基金项目: 广东省重点领域研发计划"新一代人工智能"重大专项(2021B0101400003).
通信作者: 张旭龙. E-mail: zhangxulong@ieee.org.

令人瞩目的成就，不仅在多个基准测试中刷新了记录，而且在多种下游任务中展现了强大的迁移学习和泛化能力，例如机器翻译、问答、文本摘要、对话生成等[1]。NLP 模型可以定义为使用海量的文本数据进行预训练，具有超大规模参数和强大的自然语言理解和生成能力的深度学习模型。NLP 模型通过自监督学习从无标注或弱标注的文本中学习通用的语言知识，然后通过微调或少样本学习适应特定的任务需求。这种方法不仅降低了数据标注和模型设计的成本和难度，而且提高了模型的通用性和可扩展性。目前，已经出现了许多具有代表性的 NLP 模型，例如 BERT、GPT-3、T5、Switch Transformer 等[2]，它们的参数量从数亿到数万亿不等，涵盖了多种语言和领域。这些 NLP 模型在一定程度上实现了对自然语言的“泛化”理解和生成，为人工智能走向通用人工智能提供了可能性。

尽管在文本处理和生成方面取得了成功的应用，但在音频模态 (语音、音乐、声音和 talking head) 领域，复制 NLP 模型的成功经验是有限的[3]。音频模型是指利用大量的音频数据和深度学习技术，构建能够处理复杂的音频任务的模型。音频模型可以提高音频任务的性能和效率，例如语音识别、语音合成、声纹识别、情感分析等。这些任务在人工智能领域具有广泛的应用前景，例如智能助理、智能客服、智能安防等。音频模型可以促进音频领域的跨任务和跨领域的知识迁移和共享，例如利用预训练的音频模型进行下游任务的微调或者多任务学习，或者利用音频模型进行跨媒体的信息融合和表达，例如音视频同步、音画生成等。音频模型可以推动音频领域的理论和方法的创新和发展，例如探索更有效的音频数据表示和建模方法，设计更优化的音频模型结构和训练策略，解决更具挑战性的音频问题和场景等。音频模型的研究具有重要的理论意义和实践意义。从理论意义上看，对音频模型的研究可以揭示深度学习模型在处理音频数据时所遵循的原理和机制，探索模型参数量、数据规模、计算资源、任务性能之间的关系和边界，评估模型的可靠性、可解释性、可信赖性等方面的问题，为音频领域的基础理论研究提供新的视角和思路。从实践意义上看，音频模型的研究可以促进音频领域的技术创新和应用推广，为各行各业提供智能化、自动化、高效化的解决方案，为人类社会带来巨大的经济效益和社会价值。

随着科技的不断发展，音频技术已经成为人们生活中不可或缺的一部分，音频数据的应用越来越广泛，例如音乐、电影、游戏、电话会议等各种领域。因此，开发更先进的音频技术已成为当前音频领域的重要研究方向。音频合成技术是音频技术中的一个重要领域，它可以实现人工智能音频创作、音乐自动合成、语音合成等各种应用。因此，研究下一代音频生成技术对推进音频技术的发展具有重要意义。目前，音频模型的研究主要集中在小数据量、小模型的场景中，对于大规模音频数据的处理和分析，目前还存在很多挑战。针对这些挑战，提出了基于联邦学习的音频模型训练框架，利用联邦学习解决数据异构性、通信效率、隐私保护等问题，从而为下游音频任务提供高效鲁棒的解决方案。

旨在满足社会对智能语音交互和应用日益增长的需求，并推动国家人工智能战略和产业发展。通过采用最新的音频模型技术，并结合联邦学习和音频表征学习的方法，提出了一种创新的联邦音频模型训练框架，以解决音频数据异构性、隐私性、多样性等问题，从而提高音频模型的性能和适应性。该研究的重要性在于为音频下游任务提供通用的音频预训练模型，可通过迁移学习和提示学习等技术进行快速定制化，实现多种音频任务，如语音识别、语音合成、语音转换、语音情感分析等，以满足不同领域和场景的需求。此外，该研究的必要性在于音频数据的规模和质量不足以支撑高质量的音频模型训练，同时音频数据的隐私和安全问题也制约了数据的共享和利用。为此，需要一种能够充分利用分布式数据资源，同时保护数据隐私和安全的音频模型训练方法。相关研究结果表明联邦学习和音频表征学习在音频领域具有有效性，因此具有推广应用的潜力。

本文的贡献在于：1）通过研究音频模型，解决音频数据规模化和复杂性问题，为音频技术的发展提供基础性支持；2）通过联邦学习解决数据隐私保护等问题，可以推进联邦学习在音频领域的应用；3）提出基于对比学习和提示学习的音频模型训练方法，可以提高模型的泛化能力和适应能力，为音频任务提供更好的解决方案；4）通过分布式优化算法对音频模型进行压缩，可以降低模型的复杂度和资源消耗，提高模型的部署效率和运行效率。

1　相关工作

音频模型是近年来音频领域的热门研究方向，国外有许多知名的机构和团队从事相关的研

究工作, 如谷歌、微软、Meta 等。音频模型的主要目标是利用大量的音频数据进行预训练, 学习通用的音频表征, 然后通过微调或提示学习等技术, 实现多种音频任务, 如语音识别、语音合成、语音转换、语音情感分析等。音频模型的主要优势是可以提高模型的泛化能力和多样化能力, 同时可以提高模型的适应能力和定制化能力。音频模型的主要挑战是需要解决数据异构性、通信效率、隐私保护等问题, 以及如何进行模型压缩和优化, 降低模型的复杂度和资源消耗。自监督学习 (self-supervised learning, SSL) 已成为解决许多语音处理问题的主流方法, 尤其是在有大量未标记的语音数据的情况下。HuBERT[4] 使用掩码预测和掩码连续音频信号进行训练。受向量量化技术的启发, SoundStream[5] 提出了用于携带语义信息的高层表示的分层架构。这些模型中的大多数都在紧凑且离散的空间中构建离散单元, 可以使用自回归 Transformer 来建模, 然后将预测映射回原始信号空间。Hayashi 等 [6] 利用离散的 VQ-VAE 表示来通过自回归机器翻译构建语音合成模型。AudioLM[7] 和 MusicLM[8] 采用直接建模语言的方法, 通过训练低比特率音频令牌的自回归生成模型来实现, 无需任何转录来解决一致性和高质量合成之间的权衡, 将音频合成视为语言建模任务, 并利用离散表示空间中的粗细层次音频离散单元的层次结构。最近, Nguyen 等[9] 利用离散表示的成功, 并引入了第一个端到端生成式口语对话语言模型。

联邦学习是一种分布式的机器学习方法, 可以在不暴露数据隐私的情况下, 利用多个参与方的本地数据进行协同训练。联邦学习在音频领域的应用还处于初级阶段, 国外有一些研究工作探索了联邦学习在音频任务中的可能性和优势, 如联邦语音识别[10]、联邦语音合成[11] 等。联邦学习在音频领域的主要优势是可以充分利用分散在不同设备或机构的数据资源, 提高数据利用率和模型性能, 同时保护数据隐私和安全。联邦学习在音频领域的主要挑战包括数据不平衡、通信开销、激励机制、攻击防御等问题。

对比学习和提示学习是两种自监督学习方法, 可以从无标签或弱标签的数据中学习有意义和可区分的特征表示。对比学习通过构造正负样本对, 并最大化正样本对之间的相似度, 最小化负样本对之间的相似度, 来提高特征表示的质量和泛化性。提示学习通过构造一些简单或复杂的提示, 并根据提示来调整特征表示或预测结果、提高模型的适应能力和定制化能力。对比学习和提示学习在音频领域有一些应用和探索, 如对比音频表征学习[12]、提示音频分类[13]。对比学习和提示学习在音频领域的主要优势是可以减少对人工标注数据的依赖, 同时可以适应不同领域和场景的音频任务。对比学习和提示学习在音频领域的主要挑战是需要解决特征表示的可解释性、多样性、稳定性等问题, 以及如何结合多模态信息。

目前这些相关技术的产业化还处于探索和发展阶段, 尚未形成成熟的商业模式和应用场景。国外有一些企业和机构尝试将相关技术应用于实际的音频产品和服务中, 如谷歌的语音助手、微软的语音合成器、亚马逊的语音转换器等。其主要应用领域包括智能语音交互、智能语音内容生成、智能语音内容分析等。近期有一些突破性的研究工作和实验结果出现, 显示了相关技术的潜力和价值。例如, 谷歌发布了一种基于对比学习的大规模多模态预训练模型[14], 可以实现文本到音频、图像到音频、视频到音频等多种任务; 微软发布了一种基于提示学习的大规模语音合成器[15], 可以根据不同的提示生成不同风格和情感的语音。

深入研究大音频模型、联邦学习、对比学习、提示学习等核心技术, 解决现有的技术问题和挑战, 提高技术水平和性能是今后的研究趋势。同时, 探索更多的多模态信息融合方法, 利用更丰富和更完整的信息提高音频任务的效果和体验。此外, 推动相关技术的产业化应用, 开拓更多的商业模式和应用场景, 可以满足社会对于智能语音交互和应用的日益增长的需求。国内已有一些关于音频模型相关子任务[16-18]、联邦学习[19]、对比学习[20] 和提示学习[21] 等方面的研究工作, 在联邦音频模型训练框架、超大规模数据的音频表示学习以及音频下游任务等方面研究较少。在国内产业界, 目前都还是提供简单的单一音频服务, 例如语音识别、语音合成、语音转换等, 未涵盖语音模型直接服务于下游子任务。目前比较有代表性的工作包括, UniSpeech[22](一系列用于语音的大规模自监督学习模型)、VALL-E[15](一种基于提示学习的文本转语音合成语言建模方法)和 FedP-CL[23](一种用于多模态联邦学习的原型对比学习方法)等。最近的研究工作报告了一些在音频模型或联邦学习方面的进展或成就。然而, 它们在数据质量、模型复杂度、通信效率、隐私保护等方面仍然面临一些挑战或限制。

利用音频模型、联邦学习、对比学习和提示

学习的优势，实现音频下游任务的高效和鲁棒解决方案。需要研究的方向包括：1）探索更有效和高效的联邦学习方法或算法，用于音频模型；2）为音频表示学习设计更多样化和灵活的对比学习目标或策略；3）开发更适应和个性化的提示学习方法或技术，用于音频模型微调；4）应用更先进的模型压缩或优化方法或工具，用于音频模型的部署和运行。

2　本文方法

本文提出的方法分为 4 大模块，包括：1）适用于音频模型的联邦学习模块，解决数据异构性、通信效率、隐私保护等问题；2）基于对比学习的音频模型的预训练模块，利用无标签数据提高模型的泛化能力和多样化能力；3）基于提示学习的音频模型微调模块，利用少量标注数据提高模型的适应能力和定制化能力；4）音频模型分布式优化模块进行模型压缩，降低模型的复杂度和资源消耗，提高模型的部署效率和运行效率。

2.1　音频模型联邦学习框架

为了构建音频模型，需要大量的算力资源以及数据资源，对于个人以及小型企业来说都是无法完成的任务，但是通过构建联邦音频模型可以实现数据上的联邦以及算力上的联邦，同时可以保障数据的隐私安全性，也能够借助联邦学习框架更好地统筹音频模型的构建与训练。

首先，设定音频模型的训练数据多元组，可以兼容不同任务数据集的输入格式，并可以在此多元组的基础上扩充到更多现有数据集。联邦学习的长期目标需要严格的隐私保证和低通信开销，同时保持相对较高的模型精度。同时实现所有这些目标极为具有挑战性。采用分层联邦学习的框架来解决这个问题。考虑到训练数据的统计异质性可能导致模型性能下降，该研究设计了一种运行时分布重构策略，该策略重新分配客户端并利用中介者重新安排客户端的本地训练。此外，本文设计了一种压缩校正机制，以降低通信开销而不牺牲模型性能。为进一步提供隐私保证，本文在进行本地训练时引入了差分隐私，该方法仅向完整模型的部分注入适量的噪声。分层联邦框架结构概览如图 1 所示，在模型性能、通信开销和隐私要求之间找到一个良好的平衡。

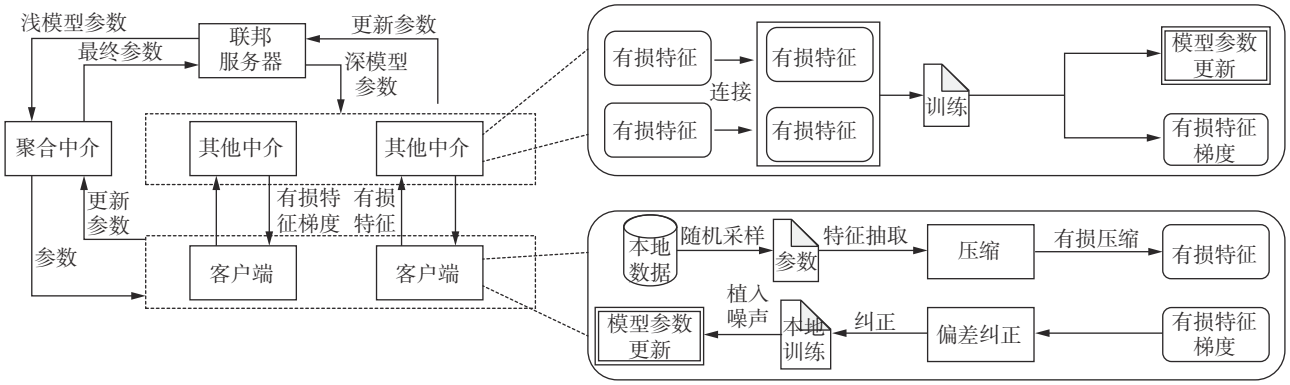

图 1　分层联邦架构概览

Fig. 1　Overview of hierarchical federated architecture

本文首先假设分层联邦中的所有组件（联邦服务器、中介、客户端）具有以下能力：1）所有组件能够遵循设计的协议规则，但都希望获取不同组件的本地数据；2）所有组件拥有任意的辅助信息来帮助推断特定客户端的私有信息，以协作构建共享模型；3）所有组件不会在训练过程中向客户端提供任何额外的信息。

在初始化阶段，联邦服务器首先将完整模型分成两个组件（浅层模型和深层模型），然后将前者分配给聚合中介，后者分配给其他中介。聚合中介将浅层模型分发给所有客户端。同时，联邦服务器初始化全局超参数。在实践中进行本地采样时，将本地数据随机排列并将其分成适当大小的小批次以提高效率。

在进行本地训练时，每个客户端都使用浅层模型提取特征，这些特征将被有损压缩器压缩并发送到相应的中介。之后，每个中介通过连接器（如图 1 所示）连接接收到的特征以获取合成特征。这个过程可以看作是从一个虚拟重建分布中采样，然后使用浅层模型进行前向传播。

2.2　基于对比学习的音频模型的预训练方法

音频是世界上最常见的信息类型之一，与文本和图像数据并列。然而，不同的音频任务通常需要精细注释的数据，这限制了可用的音频数据量，因为收集过程需要耗费大量人力。因此，设计一种有效的音频表示方式，适用于许多音频任

务, 而不需要大量的监督, 仍然是一个挑战。针对音频模型采用一种对比学习的音频−文本预训练流程, 以将自然语言描述与音频数据相结合, 学习音频表示。为了实现这个目标, 首先需要对音频−文本对进行大规模收集, 可以是任何声音配上对应的文字描述。其次, 通过构建对比语音−文本预训练模型, 考虑不同的音频编码器和文本编码器。该研究将特征融合机制和关键字到文本描述增强纳入模型设计中, 进一步使模型能够处理可变长度的音频输入并提高性能。如图 2 所示, 是基于对比学习的音频−文本预训练模型框架概览。

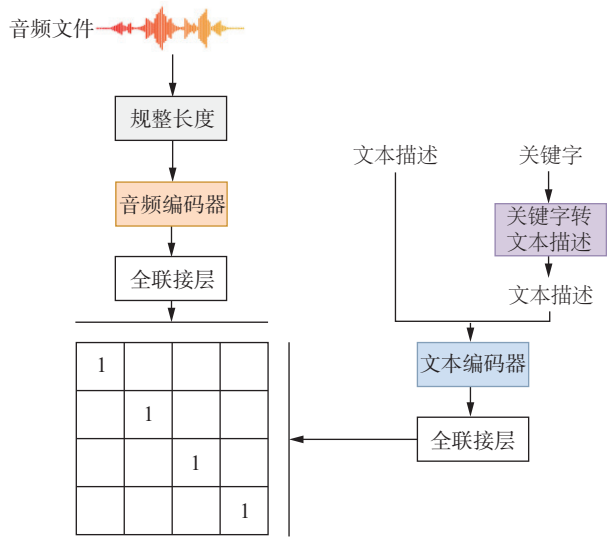

图 2　基于对比学习的音频−文本预训练模型框架

Fig. 2　Framework of contrastive learning-based audio-text pretraining model

图 2 给出了对比学习音频文本编码器模型的总体架构。本文有两个编码器分别处理音频数据和文本数据的输入, 其中文本和音频构成成对输入数据, 适用于有文本标注的音频数据集。首先, 针对文本端的文本描述和关键字, 将关键字转成文本描述, 进而跟文本描述进行连接, 一起送入文本编码器。音频端也是类似的操作, 输入的音频首先进行长度的规整, 例如裁剪或延伸等。接着送入音频编码器进行特征提取。为了将文本和音频进行特征空间中的长度对齐, 在音频编码器和文本编码器后面均加上一个全连接层, 这是由多层全联接层构成的感知机, 采用 ReLU 作为激活函数, 将编码器输出映射到相同的维度。此时获得了维度对齐之后的音频编码特征向量和文本特征编码向量。

在模型训练部分, 将成对的音频和文本嵌入进行对比学习, 该部分的 Loss 为

$$L=\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(E_i^a\cdot E_i^t/\tau)}{\sum_{j=1}^{N}\exp\left(E_i^a\cdot E_j^t/\tau\right)}+\log\frac{\exp(E_i^t\cdot E_i^a/\tau)}{\sum_{j=1}^{N}\exp\left(E_i^t\cdot E_j^a/\tau\right)}\right)\tag{1}$$

式中: E_i^a 代表所提取的音频特征; E_i^t 代表提取的文本特征; N 为文本与音频对的数量; τ 为温度参数, 用于控制对比学习中 softmax 函数的平滑程度。通过对比学习, 提出的模型可以感知到成对音频和文本之间的对应关系, 并且区分不匹配的音频和文本对之间的差异, 从而匹配音频文本关系。同时在训练模型之后, 学到的音频特征向量和文本特征向量表示可以用作不同的下游子任务。

2.3　基于提示学习的音频生成方法

大规模多模态生成建模在文本到图像生成方面取得了里程碑式的进展。本研究提出使用提示学习结合扩散模型的方法来进行音频合成微调, 这里音频合成是指对输入音频按照文本指令进行合成操作 (例如文本到语音, 或者对音频内容进行处理去除背景伴奏、增加音频特效等), 以获取按照文本指令处理后的音频, 其具有各种应用场景, 例如添加背景音效、替换背景音乐、修复不完整的音频等。遵循人类指令的音频合成工具可以更灵活地进行处理, 符合人类的期望。图 3 给出了本文提出的基于提示学习的音频生成模型的结构。

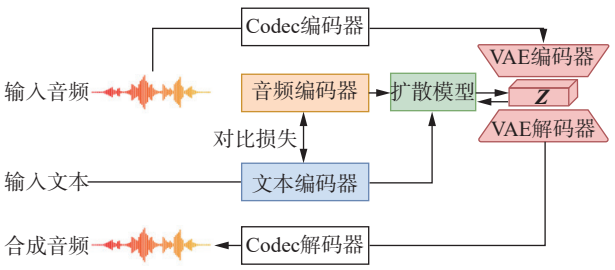

图 3　基于提示学习的音频生成模型结构

Fig. 3　Framework of prompt-based learning for audio generation fine-tuning model

图 3 由 6 个关键部分组成, 1) 文本和音频输入: 图 3 的左侧给出了两个输入流, 分别为文本指令和原始音频。文本指令用于指导音频的生成过程, 而原始音频则是合成过程的基础材料。2) 文本和音频编码器: 这两个编码器分别处理文本指令和音频输入。文本编码器将文本指令转换为特征向量, 而音频编码器则将原始音频转换为相应

的特征向量。这些特征向量是后续合成步骤的关键输入。3)特征融合:此部分给出了文本和音频特征向量的融合过程。这一过程确保了文本指令和原始音频特征的有效结合,为生成高质量的目标音频奠定了基础。4)扩散模型:图 3 中心位置的扩散模型是音频生成的核心。它接收融合后的特征向量,通过一系列复杂的运算过程,生成符合文本指令的目标音频。5)特征提取器:图 3 的右侧展示了特征提取器,它使用残差矢量量化(residual vector quantization, RVQ)和变分自编码器(variational autoencoder, VAE)技术来处理生成的音频。这一步骤进一步提高了音频质量,确保生成的音频与原始输入和文本指令高度一致。6)输出音频:图 3 的最右侧是最终的输出音频,它是整个模型处理过程的产物,给出了文本指令和原始音频输入融合、处理后的结果。

该方法的创新在于扩散模型的应用和特征提取器的创新设计。这种方法允许模型在生成过程中更细致地控制音频质量,同时保持对文本指令的高度敏感性。同时,通过使用 RVQ 和 VAE,提出的模型能够在生成过程中保持更高的音频质量和更低的数据损失。

预训练的音频文本对比模型能够在采样时提供音频特征向量和文本特征向量作为条件,从而训练具有音频特征向量的扩散模型。通过学习音频信号及其组成的潜在表示,而不对跨模态关系进行建模,在生成质量和计算效率方面都具有优势。根据输入文本描述生成音频样本。通过概率生成模型(扩散模型)估计真实条件数据分布。与传统的 STFT(short time Fourier transform)相关的频谱编码不同,本研究使用音频编解码器的 Codec 作为中间表示,利用大量且多样化的数据,使模型具有很强的上下文学习能力。其中,采用 RVQ 进行特征提取:

$$\left\{\boldsymbol{e}_j^i\right\}_{j=1}^{R}=f_{\mathrm{rvq}}(\boldsymbol{h}^i),\quad \boldsymbol{Z}^i=\sum_{j=1}^{R}\boldsymbol{e}_j^i,\quad \boldsymbol{Z}=\{\boldsymbol{Z}^i\}_{i=1}^{n}$$

式中:$\boldsymbol{e}_j^i$为第 i 个音频特征向量 $\boldsymbol{h}_i$ 经过 RVQ 编码后得到的第 j 个残差向量,R 为残差向量的数量,f_{rvq} 为 RVQ 编码函数,$\boldsymbol{Z}_i$ 为第 i 个音频特征向量 $\boldsymbol{h}_i$ 经过 RVQ 编码后得到的量化向量,n 为量化向量的个数。通过 RVQ 编码,可以将高维的音频特征向量压缩成低维的量化向量序列,从而降低模型的复杂度和计算成本。

此外,使用 VAE 将编码压缩到一个小的潜空间 $\boldsymbol{Z}$ 中。其中,VAE 由一个编码器和一个解码器组成,具体表示公式为

$$\begin{aligned}\mathcal{L}(\theta,\phi;x^{(i)})=&-D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{Z}|x^{(i)})|p_\theta(\boldsymbol{Z})\right)+\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|x^{(i)})}\left[\log p_\theta(x^{(i)}|\boldsymbol{Z})\right]\end{aligned}\tag{2}$$

式中:θ 和 ϕ 分别为解码器和编码器的参数,$x^{(i)}$ 为第 i 个输入数据样本,D_{KL} 为 Kullback-Leibler 散度,$q_\phi(\boldsymbol{Z}|x^{(i)})$ 为编码器在给定 $x^{(i)}$ 的条件下对潜在变量 $\boldsymbol{Z}$ 的近似后验分布,$p_\theta(\boldsymbol{Z})$ 为 $\boldsymbol{Z}$ 的先验分布,E 为求期望。

在训练目标中,采用重构损失、对抗损失和高斯约束损失。

2.4　音频模型分布式优化算法

联邦学习中分布式优化算法是一种在多个设备上协同训练一个全局模型的技术,可以在保护用户数据隐私的同时,利用多方的数据和算力。随机梯度下降(stochastic gradient descent, SGD)法是一种常用的分布式优化算法,它允许每个设备在本地进行多次梯度更新,然后与中心服务器进行模型平均,从而减少通信开销和提高训练效率。

考虑连续时间 SDE:

$$\mathrm{d}X_t=b(X_t)\mathrm{d}t+\sqrt{\eta\sigma(X_t)\mathrm{d}W_t}$$

式中:X_t 为模型参数在时刻 t 的取值,$b(X_t)$ 为模型参数在时刻 t 的漂移项,η 为学习率,$\sigma(X_t)$ 为模型参数在时刻 t 的扩散项。

解为(近似 SGD)

$$X_{k+1}=X_k+\eta b(X_k)+\eta\sigma(X_k)Z_k$$

式中:$b(X_k)$ 模型参数在第 k 次迭代时的梯度,$\sigma(X_k)$ 模型参数在第 k 次迭代时的随机扰动,Z_k 为标准正态分布的随机变量。

同时,考虑到概率空间与测度论,可以进一步从理论层面证明与分析优化算法本身的可行性,将联邦学习中的常见优化目标转化为

$$f(x):=\mathrm{E}f_\gamma(x)=\int_{\Omega}f_{\gamma(\omega)}(x)\mathrm{d}P(\omega)$$

式中:E 为求期望,Q 为样本空间,ω 为 Q 中的一个样本点,$f_{\gamma(\omega)}$ 为样本点 ω 出模型参数为 x 时的损失函数,$P(\omega)$ 为表示样本点 ω 的概率分布。

这里,利用实分析从 P 概率空间以及勒贝格积分对优化目标进行了重构,有利用进行数学层面的分析。

利用随机微分方程(SDE),对改进后的优化算法进行重构,得到一个 SDE 的形式。这里给出 Local SGD 的 slow-SDE:

$$d_{\zeta(t)}=\mathrm{P}_\zeta\left(\sqrt{\kappa_1}\sqrt{\boldsymbol{\Sigma}_1(\zeta)}\mathrm{d}W_t-\kappa_2\nabla^3L(\zeta)[\boldsymbol{\Sigma}_2(\zeta)]\mathrm{d}t\right.$$

式中:$\zeta(t)$为模型参数在 t 时刻的取值;P_ζ为投影操作,将参数的变化投影到一个特定的空间中;κ_1 为控制随机扰动强度的参数;$\boldsymbol{\Sigma}_1(\zeta)$为参数$\zeta$在

第一个客户端上的协方差矩阵, 用于衡量参数在不同维度上的变化程度; W_t 为 t 时刻参数权重; κ_2 为控制梯度下降强度的参数; $L(\zeta)$为损失函数; ∇^3为求三阶导数; Σ_2为参数(ζ)在第二个客户端上的协方差矩阵。

在该部分中, 考虑 Nesterov 加速:

$$V_t = \gamma V_{t-1} + \eta \nabla(\theta - \gamma V_{t-1})$$

$$\theta = \theta - V_t$$

式中: V_t 为在第 t 次迭代时参数更新的动量, γ 为动量系数。

在分布式技术的研究上, 将依据如 Multi-Level Local SGD 等方法, 构造适用于音频模型的技术方法, 其中涉及到不同分布形式的中心群、中心、客户端等, 而他们之间的通信就是联邦学习的主要通信成本。因此, 设计一个合适的通信部署是尤为重要的, 可以将不同单位之间的通信的更新规则表示为

$$X_{k+1} = (X_k - \eta G_k) T_k$$

式中: G_k 为节点在第 k 次迭代时计算的梯度, T_k 为节点之间通信的拓扑结构。

只需要改变T_k不同节点之间的定义, 就可以实现整个网络的通信。

3　实验和结果

在实验中, 验证了联邦音频模型在下游子任务音频音效转换的对比。

3.1　数据集

在公共数据集 LJSpeech[24] 中使用了 4 种环境效果, 包括浴室、洞穴、教室和画廊, 以生成实验数据集。所选环境的房间冲击响应与 LJSpeech 中的原始语音进行卷积处理。LJSpeech 数据集共有 13 100 个片段, 总时长约为 24 h。实验中对其进行预处理, 使环境效果与音频波形进行卷积处理以生成模拟环境音频。预处理后, 得到了 4 种环境效果, 并将数据集扩大到 LJSpeech 的 5 倍。

3.2　实验设置

由于音频音效转换主要是将源语音的环境效应转换为目标语音的环境效应, 将声音转换模型作为基线来比较执行环境效应切换任务的表现。AutoVC [25]、CycleGAN-VC3 [26]、SpeechSplit[27] 和 NaturalSpeech 2 [28] 的基线模型均在 LJSpeech 数据集上进行环境效应转换的重新训练。

首先进行了梅尔频谱、音调和能量的预处理, 使用 pyworld 进行音调估计。所有音频数据均被重新采样为 22.05 kHz, 使用 1 024 窗口大小和 256 跳步大小进行 STFT, 梅尔通道设置为 80。需要注意的是, 选择将梅尔频谱的最大长度设置为 1 200 进行填充。

关于实验中使用的模型配置, 将编码器改为使用卷积 1D 层以接受梅尔频谱输入, 内核大小为 5, 步长为 1, 膨胀率为 1, 填充为 2。环境效应提取器采用 U-net 架构卷积层, 包含 4 个下采样层和 4 个上采样层。下采样层使用内核大小为 2 的 1D 最大池化, 堆叠两个内核大小为 3, 填充为 1 的 1D 卷积层。上采样层使用内核大小为 2, 步长为 2 的转置 1D 卷积层, 以及与下采样层相同的两个 1D 卷积层堆叠。梯度反转层在反向函数内实现以进行负处理。环境效应分类器均使用内核大小为 3, 填充为 1 的两个 1D 卷积层堆叠, 连接一个线性层。

训练主要使用一张 Tesla V100 GPU 进行, 批次大小为 16, 总训练步数为 9×10^5, 每 1×10^4 步保存一次训练模型。使用 Adam 优化器, 将 β_1、β_2、ε 分别设置为 0.9、0.98 和 10^{-9}。

3.3　客观实验评价

表 1 给出了生成语音与特定环境下的基准真实语音之间的 Mel 倒谱失真(MCD)值。MCD 通常用于语音转换任务中, 它测量基准真实语音和合成语音之间的全局结构差异。从 MCD 的比较结果来看, 提出的模型在浴室、洞穴、教室和画廊等环境下优于基线方法。而对于环境转换, 本文提出的方法略高于 SpeechSplit。这些结果可能意味着所有系统均实现了可比较的性能水平。

表 1　不同模型的 MCD 比较

Table 1　Comparison of MCD of different models

方法	浴室	洞穴	教室	画廊
AutoVC[25]	9.56	9.32	9.31	9.29
CycleGAN-VC3 [26]	9.32	9.02	8.91	9.14
SpeechSplit [27]	8.62	8.47	8.40	8.28
NaturalSpeech 2 [28]	8.78	8.56	8.63	8.31
本文方法	8.50	8.33	8.34	8.16

3.4　主观实验评价

实验进行了听觉测试以评估转换后的语音的感知质量。在主观评价中, 邀请了 10 名听众评估结果。每名听众都被要求对语音质量进行平均意见分数评分, 并进行 ABX 测试以评估目标语音的环境相似度。对于 MOS 测试, 随机选择了每个环境的 5 个长于 2 s 且短于 6 s 的语音, 并混洗所有音频样本, 让测试者为每个语音在 1~5 的范围内给出一个评分。表 2 给出了 MOS 测试的结

果。通过比较方法和本文方法都以无成对数据的方式进行无监督重建。

表 2　不同模型的 MOS 比较

Table 2　Comparison of MOS of different models

方法	浴室	洞穴	教室	画廊
Ground truth	4.43	4.52	4.60	4.58
AutoVC [25]	2.86	3.08	3.12	3.20
CycleGAN-VC3 [26]	3.19	3.29	3.42	3.24
SpeechSplit [27]	3.58	3.63	3.76	3.79
NaturalSpeech 2 [28]	3.55	3.61	3.71	3.68
本文方法	3.61	3.67	3.75	3.81

表 2 的 MOS 测试结果可以看出，所提出的模型在浴室、洞穴、教室和画廊等环境下优于基线方法。提出的方法优于所有基线方法，但是与地面真实语音存在较大的差距，提出的方法的平均得分约为 3.7 的情况下。一方面，这表明了所提出的方法的验证，并与声音转换的相关工作相当。此外，不同环境条件下的转换结果可能学习到特定环境的混响。基线方法在环境目标下表现更差，而声音转换没有环境效应转换的目标。

在 ABX 测试中，A 和 B 分别是由基线和本文方法转换的语音，而 X 是目标环境的语音。为每个环境语音选择了 3 个句子，每对音频包含所提出的方法的转换语音和比较方法的音频，这些音频在音频对中交替出现。在计算结果时，计算测试中发生的次数的平均值。为了消除刺激顺序中的偏差，A 和 B 是随机选择所提出的方法和比较基线方法的音频标签。对于每对句子，听众被要求选择哪一个更类似于目标语音 X 的环境。听众可以选择 3 个选项，即 A、B 和相当。表 3 给出了环境效果的优先得分。

表 3　相似度偏好得分比较

Table 3　Comparison of preference score on similarity

方法	浴室	洞穴	教室	画廊
AutoVC [25]	0.06	0.08	0.10	0.06
CycleGAN-VC3 [26]	0.19	0.20	0.18	0.21
SpeechSplit [27]	0.25	0.22	0.25	0.24
NaturalSpeech 2 [28]	0.22	0.20	0.21	0.22
本文方法	0.28	0.30	0.26	0.27

如表 3 所示，特定环境效应下的 ABX 测试结果表明，本文方法表现与基线方法相当。在洞穴环境下，有 30% 的选择会落在本文方法的音频样本中。而在教室环境下，相似性略高于基线方法。这个结果与 MCD 客观评估和 MOS 评估一致。

4　结束语

本文围绕下一代音频合成技术研究，面向超大规模音频数据进行音频表征，探索用于音频生成以及音频处理的新范式，提高音频生成的质量以及鲁棒性，对于大规模音频数据语料以及标注通过构建联邦音频模型训练框架达到隐私保护的同时协作共建音频模型。通过制作模拟环境效应数据集进行语音转换实验。结果表明，本文方法能够有效地进行环境效应的转换，并在 MOS、MCD 和环境相似性等方面优于语音转换任务中的基线方法。

参考文献：

[1] LIU Pengfei, YUAN Weizhe, FU Jinlan, et al. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1−35.

[2] TRUMMER I. From BERT to GPT-3 codex[J]. Proceedings of the VLDB endowment, 2022, 15(12): 3770−3773.

[3] GHOSAL D, MAJUMDER N, MEHRISH A, et al. Text-to-audio generation using instruction-tuned LLM and latent diffusion model[EB/OL]. (2023−04−24)[2023−06−30]. http://arxiv.org/abs/2304.13731v2.

[4] 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.

[5] ZEGHIDOUR N, LUEBS A, OMRAN A, et al. SoundStream: An end-to-end neural audio codec[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 30: 495−507.

[6] HAYASHI T, WATANABE S. DiscreTalk: text-to-speech as a machine translation problem[EB/OL]. (2020−05−12)[2023−06−30]. http://arxiv.org/abs/2005.05525v1.

[7] BORSOS Z, MARINIER R, VINCENT D, et al. Audiolm: a language modeling approach to audio generation [EB/OL]. (2022−09−07)[2023−06−30]. https://arxiv.org/abs/2209.03143.

[8] AGOSTINELLI A, DENK T I, BORSOS Z, et al. Musiclm: generating music from text. [EB/OL]. (2023−01−26)[2023−06−30]. https://arxiv.org/abs/2301.11325.

[9] NGUYEN T A, KHARITONOV E, COPET J, et al. Generative spoken dialogue language modeling[J]. Transactions of the association for computational linguistics, 2023, 11: 250−266.

[10] CUI Xiaodong, LU Songtao, KINGSBURY B. Federated acoustic modeling for automatic speech recognition[C]// ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 6748−6752.

[11] HONG Zhenhou, WANG Jianzong, QU Xiaoyang, et al. Federated learning with dynamic transformer for text to speech[C]//Interspeech 2021. Brno: ISCA, 2021: 3590−3594.

[12] WU Yusong, CHEN Ke, ZHANG Tianyu, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation[EB/OL]. (2022−11−12)[2023−06−30]. http://arxiv.org/abs/2211.

06687v4.
[13] WU Shangda, YU Dingyao, TAN Xu, et al. CLaMP: contrastive language-music pre-training for cross-modal symbolic music information retrieval[EB/OL]. (2023−04−21)[2023−06−30]. http://arxiv.org/abs/2304.11029v4.
[14] WU Junru, LIANG Yi, HAN Feng, et al. Scaling multimodal pre-training via cross-modality gradient harmonization[J]. Advances in neural information processing systems, 2022, 35: 36161−36173.
[15] WANG Chengyi, CHEN Sanyuan, WU Yu, et al. Neural codec language models are zero-shot text to speech synthesizers[EB/OL]. (2023−01−05)[2023−06−30]. http://arxiv.org/abs/2301.02111v1.
[16] 谢旭康, 陈戈, 孙俊, 等. TCN-Transformer-CTC 的端到端语音识别 [J]. 计算机应用研究, 2022, 39(3): 699−703.
XIE Xukang, CHEN Ge, SUN Jun, et al. TCN-Transformer-CTC for end-to-end speech recognition[J]. Application research of computers, 2022, 39(3): 699−703.
[17] 解元, 邹涛, 孙为军, 等. 面向高混响环境的欠定卷积盲源分离算法 [J]. 通信学报, 2023, 44(2): 82−93.
XIE Yuan, ZOU Tao, SUN Weijun, et al. Algorithm of underdetermined convolutive blind source separation for high reverberation environment[J]. Journal on communications, 2023, 44(2): 82−93.
[18] 方昕, 黄泽鑫, 张聿晗, 等. 基于时域波形的半监督端到端虚假语音检测方法 [J]. 计算机应用, 2023, 43(1): 227−231.
FANG Xin, HUANG Zexin, ZHANG Yuhan, et al. Semi-supervised end-to-end fake speech detection method based on time-domain waveforms[J]. Journal of computer applications, 2023, 43(1): 227−231.
[19] 钟佳淋, 吴亚辉, 邓苏, 等. 基于改进 NSGA-Ⅲ的多目标联邦学习进化算法 [J]. 计算机科学, 2023, 50(4): 333−342.
ZHONG Jialin, WU Yahui, DENG Su, et al. Multi-objective federated learning evolutionary algorithm based on improved NSGA-Ⅲ[J]. Computer science, 2023, 50(4): 333−342.
[20] 陈洋, 廖灿辉, 张锟, 等. 基于自监督对比学习的信号调制识别算法 [J]. 系统工程与电子技术, 2023, 45(4): 1200−1206.
CHEN Yang, LIAO Canhui, ZHANG Kun, et al. A signal modulation indentification algorithm based on self-supervised contrast learning[J]. Systems engineering and electronics, 2023, 45(4): 1200−1206.
[21] 罗贤昌, 薛吟兴. 基于 BERT 的提示学习实现软件需求精确分类 [J]. 信息技术与网络安全, 2022, 41(2): 39−45.
LUO Xianchang, XUE Yinxing. Accurately classify software requirements using prompt learning on BERT[J]. Information technology and network security, 2022, 41(2): 39−45.
[22] WANG Chengyi, WU Yu, QIAN Yao, et al. UniSpeech: unified speech representation learning with labeled and unlabeled data[EB/OL]. (2021−01−19)[2023−06−30]. http://arxiv.org/abs/2101.07597v2.
[23] TAN Yue, LONG Guodong, MA Jie, et al. Federated learning from pre-trained models: a contrastive learning approach[EB/OL]. (2022−09−21)[2023−06−30]. http://arxiv.org/abs/2209.10083v1.
[24] KEITH I. The lj speech dataset[EB/OL]. [2023−06−30]. https://keithito.com/LJ- Speech-Dataset/.
[25] QIAN Kaizhi, ZHANG Yang, CHANG Shiyu, et al. Autovc: Zero-shot voice style transfer with only autoencoder loss[C]//36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 5210–5219.
[26] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion[EB/OL]. (2020−10−22)[2023−06−30]. http://arxiv.org/abs/2010.11672v1.
[27] QIAN Kaizhi, ZHANG Yang, CHANG Shiyu, et al. Unsupervised speech decomposition via triple information bottleneck[C]//Proceedings of the 37th International Conference on Machine Learning. Virtual: ACM, 2020: 7836−7846.
[28] SHEN Kai, JU Zeqian, TAN Xu, et al. NaturalSpeech 2: latent diffusion models are natural and zero-shot speech and singing synthesizers[EB/OL]. (2023−04−18)[2023−06−30]. http://arxiv.org/abs/2304.09116v3.

作者简介:

王健宗, 博士, 平安科技(深圳)有限公司副总工程师, 资深人工智能总监, 联邦学习技术部总经理, 智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后, 美国莱斯大学和华中科技大学联合培养博士, 中国计算机学会资深会员, 中国计算机学会大数据专家委员会委员, 中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习。E-mail: jzwang@188.com。

张旭龙, 博士, 平安科技(深圳)有限公司高级算法研究员, 担任清华大学深圳研究院以及中国科学技术大学先进技术研究院校外导师, 目前是 IEEE、中国自动化学会以及中国计算机学会会员, 担任联邦数据与联邦智能专委会委员, 主要研究方向为语音合成、语音转换、音频驱动虚拟人生成、音乐信息检索以及机器学习和深度学习方法在人工智能领域应用。2023 年入选上海市东方英才计划青年项目。E-mail: zhangxulong@ieee.org。

肖京, 博士, 国家特聘专家, 国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员, 兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长, 清华大学、上海交通大学、同济大学等客座教授。先后在爱普生美国研究院及美国微软公司担任高级研发管理职务, 现任平安集团首席科学家, 负责人工智能技术研发及在金融、医疗、智慧城市等领域的应用, 带领团队树立了多项传统行业智能化经营的标杆。主要研究方向为人工智能与大数据分析挖掘, 参与及承担国家级项目 8 项, 获美国授权专利 101 项, 中国发明专利 155 项。先后获 2018 年中国专利奖、2019 年吴文俊人工智能杰出贡献奖、2020 年吴文俊人工智能科技进步一等奖、2020 年上海市科技进步奖一等奖、2020 年中国人工智能十大风云人物、2021 年深圳市五一劳动奖章、2022 年深圳市最美科技工作者等荣誉。发表学术论文 249 篇。

[责任编辑: 刘冰洁]