计算机科学
COMPUTER SCIENCE

基于XTTS模型的声音克隆系统研究

王陈偲, 杨思燕, 苗启广

引用本文

王陈偲, 杨思燕, 苗启广. 基于XTTS模型的声音克隆系统研究[J]. 计算机科学, 2026, 53(5): 59-67.

WANG Chencai, YANG Siyan, MIAO Qiguang. Research on Voice Cloning System Based on XTTS Model [J]. Computer Science, 2026, 53(5): 59-67.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600187

基于 XTTS 模型的声音克隆系统研究

王陈偲[1]　杨思燕[2]　苗启广[1,3]
1 西安电子科技大学计算机科学与技术学院　西安 710071
2 陕西开放大学信息化处　西安 710119
3 西安市大数据与视觉智能重点实验室　西安 710071
(25031212025@stu.xidian.edu.cn)

摘　要　随着深度学习和语音合成技术的不断发展,语音克隆在智能语音助手、虚拟主播和无障碍通信等领域展现出广阔的应用前景。然而,现有语音克隆系统在音色相似度、交互便捷性和大规模数据处理能力等方面仍存在不足,难以满足用户对高质量、个性化语音合成的实际需求。为此,基于 XTTS 模型设计并实现了一个支持多语种语音克隆与批量文本转语音的 Web 平台,针对语言覆盖数量有限、低资源条件下音色迁移受限以及批量处理效率低的问题进行了改进。系统采用前后端分离架构,后端基于 Flask 搭建 API 接口,前端结合主流 Web 技术与 AJAX 实现异步交互,数据库采用 MySQL 管理用户与音频数据。平台集成语音克隆、文本转语音与批量处理等功能模块,具备良好的灵活性与扩展性。测试结果表明,该系统在语音自然度与音色相似度方面表现良好,具有较高的应用价值与推广潜力。

关键词:语音克隆;文本转语音;XTTS;FreeVC;Flask
中图分类号　TP311

Research on Voice Cloning System Based on XTTS Model

WANG Chencai[1], YANG Siyan[2] and MIAO Qiguang[1,3]
1 School of Computer Science and Technology, Xidian University, Xi'an 710071, China
2 Department of Information Technology, The Open University of Shaanxi, Xi'an 710119, China
3 Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an 710071, China

Abstract　With the continuous advancement of deep learning and speech synthesis technologies, voice cloning has shown broad application prospects in intelligent voice assistants, virtual anchors, and barrier-free communication. However, existing voice cloning systems still face challenges in timbre similarity, interactive efficiency, and large-scale processing capability, making it difficult to meet the growing demand for high-quality, personalized speech synthesis. To address these limitations, this paper designs and implements a Web-based platform for multilingual voice cloning and batch text-to-speech synthesis, based on the XTTS model. The system improves upon existing solutions by enhancing language coverage, reducing data dependency for timbre transfer, and optimizing batch processing efficiency. It adopts a front-end/back-end decoupled architecture, with a Flask-based RESTful API at the back end and mainstream Web technologies combined with AJAX at the front end. MySQL is used for managing user and audio data. The platform integrates voice cloning, text-to-speech, and batch synthesis modules, and demonstrates strong flexibility and scalability. Experimental results show that the system performs well in speech naturalness and timbre similarity, proving its practical value and application potential.

Keywords　Voice cloning, Text-to-speech, XTTS, FreeVC, Flask

到稿日期:2025-06-26　返修日期:2025-07-20
基金项目:陕西工商职业学院重点课题(20GA06);广西可信软件重点实验室课题(KX202047);陕西省重点研发计划(2024GH-ZDXM-47);陕西高等教育教学改革研究项目(23JG003);中国高等教育学会高等教育科学研究规划课题(24PG0101)
This work was supported by the Key Project of Shaanxi Polytechnic Institute Research Program (20GA06), Guangxi Key Laboratory of Trusted Software Project (KX202047), Key Research and Development Program of Shaanxi Province (2024GH-ZDXM-47), Higher Education Teaching Reform Research Program of Shaanxi Province (23JG003) and Research Project of the China Association of Higher Education (24PG0101).
通信作者:苗启广(qgmiao@xidian.edu.cn)

1 引言

随着深度学习、自然语言处理与语音合成技术的不断发展，语音合成已被广泛应用于智能语音助手、虚拟主播和无障碍通信等领域。语音克隆作为语音合成领域的重要分支，能够通过少量目标说话人的语音样本实现个性化的语音生成，推动语音交互向智能化方向发展。目前，主流语音克隆方法多采用端到端深度神经网络建模音色特征，以提高语音的自然度与音色一致性。

近年提出的 Tacotron[1-2]，DeepVoice[3-5]和 X-vector[6]等模型显著提升了语音合成的自然度与多说话人建模能力，并在跨语言合成方向取得了一定进展。然而，现有语音克隆系统仍面临诸多挑战，包括情感表达能力有限、跨语言音色迁移效果不理想以及模型推理时延较高等，这些问题制约了其在实际场景中的广泛应用。此外，系统部署过程复杂、交互方式单一以及缺乏对大规模任务的处理能力等因素，也在一定程度上影响了其在普通用户中的可用性。

为提升语音克隆系统的自然度与生成质量，研究者们从声码器结构、风格建模与跨语言迁移等方向提出了多种改进方法。在声码器优化方面，VocGAN[7]在 MelGAN 的基础上引入多尺度判别器，兼顾语音质量与生成速度；HiFi-GAN[8]通过将全卷积生成器与多周期判别器结合，在保证高生成效率的同时，追平了 WaveNet 的生成质量；UnivNet[9]则采用多分辨率判别器结构来增强合成波形的频谱结构，提升了语音生成的速度；CARGAN[10]通过引入部分自回归模块，弥补了周期性失真带来的音质缺陷，改善了音高一致性与听感质量。

在说话人建模与风格迁移方面，VALL-E[11]基于神经编解码器架构，能在提供文本和短时参考音频(3s)的条件下合成高质量的目标说话人语音。U-Style[12]模型通过级联 U-Net 结构对说话人与风格特征进行分离建模，提升了零样本语音克隆的表现力。OpenVoice[13]系统则支持多语言语音克隆，并在语速、情绪和口音等维度实现了灵活的风格控制。

在零样本语音克隆与跨语言迁移方面，相关研究聚焦于提升系统的泛化能力。MaskGCT[14]基于掩蔽生成变换器架构，实现在无监督条件下进行零样本文本到语音的合成；I-DEA-TTS[15]引入渐进式解耦结构，可仅通过参考语音生成具备环境特征的个性化语音；DS-TTS[16]模型结合双风格编码与动态生成机制，改善了零样本条件下的自然度与音色保真度；MiniMax-Speech[17]通过可学习说话人编码器与 Flow-VAE 解码结构，实现了高质量的跨语言个性化语音合成。面向中文场景，IndexTTS[18]系统在 XTTS 架构的基础上引入字符-拼音联合建模与 BigVGAN2 解码器，显著提升了自然度与音色保持的效果。

本文基于 XTTS(A Massively Multilingual Zero-Shot Text-to-Speech Model)[19]语音合成模型，设计并实现了一个支持在线语音克隆与文本转语音的 Web 平台，旨在提升语音克隆的可用性与交互性。第 2 章对系统进行需求分析，明确了平台的功能目标和用户角色定位；第 3 章开展系统设计，提出了整体架构及功能模块划分，并详细说明了各子系统之间的逻辑关系；第 4 章对系统的核心功能执行流程进行阐述，重点介绍了语音克隆与文本转语音模块的运行逻辑与调用机制；第 5 章结合实际案例展示了平台在多语种与个性化语音合成场景中的应用效果，验证了系统的实用性；最后总结全文。

2 系统需求分析

2.1 功能性需求

结合语音克隆平台的应用需求，系统整体划分为 5 个核心模块：语音克隆、文本转语音、批量文本转语音、留言互动与系统管理。其中，前 3 个模块构成了平台的核心功能，直接影响语音合成任务的性能表现与用户体验。

1)语音克隆模块

该模块支持两种语音克隆方式：(1)基于 XTTS 模型，用户输入文本并上传音色样本生成个性化语音；(2)基于 FreeVC(Text-Free One-Shot Voice Conversion System)模型[20]，用户上传原始语音并指定目标音色，实现音色迁移。模块支持语速调节与语种切换，可满足多样化的语音克隆需求。

2)文本转语音模块

用户输入任意文本内容，系统调用 XTTS 模型生成对应语音，支持语速、音色和语种设置，并提供试听与下载功能，适用于日常播报与语音提示等场景。

3)批量文本转语音模块

该模块支持用户上传多个文本文件并统一配置参数，系统自动完成批量语音合成，适用于教育、宣传等场景的大规模语音内容生成任务。

4)留言互动模块

注册用户可在平台上发布、查看、编辑与删除留言，系统通过权限控制确保留言内容的安全性。

5)系统管理模块

面向后台管理员，提供用户信息管理、留言审核与运行日志监控等功能，保障系统稳定运行。

2.2 非功能性需求

为保障系统的稳定性，本平台在设计过程中充分考虑了性能、安全性与可扩展性等非功能性需求，具体包括以下 3 方面。

1)高性能与低延迟：语音克隆与文本转语音模块基于 XTTS 模型实现，须支持低延迟的语音合成处理。系统通过 MySQL 索引优化与分页查询提升数据库访问的效率，前端采用 AJAX 异步通信机制，提高了页面响应速度与用户体验。

2)安全性保障机制：平台采用 Session 机制进行用户身份认证，结合哈希加密、SQL 注入防护与输入校验等措施，确保用户数据与系统操作的安全性；同时，引入日志追踪功能，实现关键行为可追溯与审计。

3)良好的可扩展性：系统采用模块化设计与 RESTful API 架构，具备良好的可维护性与扩展能力，支持后续接入其他语音克隆模型及扩展留言交互功能等，为系统的功能升级与跨模块集成提供技术保障。

2.3 系统用例分析

图 1 和图 2 分别展示了用户(含普通用户与注册用户)与管理员在系统中的用例关系，反映了不同角色在平台中的功能权限，有助于明确系统功能需求与角色划分。

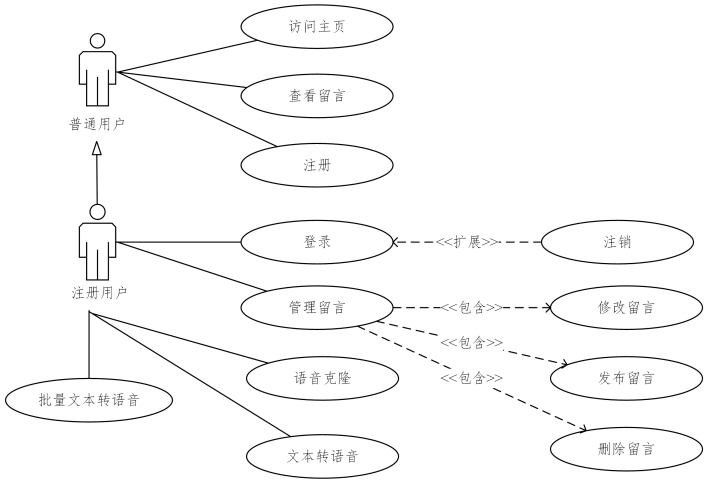


图 1　用户用例图

Fig. 1　Diagram of user use case

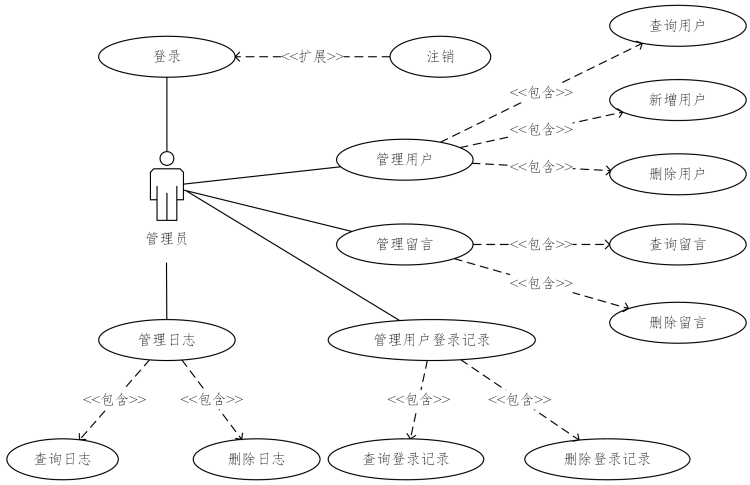


图 2　管理员用例图

Fig. 2　Diagram of administrator use case

3　系统设计

3.1　系统架构设计

本系统的整体架构如图 3 所示，由 3 部分组成：展示系统（Web 前端）、服务器（Flask 后端）与数据存储系统（MySQL 数据库）。平台采用前后端分离架构，各模块通过统一接口协同工作，实现语音克隆业务的高效处理与数据交互。

展示系统基于 HTML，CSS，JavaScript 和 Bootstrap 等前端技术栈构建，实现用户界面的动态渲染与交互控制。平台区分未注册用户、注册用户与管理员 3 类角色，分别对应不同的访问权限。前端通过 AJAX 实现与后端的异步通信，并结合 Session 机制完成用户身份验证与权限管理，提升了平台的安全性与用户交互体验。

服务器基于 Flask 框架构建，负责接收并处理用户请求，调用 XTTS 模型完成语音克隆与文本转语音等核心任务，并通过 RESTful API 实现各功能模块间的解耦，提升了系统的可扩展性。

数据存储系统采用 MySQL 数据库设计，统一管理用户信息、语音生成记录和留言内容等数据，通过结构化数据建模与索引优化，有效保障了数据的完整性与访问效率。

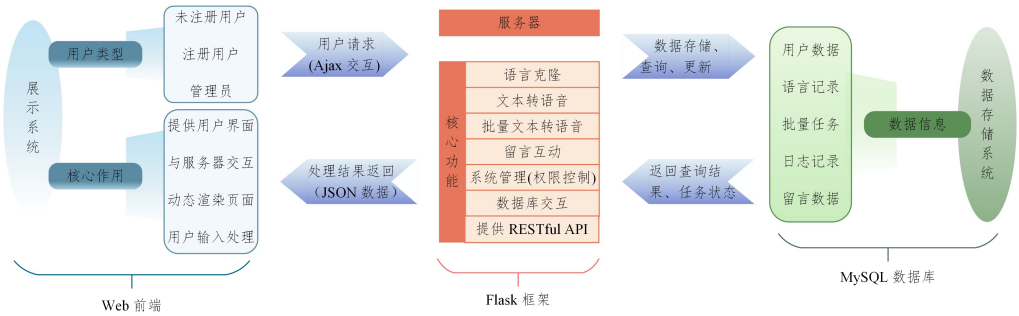


图 3　系统架构

Fig. 3　Architecture of the proposed system

3.2 系统功能模块设计

系统基于 XTTS 与 FreeVC 两种语音合成模型构建，整体划分为五大功能模块：语音克隆、文本转语音、批量文本转语音、留言互动与系统管理。其中，语音克隆与文本转语音模块构成系统的核心，提供多语种、个性化的语音合成功能；批量文本转语音模块面向大规模文本的自动转语音需求；留言互动模块实现用户间的信息交流；系统管理模块则用于平台的用户维护与数据审计。

系统功能模块如图 4 所示，后文将对主要模块进行详细介绍。

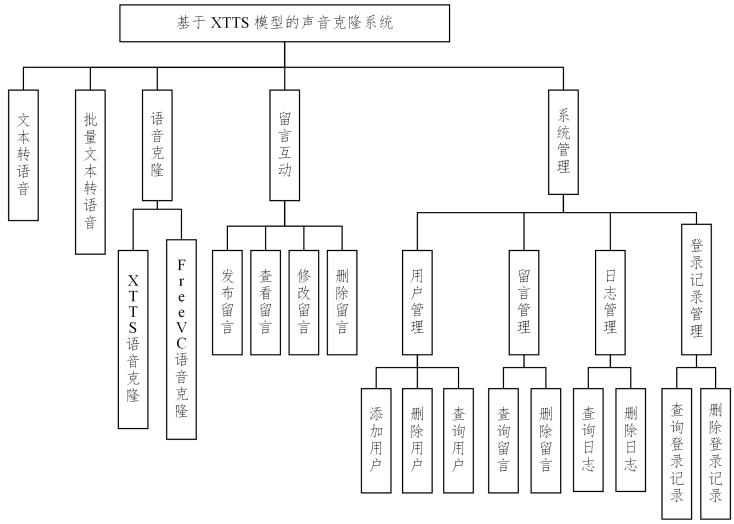


图 4 系统功能模块

Fig. 4 System functional module

3.2.1 语音克隆模块设计

语音克隆模块基于 XTTS 与 FreeVC 两种模型构建，支持用户在多种场景下实现个性化语音合成与音色迁移。图 5 展示了语音克隆模块的整体交互流程。

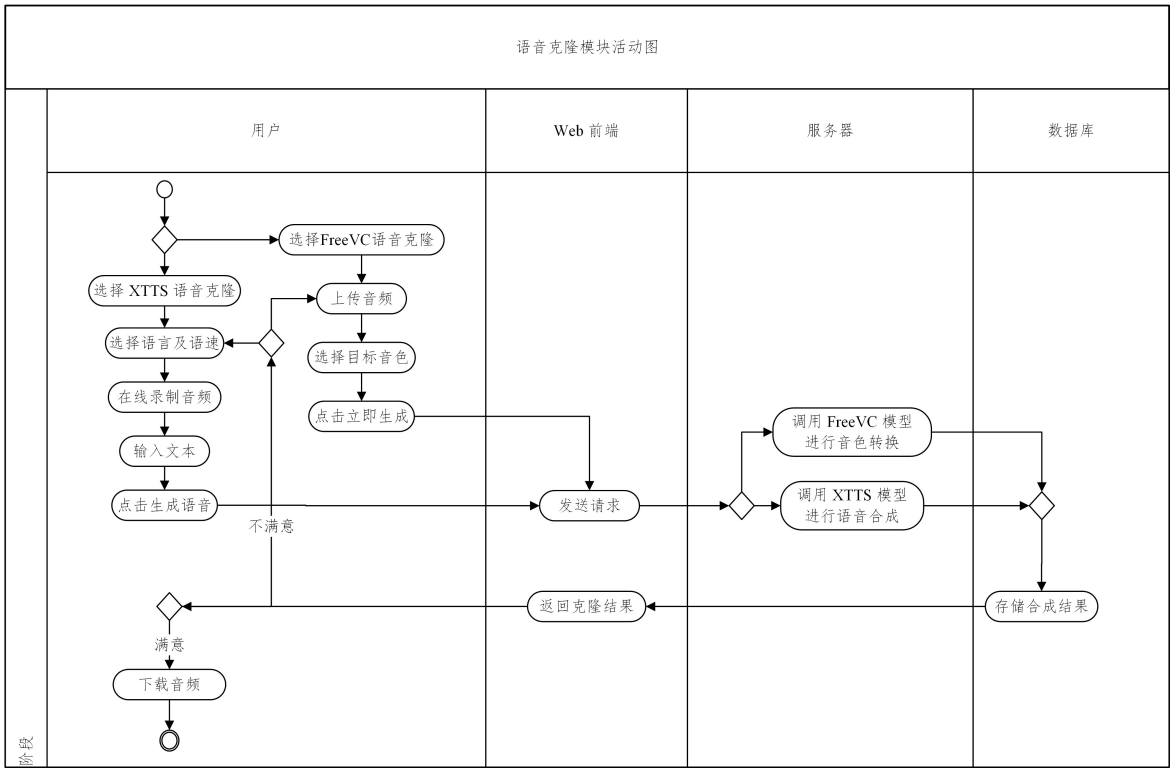


图 5 语音克隆模块活动图

Fig. 5 Activity diagram of voice cloning module

图 5 中，XTTS 模型通过用户上传或在线录制的音色样本，结合输入文本，自动提取声学特征并生成对应音色的语音，适用于多语言、多语速的定制化合成需求。FreeVC 模型则采用端到端的语音转换方式，用户上传源音频及目标音色样本后，系统在无需文本输入的前提下完成音色迁移，适合配音再加工、语音替换等应用场景。

3.2.2 文本转语音模块设计

文本转语音模块主要依托 XTTS 模型实现，将用户输入

的文本内容转换为自然流畅的语音输出。模块支持多语言合成、音色自定义与语速调节，可灵活适配多种语音合成场景与个性化需求。系统通过前端页面收集用户输入的文本与合成参数，后端在接收请求后完成语音合成任务，并记录用户操作信息及生成结果，保障后续数据的管理与使用。

图 6 展示了文本转语音模块的整体工作流程。

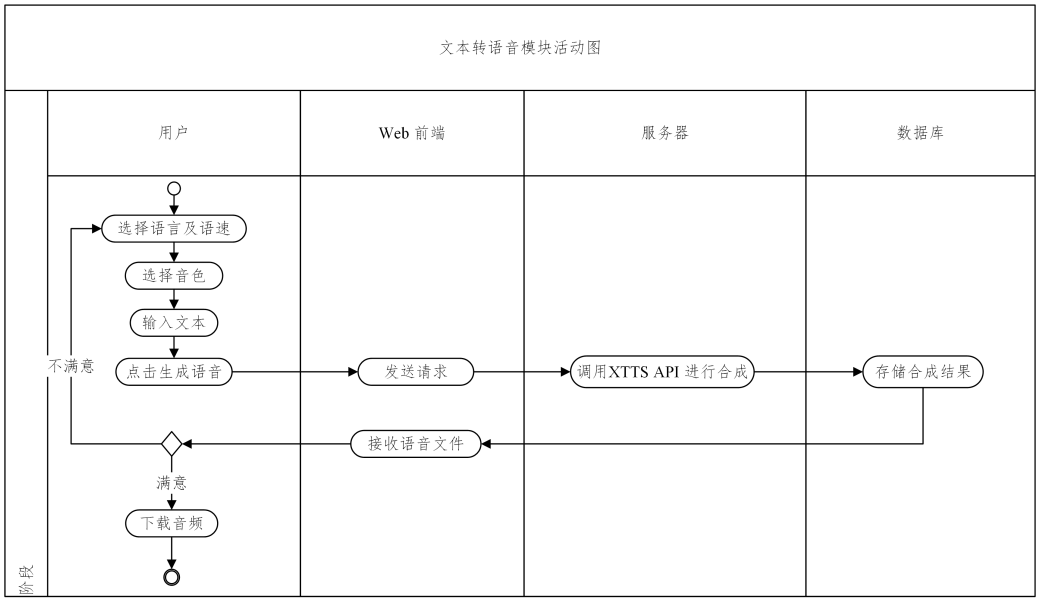


图 6 文本转语音模块活动图

Fig. 6 Activity diagram of text-to-speech module

3.2.3 批量文本转语音模块设计

批量文本转语音模块面向大规模语音合成需求，支持用户一次性导入多个文本文件，并统一配置音色、语速等合成参数。系统自动解析各文本内容，依次完成语音合成任务，并将生成的音频文件统一存储管理，提升了多文本场景下的处理效率。图 7 展示了批量文本转语音模块的执行流程。

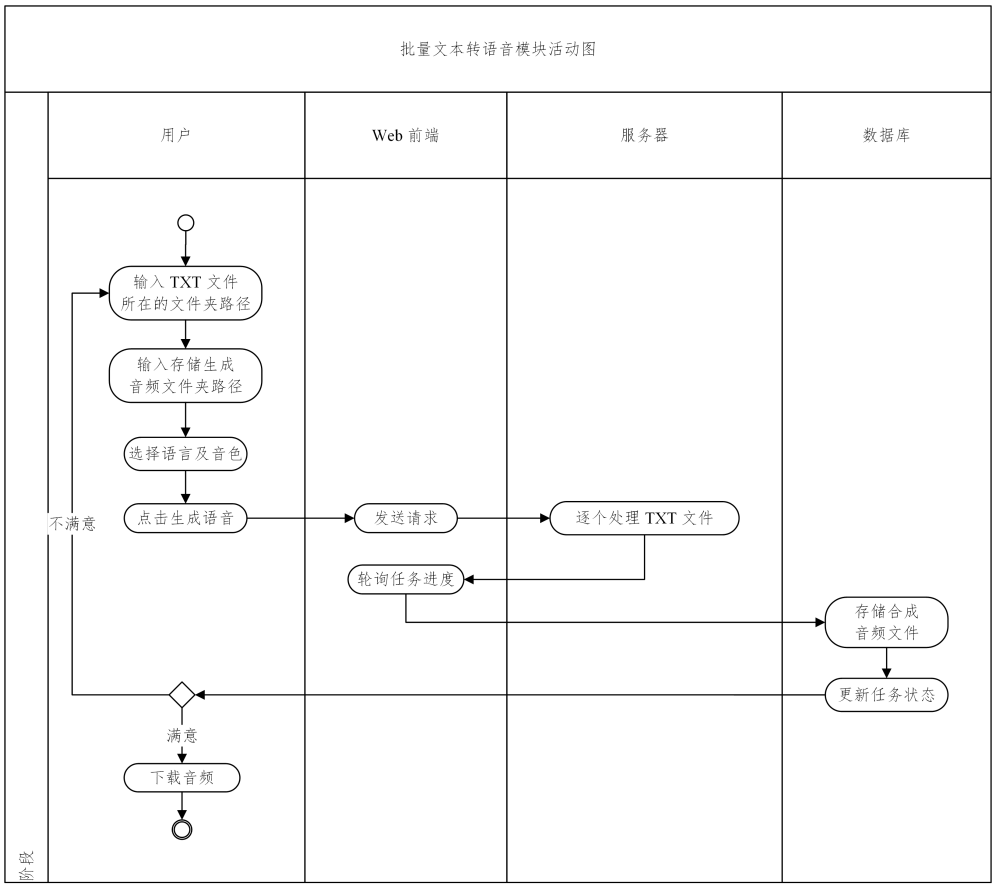


图 7 批量文本转语音模块活动图

Fig. 7 Activity diagram of batch text-to-speech module

3.3 语音克隆模型设计

3.3.1 XTTS 模型设计

XTTS 是一种支持零样本、多语言语音合成的深度学习模型，具备在 16 种语言之间进行高质量语音克隆与文本转语音的能力。该模型在 Tortoise 架构的基础上进行优化，显著提升了多语言适应能力、语音自然度与推理效率，适用于跨语种个性化语音合成场景。

XTTS 模型的训练框架由 VQ-VAE 编码器、GPT-2 编码器和 HiFi-GAN 解码器 3 部分组成，整体结构如图 8 所示。

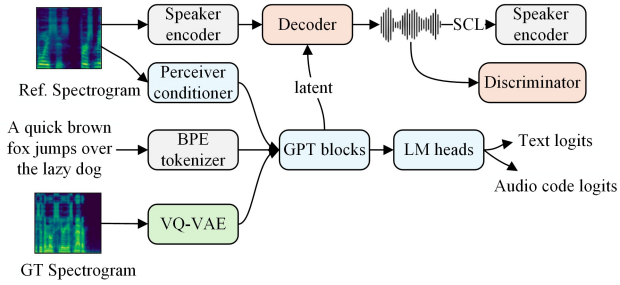


图 8 XTTS 模型的训练框架概览

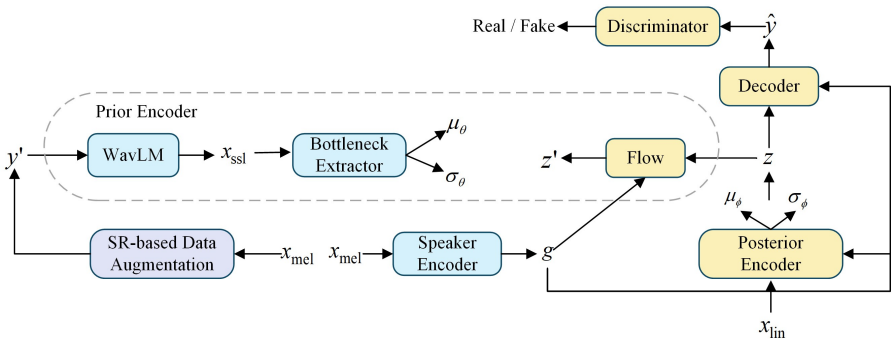
Fig. 8 Overview of XTTS model training framework

首先，VQ-VAE 对输入的 Mel 频谱图进行离散编码，使用 8192 个代码字典中的编码表示语音帧，以实现高效的信息压缩与建模；随后，GPT-2 编码器在文本输入与音频风格条件嵌入的双重引导下生成中间表示向量，用于语音内容与音色信息的融合建模；最终，HiFi-GAN 解码器接收上述表示向量，并重建出自然流畅的高保真语音输出。为增强模型的多语言泛化能力与音色保持性能，XTTS 同时引入多语言嵌入机制与 Speaker Consistency Loss(SCL)机制，有效提升了其在跨语种语音克隆任务中的表现能力[19]。

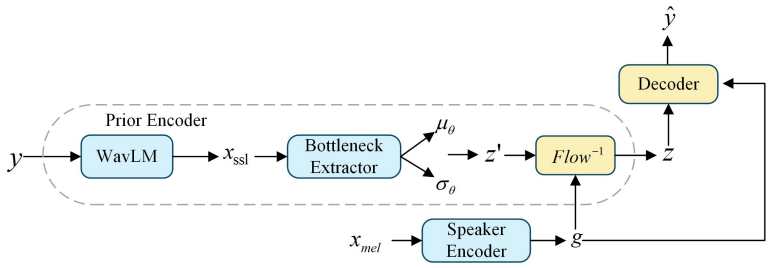
3.3.2 FreeVC 模型设计

FreeVC 是一种无需文本输入的端到端语音转换模型，基于 VITS 框架构建，能在保持语义内容不变的前提下实现高质量的音色替换。该模型引入 WavLM 预训练网络提取语义特征，并通过信息瓶颈机制去除说话人相关属性，实现内容与音色的有效解耦。说话人特征则由独立的 Speaker Encoder 提取，并结合归一化流(Flow)进行重建表达。

如图 9 所示，FreeVC 模型将训练与推理过程分离，整体结构由先验编码器、后验编码器、说话人编码器、解码器与判别器构成。在训练阶段，模型通过 WavLM 和信息瓶颈模块提取语义内容特征，结合目标音色信息，借助对抗机制实现高保真语音重建。推理阶段则输入源音频与目标音色，完成内容保持下的音色转换。为提升模型的鲁棒性与泛化能力，训练过程中还引入了基于声谱图缩放的增强策略[20]。



(a) Training procedure



(b) Inference procedure

图 9 FreeVC 模型的训练与推理流程图

Fig. 9 Training and inference flowchart of FreeVC model

图 9 中各符号的含义如下：y 表示源语音波形，y' 为增强后的语音波形，$\hat{y}$ 为转换后的语音波形，x_{mel} 表示梅尔频谱图，x_{lin} 表示线性频谱图，x_{ssl} 为自监督学习提取的特征(SSL 特征)，g 为说话人嵌入向量。

4 系统实现

4.1 文本转语音模块实现

该模块基于 XTTS 模型实现文本到语音的转换，支持多语言、多音色、个性化语速等参数配置(见图 10)，整体流程如下。

1)文本输入：用户在前端页面输入待合成的文本内容，系统支持直接输入或导入 srt 文件形式的文本资源。

2)参数设置：用户可通过界面下拉选项配置语种、语速倍数和音色参数。若启用语音克隆功能，可上传目标说话人的音频样本，或通过系统内置录音工具实时采集语音，系统将自动提取音色特征用于个性化语音合成。

3)请求提交：前端通过 AJAX 技术将文本内容及配置信

息异步发送至后端服务器，实现无刷新交互，提高用户体验与系统响应效率。

4）语音合成：服务器接收请求后，调用 XTTS 模型执行语音合成任务。

5）音频返回：合成结果以音频文件形式保存在服务器的指定路径，并将文件地址返回至前端页面，用户可在线试听或下载保存。

图 10　文本转语音界面

Fig. 10　Text-to-speech interface

4.2　语音克隆模块实现

语音克隆模块基于 FreeVC 模型构建，支持用户在无需文本输入的条件下实现音色迁移（见图 11），即将源语音内容转换为指定的目标音色，整体流程如下。

1）音频上传：用户通过前端页面上传本地音频文件（支持 WAV 和 MP3 等常见格式），将其作为语音克隆的源语音输入。

2）目标音色选择：音频上传成功后，用户可从系统预设音色库中选择目标音色样本，也可通过在线录制或上传文件的方式提供个性化音色作为参考。

3）模型调用：前端使用 AJAX 技术将源音频与目标音色参数一并提交至后端服务器，系统调用 FreeVC 模型执行音色转换操作，生成目标音色的语音。

4）音频生成：后端完成处理后，将生成的音频文件保存在服务器的指定目录，并将访问链接返回至前端页面，用户可在线试听或下载保存。

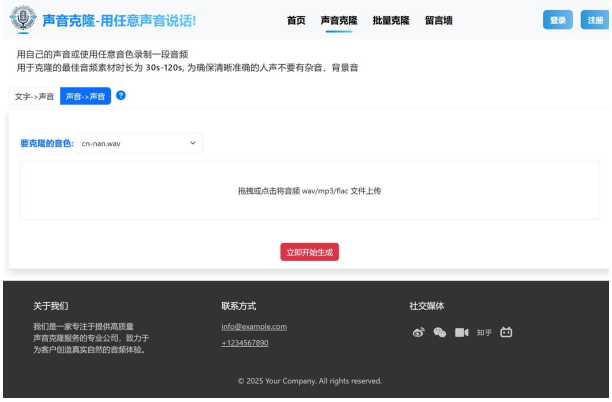


图 11　语音克隆界面

Fig. 11　Voice cloning interface

4.3　批量文本转语音模块实现

批量文本转语音模块用于处理大规模文本文件的语音合成任务，支持用户一次性上传多个文本文件并生成对应音频（见图 12）。该模块基于 XTTS 模型实现，整体流程如下。

1）输入文件路径：用户在前端页面输入待处理文本文件所在的目录路径，并设定音频输出文件夹，系统自动读取该路径下的所有文本内容。

2）参数配置：用户通过界面选择语种与音色参数，系统支持调用预设音色或上传自定义音色文件，以满足个性化合成需求。

3）任务提交：用户完成配置后，前端通过 AJAX 异步提交任务请求至后端服务器，并实时展示提交状态与任务进度反馈，提升交互可视性。

4）后端合成处理：服务器依次处理各文本文件，调用 XTTS 模型完成语音合成，并在前端同步更新任务执行状态。

5）结果返回：所有合成音频文件统一保存在用户设定的输出目录中。

图 12　批量文本转语音界面

Fig. 12　Batch text-to-speech interface

5　系统应用实例

为验证系统在实际场景中的语音合成效果，选取典型实例对文本转语音与语音克隆模块的功能进行展示与评估。

5.1　文本转语音模块应用实例

在文本转语音功能的测试中，选取以下语句作为合成内容：“只有坚定文化自信，持续推进文化建设，真正实现文化强，才能托举起中华民族伟大复兴的中国梦。”

如图 13 所示，用户在系统前端输入上述文本后，设定合成语言为中文，语速为 1.0（标准语速），并选择系统预设的音色文件“en-SteffanNeural.wav”作为目标音色。随后，系统调用 XTTS 模型执行语音合成任务，并将生成的语音保存为音频文件。

生成的语音在音质清晰度和自然度方面表现稳定，语言流畅，适用于中文播报、语音提示等标准应用场景，验证了模型在通用文本输入条件下的合成能力。

图 13 文本转语音案例

Fig. 13 Example of text-to-speech

为进一步测试文本转语音模块中语音克隆功能的表现，选取以下语句作为合成内容：“计算机科学与技术学科分别于2017年、2022年两次入选国家双一流建设学科，2024年，计算机科学首次进入全球排名前万分之一，排名全国第3。”

如图14所示，用户上传主持人董卿的一段朗读片段作为音色参考样本，语言设置为中文，语速倍数为1.0。系统在接收文本与目标音色样本后，调用XTTS模型进行语音克隆。

图 14 基于XTTS模型的语音克隆示例

Fig. 14 Voice cloning example based on XTTS model

生成结果在音色还原度方面表现良好，语音清晰，具有一定的个性化风格，能够较为准确地模拟目标音色。然而，由于XTTS的当前版本在情感表达与语调建模上的能力有限，合成语音在语气变化与情绪渲染方面尚存在不足。整体来看，该案例验证了XTTS模型具备在个性化语音合成与低资源语音克隆任务中进行实际应用的潜力。

5.2 语音克隆模块案例

在本案例中，用户上传了一段来自主持人康辉的朗读音频作为源语音输入，并选择系统预设音色“cn-nan.wav”作为目标音色，如图15所示。

图 15 基于FreeVC模型的语音克隆示例

Fig. 15 Voice cloning example based on FreeVC model

系统调用FreeVC模型完成音色转换。生成结果显示，输出语音在语义准确性、语速节奏方面均高度还原了原始语音，同时音色已成功迁移为所选目标音色。

相较于XTTS模型依赖文本与音色嵌入进行语音克隆，FreeVC采用端到端的音频到音频转换路径，具备无需文本参与的优势，更适用于配音再加工、个性化语音替换等实际应用场景。该案例验证了FreeVC模型在短时语音片段中的音色转换精度与语义保持能力，体现出其在内容解耦与音色合成方面的良好性能。

5.3 批量文本转语音模块案例

在本次测试中，输入目录包含4个文本文件(a. txt，b. txt，c. txt，d. txt)，用户选择合成语言为英文，语速默认为1.0，目标音色为系统预设的“en-MichelleNeural. wav”。任务提交后，系统依次调用XTTS模型对各文本文件进行语音合成，并将生成的音频文件保存至指定输出目录，最终输出文件命名为a. txt. mp3，b. txt. mp3，c. txt. mp3和d. txt. mp3，文件命名与原始文本一一对应，如图16所示。

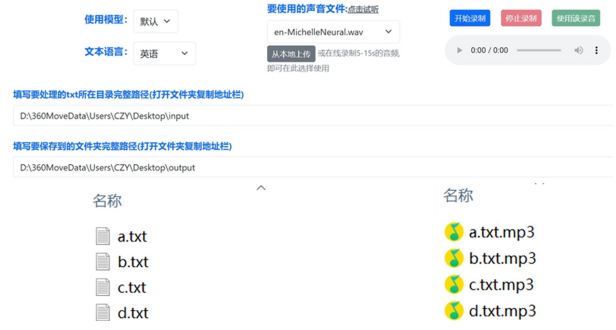


图 16 批量文本转语音处理案例

Fig. 16 Example of batch text-to-speech processing

结果表明，该模块具备良好的批量处理能力，能够高效完成多文本语音合成任务，支持统一参数配置与路径管理。合成语音在清晰度、语言节奏与稳定性方面表现良好，验证了XTTS模型在批量语音合成场景下的实用性。

结束语 本文围绕语音克隆技术的实际应用需求，设计并实现了一个基于XTTS模型的多功能语音合成平台。系统集成XTTS与FreeVC两种语音合成模型，支持文本转语音、个性化语音克隆和批量语音生成等功能，具备良好的灵活性与可扩展性。系统采用前后端分离架构，前端交互友好，后端处理稳定，整体运行效果良好。

通过实际应用案例验证，平台在合成语音的音色相似度、语义准确性与语音清晰度等方面表现优异，能够满足个性化播报、音色迁移与教育内容生成等多样化应用场景的需求。后续研究可进一步结合多模态情感建模与语音表达控制机制，提升合成语音的情感表现力与自然度，拓展系统在智能交互、虚拟人语音构建等方向的应用能力。

参 考 文 献

[1] GRAVES A. Generating sequences with recurrent neural networks [J]. arXiv:1308.0850,2013.

[2] SHEN J,PANG R,WEISS R J,et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]// Proceedings of the 2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2018.

[3] ARIK S Ö,CHRZANOWSKI M,COATES A,et al. Deep voice: Real-time neural text-to-speech[C]// Proceedings of the International Conference on Machine Learning. PMLR,2017.

[4] GIBIANSKY A,ARIK S,DIAMOS G,et al. Deep voice 2: Multi-speaker neural text-to-speech [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:2966-2974.

[5] PING W,PENG K,GIBIANSKY A,et al. Deep voice 3:Scaling text-to-speech with convolutional sequence learning [J]. arXiv:

1710.07654，2017.

[6] SNYDER D，GARCIA-ROMERO D，SELL G，et al. X-vectors：Robust dnn embeddings for speaker recognition[C]// Proceedings of the 2018 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2018.

[7] YANG J，LEE J，KIM Y，et al. VocGAN：A high-fidelity real-time vocoder with a hierarchically-nested adversarial network [J]. arXiv：2007.15256，2020.

[8] 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.

[9] JANG W，LIM D，YOON J，et al. Univnet：A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation [C]// Proceedings Interspeech 2021. 2021：2207-2211.

[10] MORRISON M，KUMAR R，KUMAR K，et al. Chunked autoregressive gan for conditional waveform synthesis [C]// International Conference on Learning Representations. 2021.

[11] CHEN S，WANG C，WU Y，et al. Neural codec language models are zero-shot text to speech synthesizers [J]. IEEE Transactions on Audio，Speech and Language Processing，2025，33：705-718.

[12] LI T，WANG Z，ZHU X，et al. U-Style：Cascading U-Nets With Multi-Level Speaker and Style Modeling for Zero-Shot Voice Cloning [J]. IEEE/ACM Transactions on Audio，Speech and Language Processing，2024，32：4026-4035.

[13] QIN Z，ZHAO W，YU X，et al. Openvoice：Versatile instant voice cloning [J]. arXiv：2312.01479，2023.

[14] WANG Y，ZHAN H，LIU L，et al. Maskgct：Zero-shot text-to-speech with masked generative codec transformer [J]. arXiv：2409.00750，2024.

[15] LU Y X，DU H P，SHENG Z Y，et al. Incremental Disentanglement for Environment-Aware Zero-Shot Text-to-Speech Synthesis[C]// Proceedings of the ICASSP 2025－2025 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2025.

[16] MENG M，YANG Z，YANG J，et al. DS-TTS：Zero-Shot Speaker Style Adaptation from Voice Clips via Dynamic Dual-Style Feature Modulation [J]. arXiv：2506.01020，2025.

[17] ZHANG B，GUO C，YANG G，et al. Minimax-speech：Intrinsic zero-shot text-to-speech with a learnable speaker encoder [J]. arXiv：2505.07916，2025.

[18] DENG W，ZHOU S，SHU J，et al. IndexTTS：An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System [J]. arXiv：2502.05512，2025.

[19] CASANOVA E，DAVIS K，GÖLGE E，et al. Xtts：a massively multilingual zero-shot text-to-speech model [J]. arXiv：2406，04904，2024.

[20] LI J，TU W，XIAO L. Freevc：Towards high-quality text-free one-shot voice conversion [C]// Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2023.

WANG Chencai，born in 2003，postgraduate. His main research interests include intelligent educational technology and so on.

MIAO Qiguang，born in 1972，Ph. D，professor，is a councillor of CCF(No. 09025D). His main research interests include computer vision，big data analysis and intelligent educational technology.

（责任编辑：柯颖）