基于TACOTRON端到端语音合成模型的改进方案

一：背景

语音合成就是将任意文本转换成语音的技术，即TTS。个典型的语音合成系统的前端部分主要是对输入文本进行分析并提取语音建模需要的信息，具体包括分词、词性标注、多音字消歧、字音转换、韵律结构与参数的预测等等。后端的部分读入前端文本分析结果，并结合文本信息对输出的语音进行建模。在合成过程中，后端会利用输入的文本信息和训练好的声学模型，生成语音信号。根据所采用的方法和框架不同，现阶段的语音生成器主要分为波形拼接、参数生成和基于波形的端到端统计合成这三种形式。现阶段语音合成发展的主要目标是进一步提高合成语音的清晰度与自然度、降低技术的复杂度等方面。

TACOTRON是一个直接从文本合成语音的神经网络架构，它将各模块放入一个黑箱，我们无需花费大量时间了解TTS中的各模块或者专业领域知识，可直接通过深度学习训练出一个TTS模型。TACOTRON作为集成的端到端TTS系统具有许多优点：它可以减少繁重的特征工程需要，更容易对各种属性（如说话者或语言）或情绪等高级功能进行丰富的调节，对新数据的适应也更容易。TACOTRON模型声码器部分使用Griffin-Lim 算法，会产生特有的人工痕迹并且合成的语音保真度较低。

Griffin-Lim算法是在仅已知幅度谱、未知相位谱的条件下重建语音的算法。Griffin-Lim算法将Seq2Seq的输出转化成被合成为波形的目标表达，使得估计得到的信号傅里叶变换的幅度值与原始信号傅里叶变换的幅度值的平方误差达到最小。通过迭代重构信号的相位信息和已知的幅度信息，得到语音信号的估算值。

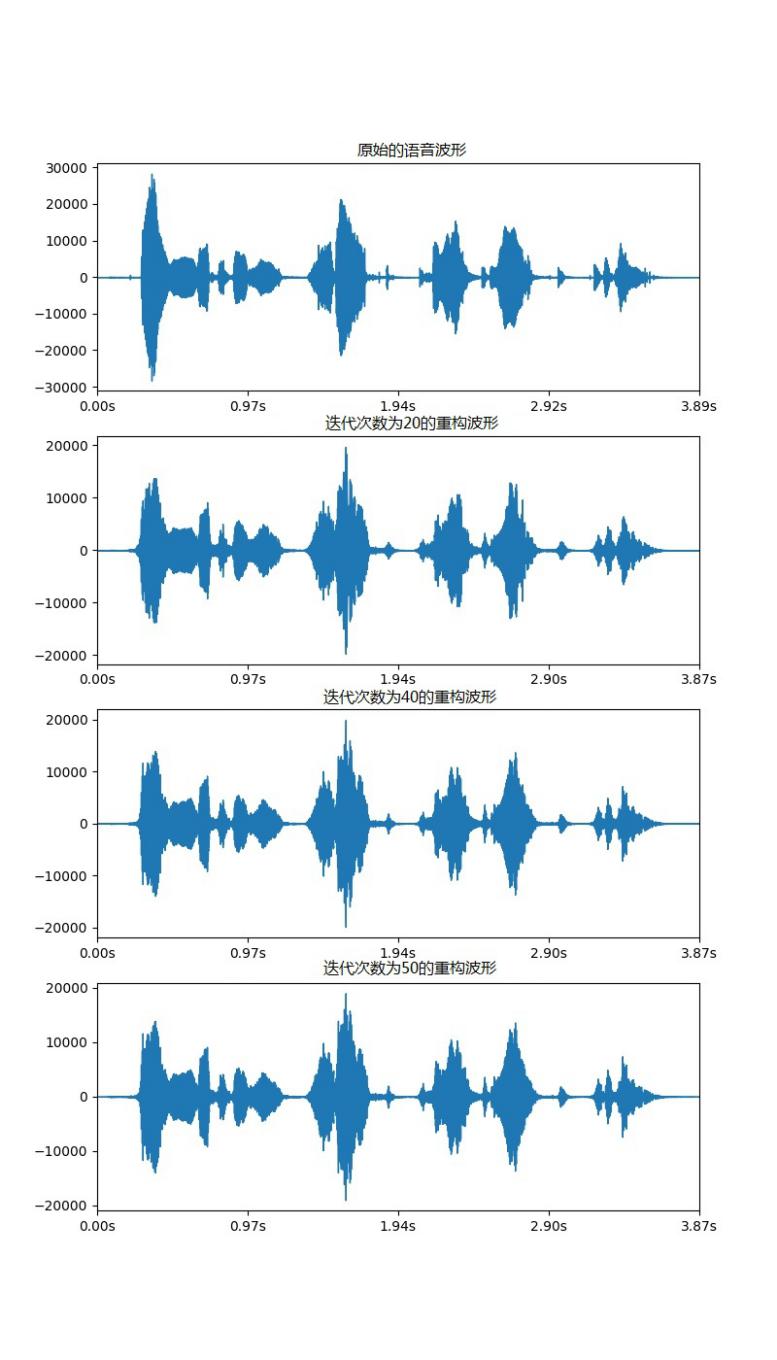


图1 原始波形与重构波形对比

二：创新点

端到端的语音合成模型相较于传统的语音合成算法在音质方面得到了前所未有的提高，没有了传统算法合成语音浓厚的“机器味”，同时还可以根据深度网络设计完备的训练字符集，使机器学会有情感的发声。但是高质量的语音合成需要消耗大量的计算资源，对于实时性要求较高的项目任务，可能还需要在质量和计算资源消耗方面做一个均衡。针对实时高质量的音频合成任务提出了WaveRNN网络模型架构来扩充TACOTRON模型的声码器部分。

WaveRNN声码器网络及其“黑科技”( Equivalent time, but higher quality)

a: 降低网络深度减少采样花费时间

b: 权重稀疏化减少采样计算时间

c: 子尺度WaveRNN减少单次训练样本数量

111111d: GRU网络层解决长序列依赖问题

图2 改进的TACOTRON结构

三：实用价值

团队设计开发了ITalk手机APP，提供了中文，英文的语音合成界面和实时播放系统，同时获取今日头条新闻检索，可以根据用户选择实现新闻标题的自动播报功能。

开发工具：Android Studio

开发语言：Java

测试平台：Android9.0

软件设计实现手机客户端可以和同在一局域网下的Python服务器端进行信息交互和指令控制，客户端和服务端程序执行流程如下：

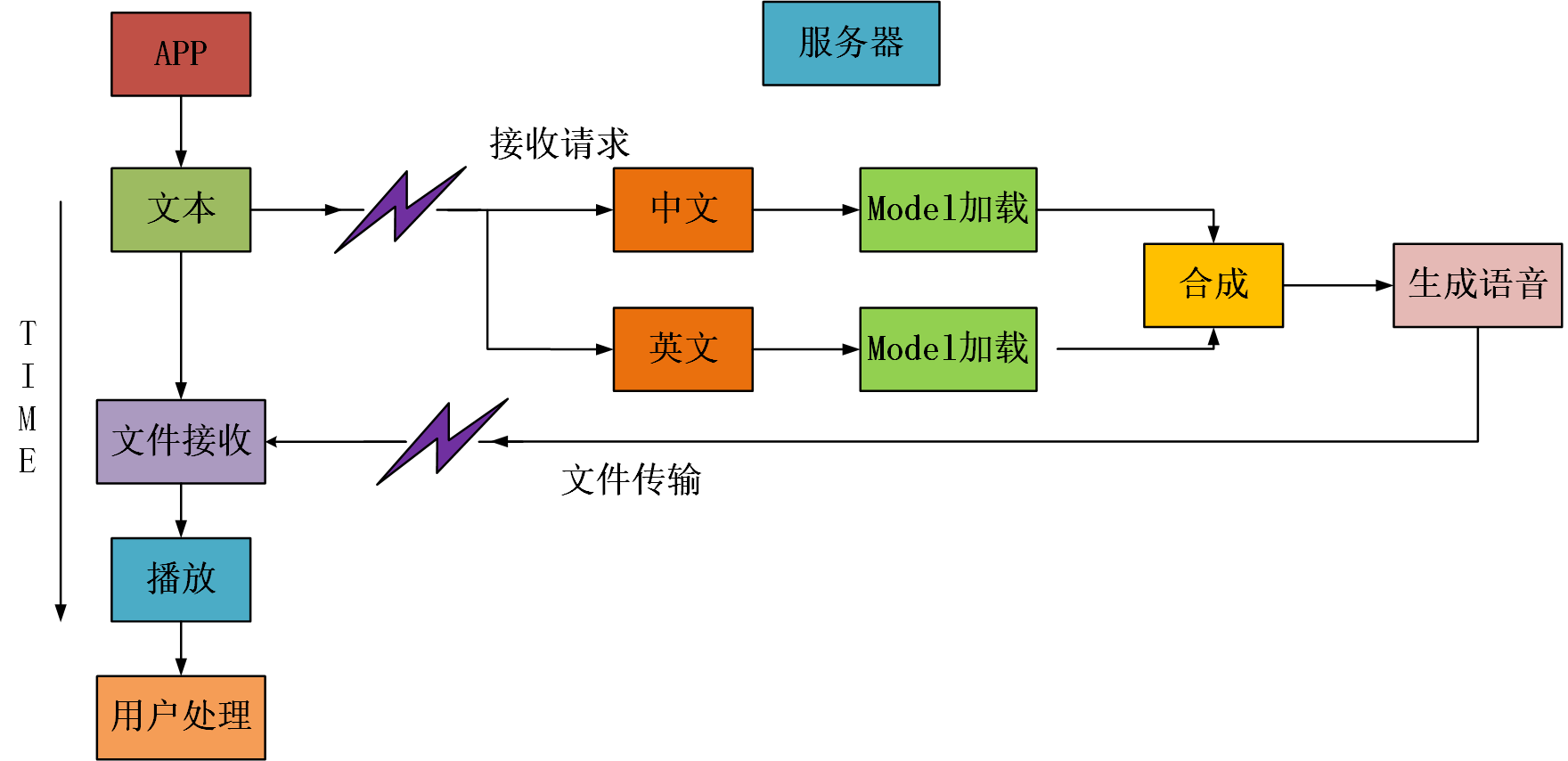


图3 APP软件系统框图

主要界面展示：





图4 APP主界面框架设计

四：总结

深度学习模型的特点很好的诠释了TTS，即文本到语音的合成流程，没有复杂的中间处理过程，而且完全的端到端合成模型，使得对各种语言的合成变得简单明晰。只需要送给模型目标语言的符号集以及对应的发音文件，模型就可以作为质量较佳的合成系统，开发者可以将注意力放在模型参数的优化和改良上来。深度学习给各行各业带来了繁荣与解决问题的新思路，语音合成也是人机交互的最基本需求，通过神经网络让语音合成日新月异，机器不但能听会说，还能分析，会思考，这是传统算法模型无法比拟的优点。我们坚信在神经网络的催化下，语音合成正在迈向新的里程碑，也准备好了迎接万物互联，万物智能时代的到来。