深度学习语音合成技术综述

张小峰[1],谢 钧[1],罗健欣[1],杨 涛[2]
1.中国人民解放军陆军工程大学 指挥控制工程学院,南京 210007
2.中国人民解放军 31121部队

摘 要:语音合成技术在人机交互中扮演着重要角色,深度学习的发展带动语音合成技术高速发展。基于深度学习的语音合成技术在合成语音的质量和速度上都超过了传统语音合成技术。从基于深度学习的声码器和声学模型出发对语音合成技术进行综述,探讨各类声码器和声学模型的工作原理及其优缺点,在此基础上对语音合成系统进行综述,系统综述经典的基于深度学习的语音合成系统,对基于深度学习的语音合成技术进行展望。

关键词:语音合成;声码器;声学模型;端到端语音合成系统
文献标志码:A **中图分类号:**TP183 doi:10.3778/j.issn.1002-8331.2101-0044

Overview of Deep Learning Speech Synthesis Technology

ZHANG Xiaofeng[1], XIE Jun[1], LUO Jianxin[1], YANG Tao[2]
1.Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China
2.Unit 31121 of PLA, China

Abstract: Speech synthesis technology plays an important role in human-machine interaction. The development of deep learning drives the rapid development of speech synthesis technology. Speech synthesis technology based on deep learning surpasses traditional speech synthesis technology in both quality and speed. This paper reviews speech synthesis technology based on deep learning vocoders and acoustic models, discusses the working principles and advantages and disadvantages of various vocoders and acoustic models, and then summarizes the speech synthesis system, systematically reviews the classic speech synthesis system based on deep learning, and finally looks forward to the speech synthesis technology based on deep learning.

Key words: speech synthesis; vocoder; acoustic model; end to end speech synthesis

语音合成是将文本转化成语音的技术,在日常生活中有着广泛的应用,比如智能客服、虚拟助手、智能阅读等[1-2]。目前,常见的语音合成方法有拼接法(concatenative speech synthesis)、参数法(parametric speech synthesis)、统计参数法(Statistical Parametric Speech Synthesis,SPSS)以及深度学习法(deep learning)等。其中,拼接法和参数法也叫传统语音合成方法。

文献[3]从前端文本处理、语音生成角度对语音合成技术进行综述,首先综述语音合成中前端文本处理技术,然后综述语音合成技术,文献[3]综述的语音合成技术有拼接法、参数法以及前馈神经网络等。文献[4]则首先简述了传统的语音合成方法,然后从深度神经网络在语音合成技术中的应用角度综述语音合成技术,比如受限玻尔兹曼机、深度置信网、循环神经网络等在语音合成中的应用,最后介绍了基于Wavenet[5]和Tacotron的语音合成技术。

不同于上述的语音合成技术综述,文章从声学模型和声码器角度,系统综述基于深度学习的语音合成技术进行,包括各种基于深度学习声学模型、声码器和端到端语音合成系统等,并分析了各自的特点、优势、劣势,以及适应的场景。文章涉及了上述综述中基于深度学习的语音合成技术,但没有涉及到传统的语音合成技术。

自2016年谷歌公司提出Wavenet声码器至今,涌现

基金项目:国家部委科技基金;江苏省自然科学基金青年基金项目(BK20150722)。
作者简介:张小峰(1992—),男,硕士研究生,研究领域为计算机网络、人工智能;谢钧(1973—),通信作者,男,博士,教授,研究领域为智能信息处理、计算机网络等,E-mail:xiejun73@189.cn;罗健欣(1984—),男,博士,讲师,CCF会员,研究领域为深度学习、智能信息处理、图形学;杨涛(1990—),硕士研究生,研究领域为图像拼接、多源视频流拼接。
收稿日期:2021-01-04 **修回日期:**2021-02-26 **文章编号:**1002-8331(2021)09-0050-10

出多种基于深度学习的语音合成技术,这些技术从合成语音质量、合成速度以及模型复杂度等方面提高语音合成技术性能。基于深度学习的语音合成系统主要有两种,一种是将深度学习应用到传统语音合成系统各个模块中建模,这种方法可以有效地合成语音,但系统有较多的模块且各个模块独立建模,系统调优比较困难,容易出现累积误差。这种系统的代表是百度公司提出的Deep Voice-1[6]和Deep Voice-2[7]。

另一种是端到端语音合成系统,这种系统旨在利用深度学习强大的特征提取能力和序列数据处理能力,摒弃各种复杂的中间环节,利用声学模型将文本转化中间表征,然后声码器将中间表征还原成语音。声学模型和声码器是端到端语音合成系统的重要组成部分,一些语音合成系统会在声学模型之前加上文本分析模块,文本分析模块对输入的文本进行预处理,比如词性标注、分词以及韵律生成等。由于文本分析是自然语言处理的内容,所以文章不进行综述。

图1是语音合成系统分类示意图,端到端语音合成系统由声学模型和声码器两个部分组成,声学模型实现文本和语音在时间上的对齐,声码器将声学模型输出还原成语音波形。由于声码器不仅可以在端到端语音合成系统中,也可以单独作为语音合成的模型,所以文章首先对基于深度学习的声码器进行综述,详细介绍经典的声码器这些工作原理,分析声码器的优缺点。然后对声学模型进行综述,在声码器和声学模型的基础上对语音合成系统进行综述。最后对基于深度学习的语音合成技术发展做出展望。

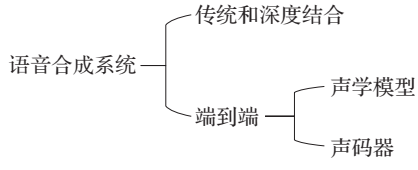

图1 语音合成系统分类

图2是近几年基于深度学习的语音合成技术发展历程,包括声码器、声学模型和端到端语音合成系统,还有一些衍生的声码器、声学模型和端到端语音合成系统没有列出。

1 声码器

声码器单独作为语音合成模型时需要和预处理模型结合使用,预处理模型为声码器提供语言学特征或声学特征等控制条件。基于深度学习的声码器按照其生成语音原理可以分为自回归式和并行式。自回归式声码器按照时间顺序生成语音,生成每一时刻的语音都依赖之前所有时刻的语音。并行式声码器并行生成语音,不再按照时间顺序。并行式声码器主要有基于概率密度蒸馏[8]声码器和基于流[9]声码器两种。基于概率密度蒸馏声码器是在自回归模型基础上结合逆自回归流[10]和概率密度蒸馏方法,采用教师学生模式训练模型,教师模型是预训练好的自回归模型,学生模型用逆自回归流方法训练,损失函数一般是教师模型和学生模型输出之间的相对熵(Kullback-Leibler散度,KL散度)。基于流的生成式模型最初用于生成图像,但基于流的声码器也可以较好地生成语音。基于流的声码器只需要一个模型、一个损失函数就可以高速生成语音,而基于概率密度蒸馏的声码器需要两个阶段训练和多个辅助损失函数。图3是声码器按照工作原理的分类,图中所列的是经典声码器,没有包含衍生的声码器。

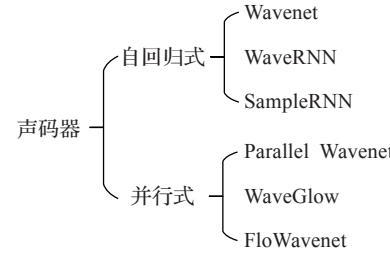

图3 基于深度学习的声码器分类

1.1 自回归式声码器

自回归声码器是较早研究的声码器,按其网络模型主要有基于卷积神经网络声码器和基于循环神经网络声码器两种。基于卷积神经网络声码器利用卷积神经网络建模,通过因果卷积、带洞卷积等达到按时序生成语音的目的。基于卷积神经网络声码器能够充分利用GPU等并行计算资源,训练速度比基于循环神经网络的声码器快。基于循环神经网络的声码器利用循环神经网络建模,比如长短时记忆网络(LSTM)[11]、门循环单元

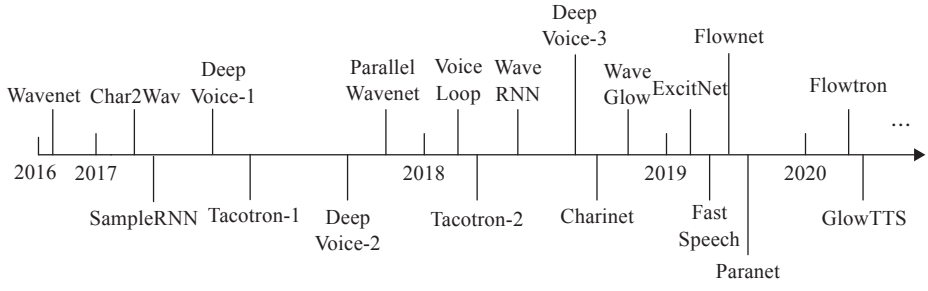

图2 基于深度学习的语音合成技术发展历程

(GRU)[12]等。

Wavenet[5]是谷歌DeepMind提出的由卷积神经网络构成的生成式模型,该模型不仅可以生成语音也可以生成图像、音乐等。Wavenet作为声码器时有Wavenet声码器[13]和条件Wavenet声码器两种模式。Wavenet声码器需要语言学模型或声学模型等提供控制条件,比如声谱图、F0基频等。图4是Wavenet结构示意图,训练前将语音序列联合概率 $x=\{x_1,x_2,\cdots,x_T\}$ 分解为各时刻条件概率的乘积,如公式(1)所示:

$$p(x)=\prod_{t=1}^{T}p(x_t|x_1,x_2,\cdots,x_{t-1}) \tag{1}$$

x 是语音波形值序列,x_t 是一个时刻波形值。训练过程中利用因果卷积和带洞卷积逐点生成语音的波形。因果卷积和带洞卷积保证模型按照时间顺序生成语音。生成阶段,模型采用逐采样点自回归方式,按公式(1)计算每个时刻的波形值。

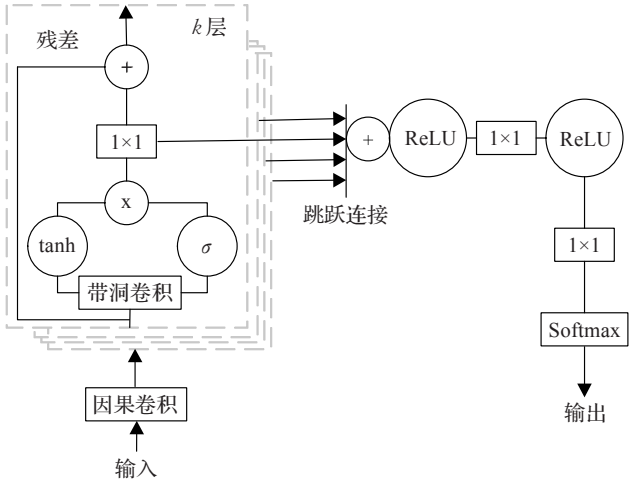

图4　Wavenet

条件Wavenet声码器在Wavenet声码器基础上接收额外的输入条件进行建模,如公式(2)所示:

$$p(x|h)=\prod_{t=1}^{T}p(x_t|x_1,x_2,\cdots,x_{t-1},h) \tag{2}$$

h 是额外的输入条件。比如在合成多说话人语音时,说话人编码就可以作为额外的输入,在文本转语音任务中,文本也可以作为额外的输入。条件Wavenet声码器有两种建模方法:全局方法和局部方法。全局方法是指模型接收单额外输入条件 h,该条件在所有的时间点上影响模型的输出。局部条件建模方法是指模型有第二种时间序列 h_t,h_t 是通过对原始数据的低采样率获得。局部条件建模效果比全局建模的效果要差一些。

Wavenet全卷积神经网络的设计可以充分利用GPU并行计算资源,训练速度比较快且合成语音质量高。但是由于采用逐点自回归方式生成语音,所以生成语音速度慢。对于16 kHz采样率,16 bit采样位的语音,每秒需要计算16 000个采样点,每个采样点有65 536种可能,这需要较大的计算量和计算耗时。虽然Wavenet采用了 μ 律压扩[14]将16 bit的采样位降低成8 bit的采样位,但是生成语音速度仍然达不到实时的要求。

虽然Wavenet存在模型较为复杂、计算量大及合成速度慢等问题,但Wavenet是基于深度学习声码器的重要组成部分,对语音合成技术的发展有着重要影响,基于Wavenet衍生出多种声码器。

对于Wavenet声码器需要额外的模型提供控制条件,文献[15]提出直接使用现有的声学特征作为Wavenet的输入特征,这样不需要对原始语音建模,从而避免了语音分析和生成过程各种先验的假设。文献[16]在文献[15]基础上研究多说话人的语音合成。

对于Wavenet合成语音有噪音的问题,文献[17]通过在Wavenet模型之前引入基于感知权重的噪音整形技术,减少Wavenet因为卷积神经网络难以捕捉语音信号动态信息而产生的噪音,但是这种做法会造成合成语音的质量下降。Excitnet[18]结合LPC(Linear Predictive Coding,LPC)声码器和Wavenet声码器优点,使用自适应预测器将与共振峰相关的频谱结构从输入语音信号解耦,然后通过Wavenet对激励信号的概率分布进行建模,实现了高效训练和生成语音。

针对Wavenet训练复杂,计算量大的问题,FastWavenet[19]在Wavenet基础上通过引入缓存,将因果卷积和带洞卷积计算过程中不变的结果保存在缓存中,需要时直接使用,这样可以节约计算的时间,提高训练和生成语音的速度。FFTNet[20]则对Wavenet结构进行调整,通过多通道卷积的结构,提高合成语音的速度同时减小模型规模且合成语音的质量较高。

不同于上述基于卷积神经网络的声码器,SampleRNN[21]采用处理序列数据常用的循环神经网络建模,SampleRNN有多层循环神经网络组成,最低层叫采样层,其余层叫帧层。各层以不同时钟速率运行,学习不同抽样级别。采样层一个时间步只学习一帧样本,而最高的帧层一个时间步可以学习较多帧样本。帧层对语音非重复帧做处理,每个帧层都是一个循环神经网络,循环神经网络可以由LSTM、GRU等组成,也可以是其他循环神经网络变体组成。采样层对单个样本数据进行处理,直接计算出语音波形。

SampleRNN和Wavenet都是直接对原始语音建模,Wavenet利用卷积神经网络学习语音前后特征的关系,而SampleRNN从循环神经网络的角度,利用多层循环神经网络学习语音不同时间维特征。SampleRNN生成语音速度比Wavenet快,模型也比Wavenet简单,合成语音质量和Wavenet相当。但是由于SampleRNN采用循环神经网络,所以训练过程中并不能充分利用计算资

源,需要采用较多的技巧加速模型训练。

不论是Wavenet还是SampleRNN都很难达到实时语音合成,WaveRNN[22]针对生成式语音合成中耗时较大的模块进行了专门分析和改进,达到超实时语音合成速度且合成语音质量也没有大的下降。WaveRNN认为生成式语音合成模型生成语音速度可以用公式(3)表示:

$$T(u)=|u|\sum_{i=1}^{N}(c(op_i)+d(op_i)) \tag{3}$$

$T(u)$是生成语音的总时间,语音有$|u|$个采样点,N表示模型神经网络层数,$c(op_i)$代表每层神经网络计算耗时,$d(op_i)$表示硬件执行程序时间。对于以上几点,WaveRNN分别使用单层循环神经网络和两层softmax层简单模型减少神经网络层数,采用裁剪法[23-24]减少每一层神经网络计算耗时,使用折叠法减少因为语音较长而需要的时间。

WaveRNN通过上述方法达到了提高语音生成速度、降低模型规模的效果。WaveRNN合成语音速度比Wavenet和SampleRNN快很多且可以部署在移动端等计算资源较少的设备上。LPCNet[25-26]在WaveRNN基础上结合线性预测模块,进一步提高了语音合成的效率,在同等大小模型的情况下,LPCNet合成语音的质量更高。

1.2 并行式声码器

自回归式声码器按照时间顺序构建语音,每一点语音值都需要依赖历史语音值,合成语音质量高,但实时性不足。不同于自回归式逐点生成语音,并行式声码器同时生成整段语音,每个语音点之间没有依赖关系。并行式声码器按其原理主要有两种:一种是基于概率密度蒸馏,利用逆自回归流和概率密度蒸馏训练和生成语音,比如Parallel Wavenet[27]和Clarinet[28]。逆自回归流是一种每个可逆函数都基于自回归神经网络的特殊的标准化流[29-30]。逆自回归流最大似然估计是自回归式,因此逆自回归流模型会和概率密度蒸馏模型结合使用。概率密度蒸馏原本是模型压缩的方法,通过优化学生模型和教师模型输出之间的误差,得到规模较小的学生模型。另一种是基于流,通过一系列可逆函数,利用简单的分布在中间表征的控制下生成语音波形,基于流的模型是目前研究的热点。

自回归模型并行输入训练数据逐点输出预测结果,特点是训练速度快但生成速度慢。逆自回归流模型相反,逐点输入训练数据并行输出预测结果,所以基于逆自回归流模型的训练速度慢,但是生成速度快。Parallel Wavenet结合逆自回归流和概率密度蒸馏方法,首先训练好自回归的教师模型,然后从教师模型中蒸馏学习出学生模型,学生模型用逆自回归流的方法训练。Parallel Wavenet损失函数是学生模型和教师模型输出之间的KL散度,通过优化KL散度使学生模型输出不断逼近教师模型输出。公式(4)是模型损失函数,$H(P_S,P_T)$是学生模型和教师模型输出的交叉熵,$H(P_S)$是学生模型输出的熵。

$$D_{KL}(P_S\|P_T)=H(P_S,P_T)-H(P_S) \tag{4}$$

仅优化KL散度,学生模型还不能很好地生成语音,Parallel Wavenet增加了感知损失、对比损失、能量损失等辅助损失函数提高学生模型生成语音质量。感知损失可以防止产生不好的发音,对比损失可以消除噪音,而能量损失协助匹配人类语音的能量。

Parallel Wavenet极大提高了语音合成速度,但实践中发现很难有效实现该模型,因为Parallel Wavenet首先假设教师和学生模型的输出均服从混合逻辑斯特分布,然后对混合逻辑斯特分布进行蒙特卡洛采样后计算两者的KL散度。蒙特卡洛采样会导致训练过程不稳定。对此,Clarinet采用高斯分布取代教师学生模型输出的混合逻辑斯特分布,两个高斯分布可以直接计算KL散度,公式(5)所示:

$$KL=(q\|p)=\ln\frac{\sigma_p}{\sigma_q}+\frac{\sigma_q^2-\sigma_p^2+(u_p-u_q)^2}{2\sigma_p^2} \tag{5}$$

q是学生模型输出的高斯分布,p是教师模型输出的高斯分布,u_q和σ_q为q均值和标准差,u_p和σ_p是p均值和标准差。为了保证训练过程中数值计算的稳定性,Clarinet还对σ_q和σ_p增加了正则项,如公式(6)所示:

$$KL^{\text{reg}}(q\|p)=KL(q\|p)+\lambda|\ln\sigma_p-\ln\sigma_q|^2 \tag{6}$$

Parallel Wavenet和Clarinet通过逆自回归流和概率密度蒸馏的方法实现了高速训练和生成语音目的,但在实际中这两种方法使用并不多,因为这两种模型训练过程比较复杂且合成效果也不稳定。

不同于Parallel Wavenet和Clarinet需要两个阶段训练和多个辅助损失函数。基于流的声码器只需要一个模型和一个损失函数就可以高速生成语音。标准化流是通过一系列可逆函数用简单分布模拟复杂分布的一种生成式模型。在语音合成中,对于原始音频x,假设存在可逆函数$f(x):x\to z$,则可以直接将x映射到先验P_z,则x的对数概率分布函数可用公式(7)表示,这样就可以用一个已知的简单分布计算一个复杂的分布的似然估计。

$$\ln P_x(x)=\ln P_z(f(x))+\ln\det\frac{\partial f(x)}{\partial x} \tag{7}$$

基于流的模型的优势是训练过程简单,不需要多次训练和多个损失函数且结果稳定。WaveGlow[31]结合Glow和Wavenet以均值为0的高斯球形分布作为先验分布,以梅尔频谱作为控制条件,通过挤压层、1×1可逆卷积和仿射耦合层[32]等拟合语音波形。FloWavenet[33]的原理和WaveGlow相似,FloWavenet认为基于流的生

成式模型要实现高效训练和生成需要满足两点:(1)计算公式(7)的雅克比矩阵的公式 f 应该简单易于计算;(2)将噪音点 z 转化成语音信号点 x 的逆函数 $x=f^{-1}(z)$ 应该容易计算。如要并行生成和易于计算,第二点必须满足。为了满足上述两点,FloWavenet使用类似文献[32]的仿射耦合层中,并在每一个模块中使用了多层流。

WaveGlow和FloWavenet实现了高速生成语音,降低了训练的难度和复杂度。但是基于可逆流的模型通常有较多的神经网络层和大量的参数,这使得基于流的模型很大,比如Wavenet和Clarinet只有1.7×10^6个的参数,而WaveGlow有200×10^6个的参数。

MCNN[34]通过多头卷积网络重建语音波形,该模型虽然不需要如Parallel Wavenet那样多次训练,但模型仍需要四个辅助损失函数,训练过程比较复杂,且由于MCNN是在完整的声谱图上还原波形,所以并不清楚能否在梅尔频谱这样的声谱图上还原波形。

文献[35]从人类发音角度建模,模型由三个部分组成,一个是源模块,用于生成基于正弦的信号,作为激励源。另一个是基于非自回归带动卷积的过滤器,过滤器将激励转化成语音波形。最后一个是条件模块,这个模块将输入的声学特征进行预处理,然后输入到前两个模块。通过这三个模型,文献[35]避免了基于概率密度蒸馏复杂的训练环节。

表1是几种声码器性能参数表,Wavenet合成语音质量高,但模型较为复杂,生成语音速度慢。SampleRNN和WaveRNN是基于循环神经网络的声码器,这两个模型比Wavenet小,合成语音质量和Wavenet相当。WaveRNN速度最快,模型最小。WaveGlow和FloWavenet是结合Wavenet和流的并行式声码器,两者的工作原理相近,结构不同。表中合成速度以实时为基准,实时是指合成一秒的语音需要一秒的时间,不足是指合成一秒语音需要超过一秒的时间。在实际应用中一般会要求语音合成系统具有超实时性,因为部署在服务器端的语音合成需要考虑网络时延问题,5G的发展会在一定程度上缓解这个问题,但不能从根本上解决问题,所以实用的语音合成系统需要可以实时合成语音。

2 声学模型

经典的语音合成系统中语言学模型和声学模型通常分开建模,这会导致模型比较复杂且需要研究人员具有专业的语言知识。随着深度学习的发展,特别是基于注意力机制[36-38]的序列到序列(seq2seq)模型[39]的提出,使得将语言学模型嵌入到声学模型中成为可能。文献[40]首次提出基于序列到序列加注意力机制[41]的声学模型,其原理是通过输入文本序列生成梅尔频谱。文献[40]声学模型可以很好地生成梅尔频谱,但需要预训练的隐马尔科夫模型辅助学习,因此文献[40]只能合成较短的语音。

和声码器相似,基于深度学习的声学模型的发展也经历了自回归式和并行式两个阶段。自回归式声学模型出现较早,生成的中间表征质量高、速度慢。并行式声学模型生成中间表征速度快,但质量会有所下降,且训练过程比较复杂。本章节从自回归和并行式两种声学模型综述。

2.1 自回归式声学模型

自回归式声学模型在语音合成中用的比较多,其基础是seq2seq加注意力机制的编码解码模型。编码器将文本转化成上下文矢量,解码器根据上下文矢量直接解码出结果。注意力机制建立起上下文矢量与输入文本之间的联系,这样可以更好地利用输入的信息。在语音合成声学模型中,注意力机制可以将解码器的注意力集中在输入的相关文本上。由于注意力机制在声学模型有着重要的作用,所以这里介绍一下注意力机制。

图5是seq2seq加注意力机制的示意图,对于输入序列 $\boldsymbol{x}=(x_1,x_2,\cdots,x_L)$ 和与之对应的输出序列 $\boldsymbol{y}=(y_1,y_2,\cdots,y_T)$,首先将输入序列编码成 $(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{Tx})$ 嵌入特征,s_t 为循环神经网络的隐层状态,如公式(8)所示,c_t 为第 t 时间步的上下文矢量,如公式(9)所示。公式(10)是解码器解码的结果。

$$s_t=RNN(\boldsymbol{s}_{t-1},\boldsymbol{y}_{t-1},\boldsymbol{c}_t) \tag{8}$$

$$\boldsymbol{c}_t=AttendContext(s_{ij},\boldsymbol{h}_j) \tag{9}$$

$$\boldsymbol{y}_t=Generate(\boldsymbol{y}_{t-1},\boldsymbol{s}_t,\boldsymbol{c}_t) \tag{10}$$

公式(9)中,上下文矢量 c_t 每次计算都和输入 h 有关,这种方式称为基于内容的注意力机制[41],这种方式

表1 基于深度学习的声码器总结

声码器	原理	主要神经网络	优、缺点	合成实时性
Wavenet	自回归	卷积神经网络	合成语音质量高,速度慢	不足
SampleRNN	自回归	循环神经网络	占用计算资源少,速度慢	不足
WaveRNN	自回归	循环神经网络	占用资源少,合成速度快	4x
Parallel Wavenet	并行式	卷积神经网络	合成速度快,训练过程不稳定	20x
WaveGlow	并行式	标准流+卷积神经网络	合成语音质量高,合成速度快,计算量大	25x
FloWavenet	并行式	标准流+卷积神经网络	合成语音质量高,合成速度快,计算量大	20x

注:4x、20x、25x指4倍、20倍、25倍超实时。

没有考虑输入文本之间的位置关系,相同的输入具有相似的权重值。另一种基于位置的注意力机制将输入文本之间的位置关系代入计算,上下文矢量计算公式如式(11)所示,这种称注意力机制为基于位置的注意力机制[37]。

$$c_t = AttendContext(s_t, c_{t-1}) \tag{11}$$

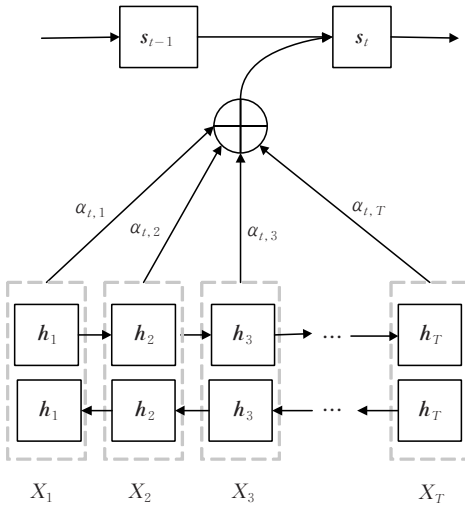

图5 循环神经网络加注意力机制示意图

2.1.1 编码器

编码器提取输入文本的特征,将输入文本转化成上下文矢量,编码器的结构一般比较简单。Tacotron-1[42-43]的编码器采用了预处理网络和CBHG[44]模块。预处理网络是一系列非线性转换层,将文本转化成嵌入矢量,CBHG模块结合注意力机制将嵌入矢量转化成上下文矢量,CBHG模块可以减少过拟合以及误发音等情况。Tacotron-2[45]编码器在Tacotron-1编码器基础上进行了简化,仅使用普通的LSTM和卷积网络,没有使用复杂的CBHG模块。Deep Voice-3[46-47]的编码器则采用全卷积进行建模,全卷积的编码器可以并行训练。

2.1.2 解码器

解码器是声学模型的关键,后期对于声学模型的改进也是集中在解码器。解码器根据上下文矢量直接解码出结果,解码的结果需要根据声码器选择。解码器可以解码出原始的声谱图,也可以解码出梅尔频谱。原始的声谱图可以保留更多的语音信息但是会有大量的冗余信息增加训练过程的不稳定性,梅尔频谱虽然会丢失大量的信息,但是通过良好设计的声码器可以还原出丢失的信息,因此解码器通常会以梅尔频谱为解码结果。

Tacotron-1采用基于内容注意力机制解码器[36],每一个时间步每一循环层都会生成一个注意力询问。Tacotron-2的解码器和Tacotron-1结构相似,但是采用基于位置的注意力机制[37],以更好地适应输入文本有重复字的情况。Deep Voice-3解码器则是由带洞卷积和基于Transformer[38]注意力机制构成,基于卷积的解码器比基于循环神经网络的声码器解码的速度要快一些。由于在合成语音过程中,对于注意力机制的错误较为敏感,所以Deep Voice-3进一步采用了[48]注意力机制。

上述的解码器可以高质量地生成中间表征,但是都存在曝光偏差[49]问题。曝光偏差是由于在训练过程中,解码器使用真实值作为每一个解码步的输入,而在预测时使用上一个解码步的输出作为本次解码步的输入,这种训练和预测的输入会造成不一致性。曝光偏差会导致合成语音出现漏词、重复词以及不完全合成等现象。

2.2 并行式声学模型

并行式声学模型并行解码出结果,并行解码的两个难点是根据输入的文本需要解码出多少帧中间表征以及并行解码出的各帧之间的依赖关系如何界定。目前常用的办法仍是概率密度蒸馏,通过预训练的教师模型指导学生模型。Fastspeech[50]通过从预训练好的自回归模型中提取文本语音对齐信息,结合Transformer、一维卷积网络[51]以及概率密度蒸馏等方法并行生成中间表征。Fastspeech解决了自回归声学模型合成速度慢、合成结果不稳定以及合成过程难以控制等问题。Paranet[52]则采用软注意力机制从自回归的教师模型中学习,并行生成中间表征。

不同于Fastspeech需要对齐模型或者Paranet依赖自回归的教师模型。GlowTTS[53]通过流和单调对齐搜索方法可以高速生成中间表征。基于流的声学模型是研究的热点,比起基于概率密度蒸馏的声学模型,基于流的声学模型结构更加简单,模型也更加稳定。

3 语音合成系统

3.1 传统和深度学习结合语音合成

将深度学习引入到经典语音合成各个模块中建模是一种有效的语音合成方法,Deep Voice-1是这种方法的代表。Deep Voice-1合成速度快,合成语音质量也较高。但系统由多个模块组成,容易产生累积误差且误差难以定位。

Deep Voice-2在Deep Voice-1基础上改进了多说话人语音合成系统,方法是将说话人矢量引入到模型中训练。目前直接将深度学习引入到经典语音合成系统各个模块中建模的研究已经不多了,Deep Voice系列的第三代Deep Voice-3已经转为采用端到端语音合成方法。

3.2 端到端语音合成系统

文献[40]是首个端到端语音合成系统,开启了端到端语音合成系统的研究。该系统需要一个预训练的隐马尔科夫对齐模块辅助声学模型学习文本与语音在时间上的对齐,其次该系统采用一些辅助的训练技巧会影响合成语音质量。

Char2Wav[54]有阅读器和声码器两部分。阅读器有编码器和解码器组成,编码器是一个双向的循环神经网络,解码器是基于注意力机制的循环神经网络,声码器是SampleRNN。

Tacotron-1语音合成系统声学模型是基于内容注意力机制的声学模型,声码器是Griffin-Lim算法,Griffin-Lim算法不是深度学习模型,文献[42]提出可以将Griffin-Lim算法替换为基于深度学习的声码器。Tacotron-2在Tacotron-1的基础上采用了Wavenet作为声码器。

Tacotron系列语音合成系统架构简单,合成语音质量高。WaveTTS[55]在Tacotron系列原有频域损失函数基础上引入时域的损失函数,这样可以提高合成语音的质量。针对Tacotron解码器采用的强制教学模式会造成曝光偏差问题,文献[56]提出用概率密度蒸馏的方法,训练教师解码器时采用强制教学方法,而训练学生解码器采用常规的自回归方式,通过优化学生解码器和教师解码器的输出误差提高学生解码器的准确性,从而减少曝光误差。文献[57]则采用对抗生成网络的方式避免曝光误差。

文献[58]使用Tacotron声学模型和Excitnet声码器组成端到端语音合成系统,该系统可以合成具有特定风格的语音。

Deep Voice-3是全卷积语音合成系统,声学模型可以生成多种中间表征,对应也有多种声码器。Deep Voice-3可以学习多说话人语音且训练时间短、合成语音速度快。

VoiceLoop[59]在文本到中间表征过程中在编码解码的基础上添加了固定大小的缓存机制,这样可以减少模型的复杂度,VoiceLoop声码器是WORLD[60]。

Clarinet是在Parallel Wavenet和Deep Voice-3基础上改进的可以高速训练和并行生成的完全意义上端到端的语音合成系统。Clarinet通过单个神经网络直接将文本转换为语音波形。Tacotron、Deep Voice-3等语音合成系统是先将文本转换为频谱,然后将频谱转换成语音。Clarinet完全从文本到原始语音波形的端到端训练,实现了对整个语音合成系统的联合优化。比起分别训练的模型,Clarinet在语音合成的自然度上有大幅提升。

文献[61]提出使用Transformer替代Tacotron-2中的注意力机制,这样编码器和解码器可以同时工作,提高了合成语音速度,基于Transform的自注意力机制更好地适应长序列依赖问题。

4 研究展望

声码器是语音合成技术的基础,声码器的发展历经了自回归式到并行式两个阶段。目前声码器已经可以满足不同场景的需求,但是高质量的声码器模型规模比较大,这意味着训练和部署模型需要较多的资源,而这些资源往往不是很充沛。所以如何在较少的资源上提高声码器的性能是研究的一个方向,目前很多研究是在声码器性能和模型复杂度之间做权衡。

端到端语音合成系统有漏发音和误发音的问题。这其中的一个原因是声学模型生成中间表征不稳定,特别是采用序列到序列加注意力机制的声学模型有曝光偏差等问题。如何解决声学模型结果不稳定是一个研究热点。一些解决的方法有将注意力机制由位置敏感改为相关位置敏感[62],采用蒸馏学习[56],增加时长控制和自适应优化策略[63]或者将改变注意力策略[64]以及应用额外的注意力机制辅助学习[65-66]等。

现阶段语音合成系统虽然可以实现高质量的语音合成,但是难以控制合成语音的风格:不同人会有不同风格,同一个人在不同状态下也会有不同的语音特点。目前常用的办法有预先从语音中提出可以表征韵律变化的隐变量,将隐变量代入语音合成模型中[67-69]训练、引入新的文本特征[70]或者引入句法、语义结构等信息[71]。

现有的语音合成系统都需要大量的语音文本数据,但现实中这些数据往往很难获得,所以如何能在较小的数据集上训练出高质量语音合成模型是一个研究方向。现有的方法有基于迁移学习[72]和半监督学习方法[73]等。

合成的语音仔细识别仍可以区分出自然语言和合成语音,这是因为合成语音过程中存在声学模型不准确、过度平滑以及声码器不准确等问题。所以如何能降低这些问题对合成语音的影响是研究的一个方向,文献[74-75]提出在合成语音过程中增加语音增强的模块,比如声学特征增强、语音增强等。

一句话中可能会包含多种语言,比如汉语和英语。这种跨语种的语音合成也是一个研究方向,一些做法是找到一种可以表示相关语种特征的矢量,然后将特征矢量嵌入语音合成模型[76]中训练。

5 结束语

基于深度学习的语音合成技术无论是在合成语音质量上还是速度上都取得了重大的进展。声码器是语音合成技术的基础,基于深度学习的声码器历经了自回归式和并行式两个阶段。自回归式可以生成高质量的语音,但速度慢,并行式可以快速生成语音,但存在着模型较大,训练过程比较复杂等问题。声学模型将文本转化成中间表征,和声码器相似,声学模型也经历了自回归式和并行式两个阶段。端到端语音合成系统将声学模型和声码器结合建模,直接建立起从文本到语音的合成,降低了语音合成研究对专业知识的要求,是目前研究较多的语音合成方法。

语音合成应用场景丰富,不同场景对语音合成的要求也不尽相同。满足在不同场景下的要求将是语音合成发展的一大考验。

参考文献:

[1] PURINGTON A,TAFT J G,SANNON S,et al.Alexa is my new bff:social roles,user satisfaction,and personification of the amazon echo[C]//Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems,2017:2853-2859.

[2] 张斌,全昌勤,任福继.语音合成方法和发展综述[J].小型微型计算机系统,2016,37(1):186-192.

[3] YIN Z.An overview of speech synthesis technology[C]// 2018 Eighth International Conference on Instrumentation & Measurement,Computer,Communication and Control (IMCCC),2020.

[4] NING Y,HE S,WU Z,et al.A review of deep learning based speech synthesis[J].Applied Sciences,2019,9(19):4050.

[5] OORD A V D,DIELEMAN S,ZEN H,et al.Wavenet:a generative model for raw audio[J].arXiv:1609.03499,2016.

[6] ARIK S O,CHRZANOWSKI M,COATES A,et al.Deep voice:real-time neural text-to-speech[C]//International Conference on Machine Learning,2017:195-204.

[7] ARIK S,DIAMOS G,GIBIANSKY A,et al.Deep voice 2:multi-speaker neural text-to-speech[C]//Neural Information Processing Systems,2017:2962-2970.

[8] HINTON G,VINYALS O,DEAN J.Distilling the knowledge in a neural network[J].Computer Science 2015,14(7):38-39.

[9] KINGMA D P,DHARIWAL P.Glow:generative flow with invertible 1x1 convolutions[C]//Neural Information Processing Systems,2018:10215-10224.

[10] KINGMA D P,SALIMANS T,JOZEFOWICZ R,et al. Improving variational inference with inverse autoregressive flow[C]//Neural Information Processing Systems, 2016:4743-4751.

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

[12] JUNYOUNG C,CAGLAR G,KYUNGHYUN C,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J].arXiv:1412.3555,2014.

[13] ADIGA N,TSIARAS V,STYLIANOU Y.On the use of wavenet as a statistical vocoder[C]//2018 IEEE International Conference on Acoustic,Speech and Signal Processing,2018:5674-5678.

[14] ITU-T.Recommendation G.711.Pulse code modulation (PCM) of voice frequencies[S/OL].1988.http://read.pudn.com/downloads110/ebook/456655/ITU-T-G711.pdf.

[15] TAMAMORI A,HAYASHI T,KOBAYASHI K,et al. Speaker-dependent wavenet vocoder[C]//Interspeech 2017, 2017:1118-1122.

[16] HAYASHI T,TAMAMORI A,KOBAYASHI K,et al. An investigation of multi-speaker training for wavenet vocoder[J].2017 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU),Okinawa,Japan, 2017:712-718.

[17] TACHIBANA K,TODA T,SHIGA Y et al.An investigation of noise shaping with perceptual weighting for wavenet-based speech generation[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Calgary,AB,Canada,2018:5664-5668.

[18] SONG E,BYUN K,KANG H G.Excitnet vocoder:a neural excitation model for parametric speech synthesis systems[C]//2019 27th European Signal Processing Conference(EUSIPCO),A Coruna,Spain,2019:1-5.

[19] LE P T,KHORRAMI P,CHANG S,et al.Fast wavenet generation algorithm[J].arXiv:1611.09482,2016.

[20] JIN Z,FINKELSTEIN A,MYSORE G J,et al.Fftnet:a real-time speaker-dependent neural vocoder[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2018:2251-2255.

[21] MEHRI S,KUMAR K,GULRAJANI I,et al.Samplernn: an unconditional end-to-end neural audio generation model[J].arXiv:1612.07837,2016.

[22] NAL K,ERICH E,KAREN S,et al.Efficient neural audio synthesis[C]//Proceedings of the 35th International Conference on Machine Learning,Stockholm,Sweden,2018.

[23] NARANG S,DIAMOS G,ELSEN F,et al.Exploring sparsity in recurrent neural networks[C]//International Conference on Learning Representations,2017.

[24] NARANG S,UNDERSANDER E,DIAMOS G,et al. Block-sparse recurrent neural networks[J].arXiv:1711.02782,2017.

[25] VALIN J,SKOGLUND J.Lpcnet:improving neural speech synthesis through linear prediction[C]//International Conference on Acoustics Speech and Signal Processing, 2019:5891-5895.

[26] 陈小东,宋文爱,刘晓峰.基于LPCNet的语音合成方法研究[J].计算机与数字工程,2020,48(5):1143-1147.

[27] DEN OORD A V,LI Y,BABUSCHKININ I,et al. Parallel wavenet:fast high-fidelity speech synthesis[C]// International Conference on Machine Learning,2018: 3915-3923.

[28] PING W,PENG K,CHEN J,et al.Clarinet:parallel wave generation in end-to-end text-to-speech[C]//International Conference on Learning Representations,2019.

[29] REZENDE D J,MOHAMED S.Variational inference

with normalizing flows[C]//International Conference on Machine Learning,2015:1530-1538.
[30] DINH L,KRUEGER D,BENGIO Y,et al.Nice:nonlinear independent components estimation[J].arXiv:1410.8516,2014.
[31] PRENGER R,VALLE R,CATANZARO B,et al.Waveglow:a flow-based generative network for speech synthesis[C]//International Conference on Acoustics Speech and Signal Processing,2019:3617-3621.
[32] DINH L,SOHLDICKSTEIN J,BENGIO S,et al.Density estimation using real nvp[J].arXiv:1605.08803,2016.
[33] KIM S,LEE S G,SONG J,et al.Flowavenet:a generative flow for raw audio[C]//International Conference on Machine Learning,2019:3370-3378.
[34] ARIK S Ö,JUN H,DIAMOS G.Fast spectrogram inversion using multi-head convolutional neural networks[J].IEEE Signal Processing Letters,2018,26(1):94-98.
[35] WANG X,TAKAKI S,YAMAGISHI J.Neural source-filter waveform models for statistical parametric speech synthesis[J].IEEE/ACM Transactions on Audio,Speech,and Language Processing,2019,28:402-415.
[36] VINYALS O,KAISER L,KOO T,et al.Grammar as a foreign language[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems,2015:2773-2781.
[37] CHOROWSKI J,BAHDANAU D,SERDYUK D,et al.Attention-based models for speech recognition[C]//Neural Information Processing Systems,2015:577-585.
[38] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:5998-6008.
[39] SUTSKEVER I,VINYALS O,LE Q V,et al.Sequence to sequence learning with neural networks[C]//Neural Information Processing Systems,2014:3104-3112.
[40] WANG W,SHUANG X,BO X.First step towards end-to-end parametric tts synthesis:generating spectral parameters with neural attention[C]//Proceedings Interspeech,2016:2243-2247.
[41] BAHDANAU D,CHO K,BENGIO Y,et al.Neural machine translation by jointly learning to align and translate[J].arXiv:1409.0473v7,2014.
[42] WANG Y,SKERRY-RAN R,STANTON D,et al.Tacotron:towards end-to-end speech synthesis[C]//Conference of the International Speech Communication Association,2017:4006-4010.
[43] 张小峰,谢钧,罗健欣,等.深度学习语音合成技术研究[J].计算机时代,2020(9).
[44] LEE J,CHO K,HOFMANN T.Fully character-level neural machine translation without explicit segmentation[J].Transactions of the Association for Computational Linguistics,2017,5:365-378.
[45] SHEN J,PANG R,WEISS R,et al.Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//International Conference on Acoustics,Speech,and Signal processing,2018:4779-4783.
[46] PING W,PENG K,GIBIANSKY A,et al.Deep voice 3:scaling text-to-speech with convolutional sequence learning[J].arXiv:1710.07654v3,2017.
[47] PING W,PENG K,GIBIANSKY A,et al.Deep voice 3:2000-speaker neural text-to-Speech.[C]//International Conference on Learning Representations,2018.
[48] RAFFEL C,LUONG M T,LIU P J,et al.Online and linear-time attention by enforcing monotonic alignments[C]//International Conference on Machine Learning,2017:2837-2846.
[49] SCHMIDT F.Generalization in generation:a closer look at exposure bias[C]//Proceedings of the 3rd Workshop on Neural Generation and Translation,2019.
[50] REN Y,RUAN Y,TAN X,et al.Fastspeech:fast,robust and controllable text to speech[C]//Neural Information Processing Systems,2019:3171-3180.
[51] GEHRING J,AULI M.Convolutional sequence to sequence learning[C]//International Conference on Machine Learning,2017:1243-1252.
[52] PENG K,PING W,SONG Z,et al.Parallel neural text-to-speech[J].arXiv:1905.08459v1,2019.
[53] KIM J,KIM S,KONG J,et al.Glow-tts:a generative flow for text-to-speech via monotonic alignment search[J]arXiv:2005.11129,2020.
[54] SOTELO J,MEHRI S,KUMAR K,et al.Char2wav:end-to-end speech synthesis[J].arXiv:1703.10135,2017.
[55] LIU R,SISMAN B,BAO F,et al.Wavetts:tacotron-based tts with joint time-frequency domain loss[J].arXiv:2002.00417v1,2020.
[56] LIU R,SISMAN B,LI J,et al.Teacher-student training for robust tacotron-based tts[J].arXiv:1911.02839,2019.
[57] GUO H,SOONG F K,HE L,et al.A new gan-based end-to-end tts training algorithm[J].arXiv:1904.04775,2019.
[58] KWON O,SONG E,KIM J M,et al.Effective parameter estimation methods for an excitnet model in generative text to speech systems[J].arXiv:1905.08486,2019.
[59] YAIGMAN Y,WOLF L,POLYAK A,et al.Voiceloop:voice fitting and synthesis via a phonological loop[J].arXiv:1707.06588,2017.
[60] MORISE M,YOKOMORI F,OZAWA K,et al.World:a vocoder-based high-quality speech synthesis system for real-time applications[J].IEICE Transactions on

Information and Systems,2016,99(7):1877-1884.

[61] LI N,LIU S,LIU Y,et al.Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:6706-6713.

[62] BATTENBERG E,SKERRY-RYAN R J,MARIOORYAD S,et al.Location-relative attention mechanisms for robust long-form speech synthesis[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:6194-6198.

[63] FU R,WEN Z,YI J,et al.Focusing on attention:prosody transfer and adaptative optimization strategy for multi-speaker end-to-end speech synthesis[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Barcelona,Spain,2020:6709-6713.

[64] YASUDA Y,WANG X,YAMAGISHI J.Effect of choice of probability distribution,randomness,and search methods for alignment modeling in sequence-to-sequence text-to-speech synthesis using hard alignment[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:6724-6728.

[65] HE M,DENG Y,HE L.Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts[C]//Proceedings of Interspeech 2019,2019:1293-1297.

[66] ZHU X,ZHANG Y,YANG S,et al.Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis[J].IEEE Access,2019,7:65955-65964.

[67] SUN G,ZHANG Y,WESISS R J,et al.Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and auto-regressive prosody prior[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:6699-6703.

[68] UM S Y,OH S,BYUN K,et al.Emotional speech synthesis with rich and granularized control[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:7254-7258.

[69] VALLE R,LI J,PRENGER R,et al.Mellotron:multi-speaker expressive voice synthesis by conditioning on rhythm,pitch and global style tokens[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:6189-6193.

[70] XIAO Y,HE L,MING H,et al.Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:6704-6708.

[71] SUN A,WANG J,CHENG N,et al.Graphtts:graph-to-sequence modelling in neural text-to-speech[J].arXiv:2003.01924,2020.

[72] JIA Y,ZHANG Y,WEISS R,et al.Transfer learning from speaker verification to multi-speaker text-to-speech synthesis[C]//Neural Information Processing Systems,2018.

[73] INOUE K,HARA S,ABE M,et al.Semi-supervised speaker adaptation for end-to-end speech synthesis with pretrained models[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:7634-7638.

[74] TANAKA K,KANEKO T,HOJO N,et al.Wavecyclegan:Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks[C]//2018 26th European Signal Processing Conference,2018.

[75] TANAKA K,KAMEOKA H,KANEKO T,et al.Wavecyclegan2:time-domain neural post-filter for speech waveform generation[J].arXiv:1904.02892v1,2019.

[76] ZHOU X,TIAN X,LEE G,et al.End-to-end code-switching tts with cross-lingual language model[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing,2020:7614-7618.