doi:10.3969/j.issn.1001-4616.2025.05.015

一种基于非自回归模型的文本转语音方法

郭璐璐,高　尚

(江苏科技大学计算机学院,江苏 镇江 212100)

[**摘要**]　文本转语音(Text-to-Speech,TTS)是一种将给定文本合成为语音的技术,具有广泛的应用前景. 相比于自回归的 TTS 模型,非自回归的 TTS 模型在语音合成速度上有显著提升. 然而,非自回归模型在长序列的语音合成任务中其合成速度和语音质量仍有提升空间. 为此,本文提出了一种基于非自回归的 EnhanceSpeech 模型. 首先,该模型利用可学习的外部记忆向量简化注意力机制计算方式,有效减少了计算复杂度和内存占用,并提升了模型的推理速度. 其次,通过引入基于分层挤压注意力的后处理网络,利用二维卷积将梅尔频谱图生成过程视为图像处理,显著提升了梅尔频谱图的生成质量. 实验结果表明,EnhanceSpeech 模型与自回归模型相比生成速度提高了 60 倍以上. 此外,与同类非自回归模型相比,本文方法的性能突出,更接近领先的自回归模型水平.

[**关键词**]　语音合成,自回归模型,非自回归模型,注意力机制,后处理网络
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2025)05-0129-10

A Text-to-Speech Method Based on Non-Autoregressive Model

Guo Lulu,Gao Shang

(School of Computer,Jiangsu University of Science and Technology,Zhenjiang 212100,China)

Abstract:Text-to-Speech (TTS) is a technology that synthesizes given text into speech and has a wide range of application prospects. Compared with the autoregressive TTS model,the non-autoregressive TTS model has significantly improved the speech synthesis speed. However,there is still room for improvement in the synthesis speed and speech quality of non-autoregressive models in long-sequence speech synthesis tasks. To this end,an EnhanceSpeech model based on non-autoregression is proposed. First,the model uses learnable external memory vectors to simplify the calculation of the attention mechanism,effectively reducing computational complexity and memory usage,and improving the model's inference speed. Secondly,by introducing a post-processing network based on hierarchical squeeze attention and using two-dimensional convolution to treat the mel-spectrogram generation process as image processing,the quality of mel-spectrogram generation is significantly improved. Experimental results reveal that the EnhanceSpeech model is over 60 times faster than its autoregressive counterparts. Moreover,it outperforms other non-autoregressive methods,bringing its performance closer to that of top-tier autoregressive models.

Key words:speech synthesis,autoregressive model,non-autoregressive model,attention mechanisms,post-processing network

文本转语音(Text-to-Speech,TTS)技术正逐渐成为人们生活中不可或缺的一部分,其目标是将文本合成为可理解且自然的语音[1-4]. 传统的 TTS 系统依赖规则和语音片段拼接,而基于深度学习的方法能够学习复杂的文本与语音映射关系,生成更为自然流畅的语音输出[5]. 目前,主流的深度神经网络声学建模方法采用端到端的(sequence-to-sequence,Seq2Seq)模型,将文本的语言学特征序列直接转换为声学特征序列,再通过神经声码器生成音频波形[6-9]. 然而,传统 Seq2Seq 模型的解码器在样本个性化处理能力上存在局限性,从而限制了解码效果. 近年来,诸如 Tacotron 和 Transformer TTS 等自回归声学模型按顺序生成声学特征,但在推理速度和语音控制性方面存在一定局限性[10-12]. 相较之下,FastSpeech2 采用非自回归生成策略,以音素序列作为输入,提升了推理速度. 然而,由于语音合成涉及序列生成,且输出序列通常较长,因此仍需提升声学特征的生成效率,并解决声谱图中可能存在的噪音问题,以进一步提高语音合成的质量[13-15].

收稿日期:2024-09-09.
基金项目:国家自然科学基金项目(62376109).
通讯作者:高尚,博士,教授,研究方向:模式识别. E-mail:gao_shang@ just.edu.cn

为解决上述问题,本文提出了 EnhanceSpeech 模型. 通过引入外部编码器和解码器以及可学习的外部注意力机制(external attention,EAttention)来改进语音合成效果. 外部注意力允许对文本中不同元素的查询向量与可学习的键和值记忆向量进行计算,这些记忆向量独立于单个样本并在整个文本元素之间共享,从而通过一个中介促进元素之间的交互,提升模型的语音合成速度. 此外,本文借鉴了 Tacotron 模型的思想,在解码器后引入了基于分层挤压注意力的后处理网络(layered squeeze attention-based postnet,LSA-Postnet),用于修正和增强生成的声谱图,以生成更为清晰和自然的语音输出.

1 相关概念

1.1 文本转语音

文本转语音(TTS)技术是人工智能领域的热门研究课题,旨在生成自然且易于理解的语音[1,3,7]. 该技术的发展经历了多个阶段,从早期的拼接合成和统计参数合成,逐渐过渡到基于神经网络的端到端模型. 这些技术首先将文本转换为声学特征(如梅尔频谱图),然后再生成音频波形[16-18]. 然而,采用自回归方式生成梅尔频谱图(mel-spectrogram,Mel)将导致推理速度较慢. FastSpeech 1 和 FastSpeech 2 模型通过非自回归方式生成声谱图[19-20],显著提高了文本转语音的效率,但合成的语音中仍存在噪音问题. 本文提出的 EnhanceSpeech 模型不仅提升了生成效率,还通过优化特征处理和后处理网络,实现了更自然流畅的语音合成效果.

1.2 非自回归序列生成

传统的 TTS 系统通常采用自回归模型生成声学特征(如梅尔频谱图),每个时间步的生成依赖于前面的结果[21]. 虽然自回归模型能产生高质量语音,但推理速度慢且容易累积错误,导致语音不连贯或重复. 相比之下,非自回归模型可以并行生成序列,显著提升推理速度[22]. EnhanceSpeech 作为一种非自回归模型,进行了两方面的改进:首先,通过减少自注意力机制的参数量,在保证生成音频质量的前提下,实现了轻量级模型的推理. 其次,针对复杂语法结构或生僻词汇导致的生成质量问题,EnhanceSpeech 引入了基于分层挤压注意力的后处理网络对声学特征进行降噪和平滑处理,提高语音的整体质量和自然度.

1.3 TTS 任务中的注意力机制

注意力机制用于加权输入文本的不同部分,以便模型在生成音频帧时能更有效地关注相关文本信息[6]. TransformerTTS 模型中,通过将声学特征和语言学特征进行加权融合,获得音频帧的上下文表示[20]. FastSpeech 1/2 的 Transformer 前馈模块结合了自注意力网络和一维卷积网络,自注意力网络利用多头注意力提取交叉位置信息[16,21]. 本文引入的外部注意力机制计算自注意力和较小的可学习记忆向量之间的关系,这些记忆向量捕获文本的全局上下文. 外部注意力不依赖于语义信息,可通过端到端的反向传播算法进行优化,无需迭代算法. 该注意力机制在降低计算成本的同时,取得了与原始自注意力及其一些变体相当或更好的效果.

1.4 后处理网络

后处理网络在语音合成中起着关键作用,通过多种技术手段调整和优化声学特征或音频,提升生成语音的质量、清晰度和自然度[3,22-23]. 这些技术包括降噪、去混响、频谱平滑、音量调节和音色优化等[24-25]. 后处理网络综合应用这些技术,显著提升语音合成系统的性能和用户体验,使生成的语音更真实、流畅和易懂. Tacotron[4] 和 TransformerTTS[20] 中的后处理网络由一维卷积堆叠组成,通过残差连接将输出叠加到原始声谱图上,生成最终的声谱图,有效降低了声谱图中的噪音[26-27]. 本文采用基于分层挤压注意力的后处理网络,以二维卷积方式对生成的声谱图进行优化,以可学习的方式增强和降噪,从而显著改善语音的质量和自然度.

2 研究方法

2.1 概述

本文在 FastSpeech2 模型的基础上,提出了一种新的语音合成模型——EnhanceSpeech,如图 1 所示. 该模型通过引入创新的可学习外部注意力机制和基于分层挤压注意力的后处理网络,对声谱图进行精细化修正和增强,旨在显著提升语音输出的清晰度和自然度. 具体地,可学习的外部注意力机制被应用于计

算文本元素的查询向量与可学习的键值记忆向量之间的关联. 这些记忆向量在整个文本元素范围内共享,并通过中介方式促进元素之间的交互,从而提升了模型的生成效率. 随后,基于分层挤压注意力机制的后处理网络对解码器生成的声谱图进行降噪和特征增强,进一步提升了语音合成的质量.

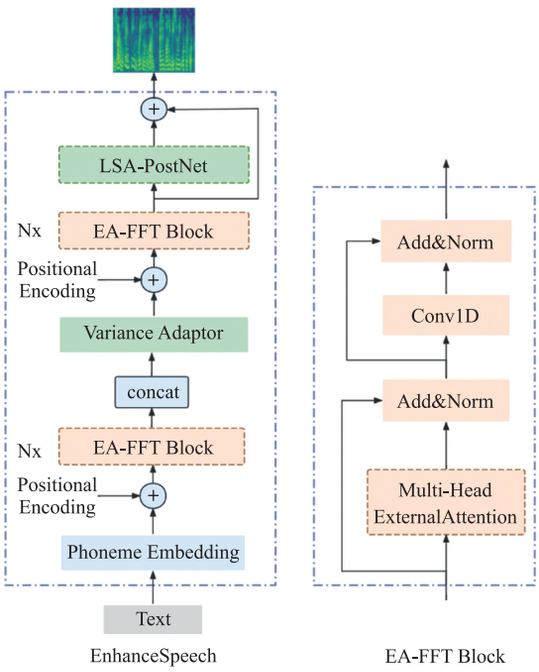


图 1　EnhanceSpeech 模型的总体框架

Fig. 1　General framework of the EnhanceSpeech model

2.2　可学习的外部注意力机制

在先前的研究中,编码器和解码器的自注意力机制仅计算输入序列内部元素之间的关系,而忽略了不同输入序列间潜在的共享关系. 这种限制可能导致在处理多个文本的语言学特征时,错过重要的共享特征或模式,从而缺乏泛化性. 因此,本文采用可学习的外部注意力机制(external attention,EAttention)来促进样本间的信息交互,同时降低注意力计算的复杂度,以提升文本嵌入表示的质量和效果. 具体地,该外部注意力机制使用两个可学习的外部记忆向量 M_k 和 M_v 来替代自注意机制中输入向量线性映射得到的键和值向量. 其中,$M_k \in R^{d\times m}$ 和 $M_v \in R^{m\times d}$ 是与输入无关的可学习参数,在整个训练集中共享. 由于编码器和解码器中的注意力处理步骤一致,本文选择使用编码器中的外部注意力为例进行介绍. 以文本序列 $x=\{x_1,x_2,\cdots,x_N\}$ 为例,通过 Embedding 层和位置编码得到高维嵌入向量 $H \in R^{N\times d}$,d 代表的是每个字符的隐藏向量维度,只需要一次线性映射产生查询向量 Q,如式(1)所示:

$$Q=HW^Q. \tag{1}$$

然后,将查询向量与 M_k 的转置,即 M_k^{T} 进行相似度计算得到注意力权重 $A \in R^{N\times m}$,如式(2)所示:

$$A=\mathrm{Norm}(\mathrm{softmax}(QM_k^{\mathrm{T}})). \tag{2}$$

最后,将注意力权重与值记忆向量进行点积运算得到注意力向量 $O \in R^{N\times d}$,如式(3)所示:

$$O=AM_v. \tag{3}$$

可学习记忆向量旨在从整个数据集中提取鉴别力最强的特征,捕捉信息量最大部分,并排除其他样本的干扰. 不同于传统注意力机制,该机制不试图重建输入特征,也不应用稀疏正则化,而是通过可学习权重自适应地动态调整样本间的关注度.

2.3　基于分层挤压注意力的后处理网络

在语音合成领域,声学模型解码后将多个时间步的声学特征拼接成完整的声谱图,这直接影响最终生成语音的质量. 本文采用二维卷积技术对声谱图进行处理,充分利用了其空间特征,深入挖掘局部模式和结构信息. 与传统的一维卷积方法相比,这种新方法在降低噪声的同时,更有效地保留了声音细节,展现出更优越的性能. 为了优化解码器生成的声谱图,本文提出了一种基于分层挤压注意力机制的后处理网

络. 该网络结构如图 2 所示,首先,通过挤压和连接模块(squeeze and concat module,SC module)获得声谱图的层级多尺度特征图. 然后,通过压缩与激励权重模块(squeeze and excitation weight model,SEWeight module)提取不同层级特征图的注意力,得到层级注意力向量. 接下来,将逐层乘积操作应用于注意力向量和多层特征图. 最后,将多层声谱图信息与原始声谱图信息进行残差连接,得到最终输出. 下文是 SC module 和 SEWeight model 的具体实现细节.

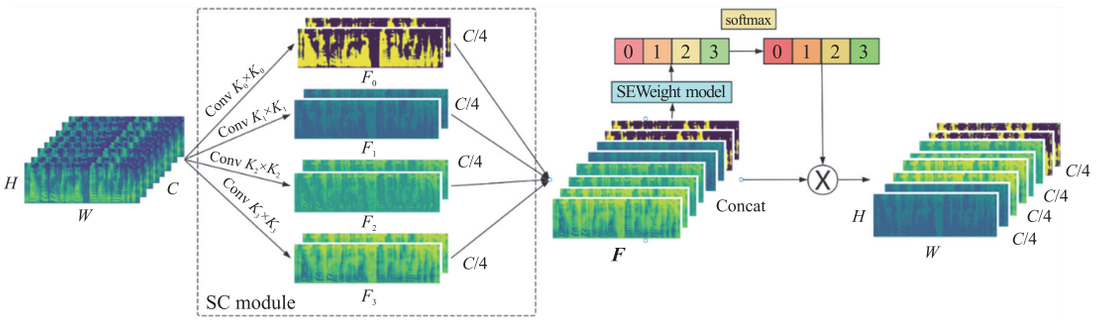


图 2　基于分层挤压注意力的后处理网络

Fig. 2　Post-processing network based on hierarchical squeezed attention

2.3.1　SC module

为了增强声学特征在不同层次上的表达能力,本文采用多通道的方法生成声谱图表示. 对于尺寸为 $H\times W$ 的声谱图 X,通过设置通道数 C,将其扩展为 $H\times W\times C$ 的高维张量表示. 这种升维操作允许更深入地分析不同层次的声学特征. 每层的特征图都具有相同的通道数,共有 S 层. 每一层的特征图都具有公共通道维数 $C'=C/S$. 值得注意的是,层数 C 应当可以被 S 整除. 在每层中,模型独立学习不同尺度的空间信息,并通过局部交互方式建立通道之间的联系. 为有效处理不同核尺度的输入张量而不增加计算成本,本文引入组卷积方法,利用硬件并行计算能力提高计算效率. 多尺度核与组大小的关系如式(4)所示:

$$G=2^{\frac{K-1}{2}}, \tag{4}$$

式中,K 表示的是卷积核大小,G 是组大小. 对于 $i=0,1,\cdots S-1$ 的 S 层的卷积层,所对应的多尺度特征图生成函数如式(5)所示:

$$F_i=\mathrm{Conv}(k_i\times k_i,G_i)(X), \tag{5}$$

式中,第 i 层的卷积核大小 k_i 可由公式 $k_i=2\times(i+1)+1$ 计算得到. 同时,每个尺度的组大小 G_i,通过式(6)计算,这导致组大小随着 k_i 的增加呈指数级增长. $F_i\in R^{H\times W\times C'}$ 代表该尺度下的特征图. 随着层数的增加,卷积核和组大小逐渐扩展,有助于模型在不同尺度上有效捕捉特征. 通过设定不同的尺度大小,模型能够探测到多种尺寸的目标. 最终,将所有层的不同尺度特征图连接起来,形成整体特征图.

$$F=\mathrm{Concat}([F_0,F_1,\cdots,F_{S-1}]), \tag{6}$$

式中,$F\in R^{H\times W\times C}$ 表示捕捉到不同目标的多尺度特征图. 通过这种多层表示,后处理网络能够更精确地定位并处理声谱图中的噪音和重要信息,从而进一步提高语音合成的质量.

2.3.2　SEWeight model

通道注意机制允许网络根据需要加权每个通道的重要性,从而产生更丰富的输出信息. SC module 融合了多种层次空间信息的不同目标尺度特征图. SEWeight 模块通过挤压和计算注意力权重来编码全局信息和自适应调整通道关系,具体结构如图 3 所示:

首先,对多尺度特征图 F 使用式(7)进行自适应二维平均池化.

$$g_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F_c(i,j). \tag{7}$$

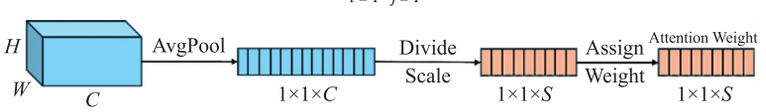


图 3　SEWeight 模型结构

Fig. 3　The structure of SEWeight model

最后得到的 $g_c \in R^{1\times1\times C}$ 代表不同通道的特征图的整体表示. 由于在 SCModule 中,相同尺度下的特征图关注的细节特征类似,因此本文将每个尺度视为权重分配的单位. 在这里,本文进一步将 g_c 分为 S 个空间:

$$h_i=\frac{1}{C'}\sum_{j=0}^{C'-1} g_c[i*C'+j]. \tag{8}$$

整个多尺度通道特征向量以串联方式得到:

$$h_c=h_0\oplus h_1\oplus\cdots\oplus h_{S-1}, \tag{9}$$

式中,$\oplus$ 是连接操作符,用于将结果连接起来,得到 $h_c \in R^{1\times1\times S}$,表示通过挤压操作得到不同尺度的特征向量表示. 在 SE 块中,第 c 个通道的注意力权重计算步骤如式(10)所示:

$$w_c=\sigma(W_1\delta(W_0(h_c))), \tag{10}$$

式中,W_0、W_1 分别代表两个全连接层,它们通过有效组合通道间的线性信息,促进高通道和低通道维度信息的交互. 其中,δ 表示 ReLU 激活函数,而 σ 则代表 Sigmoid 激活函数. 通过这些激活函数,实现了通道间交互后的权重分配,从而更有效地提取信息. 通过跨通道的软注意机制,自适应地选择不同的空间尺度,其软分配权重如式(11)所示:

$$att_c=\mathrm{Softmax}(w_c), \tag{11}$$

式中,$att_c \in R^{1\times1\times S}$ 表示注意力交互后的多尺度软分配权重. 然后,将这些多尺度软分配权重与相应尺度的 F 相乘,如式(12)所示:

$$Y_i=F_i\odot att_i, i=0,1,\cdots,S-1, \tag{12}$$

式中,Y_i 表示通过多尺度通道注意力权重后的特征图. 这使得声谱图能够根据不同尺度以可学习的方式增强重要特征,同时弱化噪音信息. 为了保证原有解码后声谱图的信息完整性,将获得注意力权重后的特征图与原始声谱图进行残差相加,如式(13)所示:

$$Mel=X+Y, \tag{13}$$

式中,X 表示原始声谱图,Y 表示经过增强重要特征后的声谱图.

3 实验验证

3.1 数据集与评价指标

3.1.1 数据集

为验证所提方法有效性,本文使用了常用 LJSpeech[28] 和 CSMSC[29] 数据集评估试验. 其中,LJSpeech 包含 13 100 个音频片段及相应的文本记录,而 CSMSC 包含 10 000 个音频片段及相应的文本标注. 根据 FastSpeech 2[21] 的方式,本文将两个数据集各自划分为训练集、验证集和测试集. 具体地,94% 的音频切片被用作训练集,3% 的音频切片被用作验证集,剩余的 3% 的音频切片被用作测试集. 音频数据的采样率为 22 050 Hz,转换后的特征表示为 80 维的梅尔频谱图,帧大小为 1 024.

针对数据集中的英文字符,本文使用 G2P[30] 工具将原始的英文字符序列转换为对应的音素序列. 针对数据集中的中文字符,使用 G2PC[31] 工具将所有的文本标注从中文字符形式转换为汉语拼音形式.

3.1.2 评价指标

本文评估了模型在 MOS[32]、RTF[33]、Peak Mem[34] 和 Params[34] 四个指标上的性能. 具体地,MOS 用于评估音频质量,分数越高表示质量越好. RTF 表示实际处理时间与语音长度之比,用于衡量模型的推理效率. RTF 值越小,则代表模型合成声谱图的速度越快. Peak Mem 是峰值内存,描述模型在运行过程中所需的最大内存量. 而 Params 是模型参数量,用于衡量模型大小.

3.1.3 对比方法

为验证本文所提方法的有效性,实验选择比较的模型包括(1)真实音频(ground truth,GT):作为基准的真实音频样本;(2)GT(Mel+HiFi-GAN[35]),即首先将真实音频转换为梅尔频谱图,然后使用 HiFi-GAN 将梅尔频谱图转换回音频;(3)Tacotron 2[4]:一种常用的基于端到端训练的自回归语音合成模型,常作为基准模型;(4)TransformTTS[20]:使用 Transformer 架构进行语音合成的自回归模型;(5)FastSpeech[16]:一种非自回归的语音合成模型,旨在提高语音合成的速度;(6)FastSpeech2[21]:FastSpeech 的改进版本,去除了师生架构,使用真实语音作为标签进行训练;(7)Glow-TTS[36]:基于 Glow 生成的非自回归模型的语音合

成系统. 其中 GT(真实音频)和 GT(Mel+HiFi-GAN)没有使用模型,因此只测量 MOS 指标.

3.1.4 实施细节

本文使用 2 块 NVIDIA RTX 3060Ti GPU 进行训练,每个模型的训练批次大小为 64,并采用相同的 Adam 优化器,设定参数为 $\beta1=0.9$,$\beta2=0.98$,$\varepsilon=10-9$. 所有模型在 LJSpeech 数据集上进行了 50×100 次迭代训练,已达到收敛. 随后,模型生成的梅尔频谱图经过预训练的 HiFi-GAN 声码器转换为音频样本,并在测试集上评估. 对于 EnhanceSpeech 模型,EAttention 中注意力单元的隐藏层维度设置为 64. 在 LSA-Postnet 模块中,channel 设置为 16,reduction 设置为 4,多尺度数量 S 设置为 4.

3.2 实验结果

如表 1 所示,本文将 EnhanceSpeech 模型与当前主流模型进行比较测试,评估其在 MOS、RTF、Peak Mem 及 Params 上的性能. 首先,在 MOS 中,EnhanceSpeech 在两个数据集上的性能分别比非自回归模型提升了 1%~3.4%和 1%~1.8%,与自回归模型的差距控制在 2.6%以内. 在推理效率方面,EnhanceSpeech 表现卓越. 其 RTF 值降低至 1.99×10^{-3},意味着音频生成速度获得了巨大提升. 与自回归模型相比,其速度提升了约 58 至 481 倍;与主流非自回归模型相比,也展现出领先的速度优势. 然后,在 Peak Mem 方面,所提模型仅高于 Tacotron2 模型,但与其他主流的非自回归模型相比降低了 5%~12.1%,与 TransformerTTS 相比降低了 7.5%. 最后,在 Params 上,EnhanceSpeech 虽然使用了后处理网络来增强声谱图的质量,其参数量依然较小. 这表明本文所提方法在性能和参数量之间能够实现良好的平衡. 综上所述,EnhanceSpeech 在音频质量、推理速度、内存占用和模型参数指标上整体表现较优,表明了模型结构的合理性,展示了其作为高效非自回归语音合成模型的潜力.

表 1 实验中不同模型生成音频的各指标比较

Table 1 Comparison of performance metrics for different models on audio generation

Method	MOS		RTF(10^{-3})	Peak Mem	Params
	LJSpeech	CSMSC			
GT	4.465	4.543			
GT(Mel+HiFi-GAN)	4.315	4.421			
Tacotron2	3.825	3.924	118.00	61.80	28.2
TransformerTTS	3.845	3.915	960.00	119.36	24.2
FastSpeech	3.675	3.774	2.01	116.19	23.5
FastSpeech2	3.775	3.776	2.03	125.69	27.0
Glow-TTS	3.745	3.767	17.20	116.59	28.6
EnhanceSpeech	3.801	3.835	1.99	110.40	26.2

注:RTF 表示实时因子,系统所需秒(连同 Hifi-GAN 声码器)来合成 1 s 音频.

3.3 方法有效性讨论

以 TTS 中常用的 LJSpeech 数据集为例,下文讨论了所提方法的有效性,包括模块有效性分析、外部注意力机制分析、后处理网络分析、LOSS 分析以及超参数分析.

3.3.1 模块有效性分析

如表 2 所示,本节进行消融研究以验证 EnhanceSpeech 设计的合理性. 其中,-EA 表示去除外部注意力,仅保留其他改进部分;-LSA-Postnet 表示去除后处理网络,仅保留其他改进部分;-EA+LSA-Postnet 表示原始的 FastSpeech2 模型.

表 2 模型中各模块影响分析

Table 2 The impact analysis of each module in model

Method	MOS	RTF(10^{-3})	Peak Mem	Params
EnhanceSpeech	0.000	0.000	0.0	0.0
-EA	-0.012	0.220	25.4	1.0
-LSA-Postnet	-0.023	-0.100	-14.3	-0.3
-EA+LSA-Postnet	0.026	0.090	14.4	0.8

-EA 模型去除了外部注意力,结果显示 MOS 评分有所下降,RTF 降低,Peak Mem 占用显著增加,Params 增加. 这表明外部注意力在提高音频质量和减少内存占用方面发挥了重要作用.-LSA-Postnet 模型

去除了后处理网络,MOS 评分略有下降,但 RTF 提升,Peak Mem 占用显著降低,Params 减少. 这表明后处理网络对音频质量有提升作用,但在推理效率和内存占用上付出了一定的代价.

-EA+LSA-Postnet 模型即原始 FastSpeech 2 模型,在所有指标上表现都不如 EnhanceSpeech,尤其在 RTF 和 Peak Mem 方面. 这验证了外部注意力和后处理网络的结合显著提升了模型整体性能.

综上,EnhanceSpeech 模型通过整合外部注意力和基于分层挤压注意力的后处理网络,在保持低参数量和内存占用的同时,显著提升了音频质量和推理效率,展现了其在语音合成中的优势.

3.3.2 外部注意力机制的有效性分析

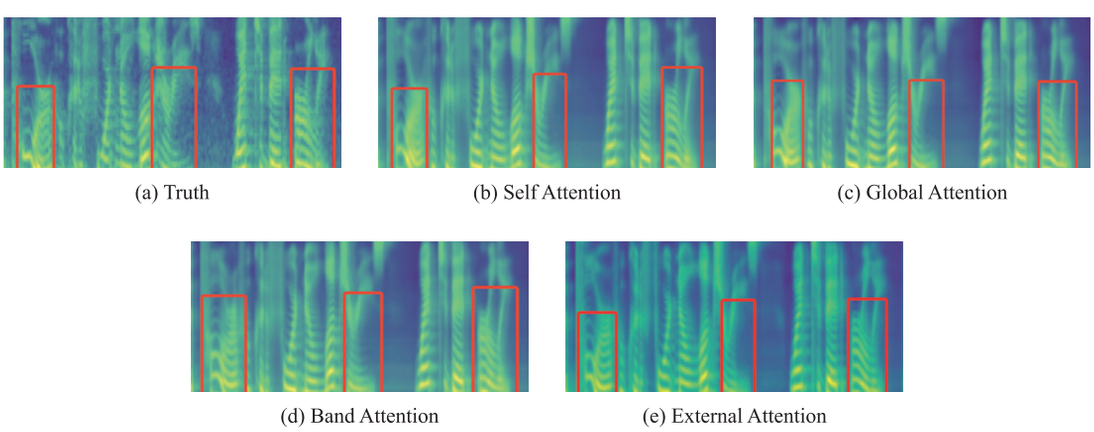
为了研究可学习的外部注意力机制在捕捉元素关联性方面的有效性,表 3 展示了不同注意力机制下编码器和解码器的性能. 本文比较了自注意力(Self Attention[37])与两种静态稀疏注意力机制(Global Attention[38]和 Band Attention[39]). 结果显示,外部注意力机制表现最佳. 与自注意力机制相比,外部注意力机制在 MOS 上提升了 0.8%,RTF 提升了 15.5%,Peak Mem 占用降低了 22.5%,模型 Params 降低了 4.1%. 这表明了自注意力机制构成的编码器和解码器可能包含与任务无关的冗余信息,难以有效捕捉关键信息. 此外,与 Global Attention 和 Band Attention 两种静态稀疏注意力相比,外部注意力机制在 MOS 上分别提升了 0.4%和 2.3%,RTF 分别提升了 4%和 6.6%,Peak Mem 占用分别降低了 23.8%和 21.9%,模型 Params 分别降低了 9.4%和 5.1%. 尽管静态方法能够减少元素之间的连接,但它们捕捉到的元素关联性是固定的,无法根据数据可学习地增强重要特征. 相比之下,本文提出的外部注意力在减少参数量的同时,可学习地构建元素之间的重要关联,极大地保留了自注意力机制的性能.

表 3 使用不同注意机制的验证结果

Table 3 Verification results using different attention mechanisms

Methods	MOS	RTF(10^{-3})	Peak Mem	Params
Self Attention	3.775	2.00	124.8	27.0
Global Attention	3.763	1.76	126.3	28.6
Band Attention	3.694	1.81	123.7	27.3
External Attention	3.778	1.69	96.1	25.9

为了可视化不同注意力机制生成的声谱图,本文对四种注意力机制生成的音频进行了主观评测,并选取了主观评测中 MOS 分数差异显著的音频样本进行声谱图分析. 具体如图 4 所示.



(a) Truth (b) Self Attention (c) Global Attention

(d) Band Attention (e) External Attention

图 4 不同注意力机制下的声谱图对比分析

Fig. 4 Comparative analysis of spectrograms under different attention mechanisms

对比分析不同注意力机制生成的声谱图发现,两种稀疏注意力机制生成的声谱图存在模糊部分. 当将自注意力替换为外部注意力后,声谱图特征并未发生明显变化,但其能降低了计算复杂度,加快了特征处理速度.

3.3.3 后处理网络的有效性分析

后处理网络在语音合成中提升生成语音的自然度和清晰度. 为验证基于分层挤压注意力的后处理网络的有效性,本文将外部注意力(EA)、一维卷积后处理网络(Conv1d-Postnet)、以及基于分层挤压注意力的后处理网络(LSA-Postnet)之间不同组合进行比较. 如表 4 所示,相比于 EA-Conv1d-Postnet,EA+LSA-Postnett 在

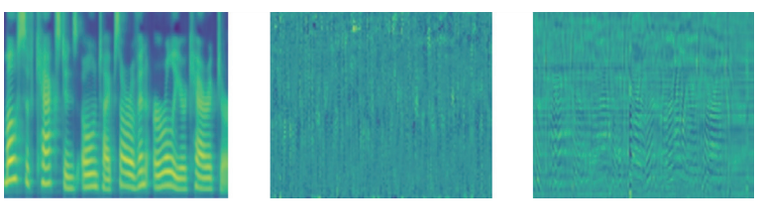
MOS 和 RTF 上得到了一定的提升,在模型 Params 上变化不大,仅在 Peak Mem 占用上略有增加. 综上所述,本文所提的后处理网络整体上表现较好,其不仅有效捕捉多尺度特征,同时提高了运行效率.

表 4　不同后处理网络对于性能的影响

Table 4　The effect of model performance based on two post-processing networks

Methods	MOS	RTF(10^{-3})	Peak Mem	Params
EA+Conv1d-Postnet	3.792	2.38	105.60	26.1
EA+LSA-Postnet	3.801	1.91	110.40	26.2

如图 5 所示,为直观展示后处理网络对声谱图的优化效果,本文对比了真实梅尔频谱图、使用一维卷积堆叠的后处理网络(Conv1d-Postnet)生成的梅尔频谱图,以及基于分层挤压注意力的后处理网络(LSA-Postnet)生成的梅尔频谱图.



(a) GT　　(b) Conv1d-Postnet　　(c) LSA-Postnet

图 5　两种后处理网络的声谱图对比分析

Fig. 5　Comparative analysis of spectrograms of two post-processing networks

图 5 中,Conv1d-Postnet 生成的梅尔频谱图通过平均池化处理噪音,但未能充分考虑波形的复杂特征,导致声谱图较为模糊. 相比之下,本文提出的 LSA-Postnet 通过分层挤压注意力机制,自适应地分配频谱图中特征的权重,显著增强关键语音特征,同时弱化噪音特征,比传统后处理网络更具自适应性.

3.3.4　Loss 分析

本文在 FastSpeech2 模型基础上提出了 EnhanceSpeech 模型,其拥有更好的拟合速度与更低的 loss 值. 如图 6 所示,EnhanceSpeech 模型在训练和验证阶段收敛速度高于 FastSpeech2 模型,这表明了本文提出的 EAttention 和 LSA-Postnet 两个模块能够加速模型对真实特征图的拟合和提升训练速度.

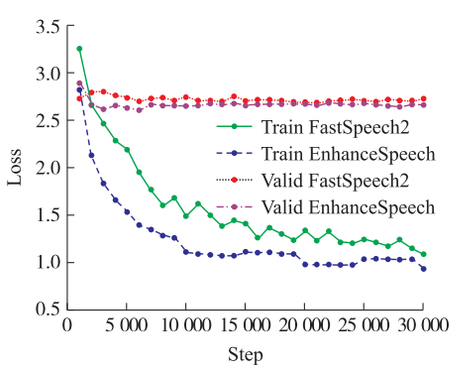


图 6　损失函数有效性验证

Fig. 6　Verification of the validity of the loss function

3.3.5　超参数分析

如表 5 和表 6 所示,本文验证了 EAttention 模块中的注意力单元隐藏层维度(S)和 LSA-PostNet 模块的通道数(channel)设置的合理性.

表 5　不同 S 值对模型性能的影响

Table 5　The impact of different S values on model performance

S	MOS	RTF(10^{-3})	Peak Mem	Params
32	3. 754	1. 21	105. 40	25. 2
64	3. 801	1. 91	110. 40	26. 2
128	3. 805	2. 23	122. 90	27. 1
256	3. 809	2. 49	129. 80	28. 2

表 6　不同 channel 值对模型性能的影响

Table 6　The impact of different channel values on model performance

channel	MOS	RTF(10^{-3})	Peak Mem	Params
8	3. 788	1. 03	104. 4	25. 8
16	3. 801	1. 91	110. 4	26. 2
32	3. 811	3. 91	133. 4	26. 8
64	3. 814	8. 97	180. 4	27. 1

表 5 中,固定 channel 为 16,随着 S 从 32 逐步增加到 128 和 256,模型在 MOS 上分别提升了 1.3%和 1.4%,RTF 分别降低了 84.2%和 105%,Peak Mem 占用分别增加了 16.6%和 23.1%,模型 Params 分别增加了 7.5%和 11.9%. 这表明增加注意力单元的维度有助于提升模型的性能,但也导致了更高的计算成本.

表 6 中,在固定 S 为 64 的情况下,将 channel 从 8 逐步增加到 64,模型在 MOS 上分别提升了 0.6%和 0.68%,RTF 分别降低了 279%和 770%,Peak Mem 占用分别增加了 27.8%和 73%,模型 Params 分别增加了 3.8%和 4.6%. 这表明增加 channel 数量改善了模型的性能,但同时也增加了计算资源消耗和模型的复杂性. 因此,本文所设置的 S=64 和 channel=16 兼顾了模型性能和计算成本.

4 结论

本文提出了一种创新的文本到语音(TTS)模型 EnhanceSpeech,该模型通过集成两个即插即用的创新模块,显著提升了语音合成的性能与音质. 首先,EnhanceSpeech 引入了一种可学习的外部注意力机制,有效减少计算量和内存占用,从而在实际语音合成应用中表现出显著的性能优势. 其次,EnhanceSpeech 利用基于分层挤压注意力的后处理网络生成梅尔频谱图,将其视作图像处理,显著提升了梅尔频谱图的生成质量. 同时该模块能够更好地处理噪音,并专注于细节化梅尔频谱图的波形特征,进而提升了合成语音的整体质量. 未来工作中,将继续提高合成语音的质量和速度,并致力于解决文本转语音过程中的多音字及参数优化问题.

[参考文献]

[1] ARIK S O, KLIEGL M, CHILD R, et al. Convolutional recurrent neural networks for small-footprint keyword spottin[C]// Proceedings of the 18th Annual Conference of the International Speech Communication Association(INTERSPEECH 2017). Stockholm, Sweden: ISCA, 2017: 1606-1610.

[2] 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(1): 6706-6713.

[3] SHEN J, PANG R, WEISS R, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Calgary, Canada: IEEE, 2018: 4779-4783.

[4] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[C]//18th Annual Conference of the International-Speech-Communication-Association. Stockholm, Sweden: International Speech Communication Association, 2017: 4006-4010.

[5] LIU J, LI C, REN Y, et al. DiffSinger: Singing voice synthesis via shallow diffusion mechanism[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 36(10): 11020-11028.

[6] PENG Y, LIU B. Attention-based neural network for short-text question answering[J]. Proceedings of the 2018 2nd international conference on deep learning technologies, 2018: 21-26.

[7] REN Y, LIU J, TAN X, et al. SimulSpeech: end-to-end simultaneous speech to text translation[C]//58th Annual Meeting of the Association-for-Computational-Linguistics. Electric Network: ACL, 2020: 3787-3796.

[8] YANG B, ZHONG J, LIU S. Pre-Trained text representations for improving front-end text processing in mandarin Text-to-Speech synthesis[C]//Interspeech Conference. Graz, Austria: International Speech Communication Association, 2019: 4480-4484.

[9] REN Y, TAN X, QIN T, et al. Almost unsupervised text to speech and automatic speech recognition[C]//36th International Conference on Machine Learning. Long Beach, CA: JMLR, 2019: 97.

[10] LEE Y, SHIN J, JUNG K. Bidirectional variational inference for non-autoregressive Text-to-Speech[C/OL]//International Conference on Learning Representations. Online, 2020. https://openreview.net/forum? id=S1g_G1HwDB.

[11] HAYASHI T, YAMAMOTO R, YOSHIMURA T, et al. ESPnet2-TTS: extending the edge of TTS research[J]. arXiv Preprint arXiv: 2110.07840, 2021.

[12] JEONG M, KIM H, CHEON S J, et al. Diff-TTS: a denoising diffusion model for Text-to-Speech[C]//Interspeech Conference. Brno, Czech Republic: International Speech Communication Association, 2021: 3605-3609.

[13] LIM D, JANG W, O G, et al. JDI-t: jointly trained duration informed transformer for Text-to-Speech without explicit alignment [C]//Interspeech Conference. Shanghai, China: International Speech Communication Association, 2020: 4004-4008.

[14] KIM G, HONG S, FRANZ M, et al. Improving cross-platform binary analysis using representation learning via graph alignment

[C]//Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis. New York, NY, USA: Association for Computing Machinery, 2022: 151-163.
[15] HUANG W C, WU Y C, HAYASHI T. Any-to-One sequence-to-sequence voice conversion using self-supervised discrete speech representations[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 5944-5948.
[16] REN Y, RUAN Y, TAN X, et al. Fastspeech: fast, robust and controllable text to speech[C]//33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing System, 2019: 32.
[17] YU J, XU Z, HE X, et al. DIA-TTS: deep-inherited attention-based Text-to-Speech synthesizer[J]. Entropy, 2022, 25(1): 41.
[18] ZHOU K, SISMAN B, LI H. Limited data emotional voice conversion leveraging Text-to-Speech: two-stage sequence-to-sequence training[C]//Interspeech Conference. Brno, Czech Republic: International Speech Communication Association, 2021: 811-815.
[19] LI N, LIU Y, WU Y, et al. RobuTrans: a robust transformer-based Text-to-Speech model[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 8228-8235.
[20] OKAMOTO T, TODA T, SHIGA Y, et al. Transformer-based Text-to-Speech with weighted forced attention[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6729-6733.
[21] REN Y, HU C, TAN X, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[C/OL]//International Conference on Learning Representations. Online, 2021. https://openreview.net/forum? id=pi-n2_533c.
[22] LIAN J, ZHANG C, ANUMANCHIPALLI G K, et al. Unsupervised TTS acoustic modeling for TTS with conditional disentangled sequential VAE[J]. IEEE/ACM transactions on audio, speech, and language processing, 2023, 31: 2548-2557.
[23] JING X, CHANG Y, YANG Z. U-DiT TTS: U-Diffusion vision transformer for Text-to-Speech[C]//Proceedings of the 49th DAGA Conference on Acoustics. Hamburg, Germany: VDE VERLAG GMBH, 2023: 110-113.
[24] JIANG Z, REN Y, YE Z, et al. Mega-TTS: zero-shot Text-to-Speech at scale with intrinsic inductive bias[J]. arXiv Preprint arXiv: 2306.03509, 2023.
[25] JIANG Z, LIU J, REN Y, et al. Mega-TTS 2: boosting prompting mechanisms for zero-shot speech synthesis[J]. arXiv Preprint arXiv: 2307.07218, 2023.
[26] LIU H, HUANG R, LIN X, et al. ViT-TTS: visual Text-to-Speech with scalable diffusion transformer[C]//Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023: 15957-15969.
[27] LEE K, KIM D W, KIM J, et al. DiTTo-TTS: efficient and scalable zero-shot Text-to-Speech with diffusion transformer[J]. arXiv Preprint arXiv: 2406.11427, 2024.
[28] ITO K, JOHNSON L. The LJ Speech Dataset[DB/OL]. (2017). https://keithito.com/LJ-Speech-Dataset/.
[29] DATABAKER. Chinese standard mandarin speech corpus[DB/OL]. (2019). https://www.data-baker.com/.
[30] PARK H, KIM Y, KIM J, et al. g2p[J]. Journal of arrhythmia, 2019, 35(4): 593-601.
[31] PARK K. G2PC: A grapheme-to-phoneme converter for Chinese [DB/OL]. (2019). https://github.com/kyubyong/g2pc.
[32] LI X, CHENG Z Q, HE J Y, et al. MM-TTS: a unified framework for multimodal, prompt-induced emotional Text-to-Speech synthesis[J]. arXiv Preprint arXiv: 2404.18398, 2024.
[33] GUAN W, SU Q, ZHOU H, et al. Reflow-TTS: a rectified flow model for high-fidelity Text-to-Speech[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea: IEEE, 2024: 10501-10505.
[34] RAITIO T, LI J, SESHADRI S. Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7587-7591.
[35] 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.
[36] KIM J, KIM S, KONG J, et al. Glow-tts: a generative flow for Text-to-Speech via monotonic alignment search[C]//34th Conference on Neural Information Processing Systems. Electric Network: Neural Information Processing System. 2020: 33.
[37] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates, Inc., 2017, 30: 5998-6008.
[38] CHEN Z, WU G, GAO H, et al. Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation[J]. Expert systems with applications, 2023, 232: 120828.
[39] LI W, HOU Z, ZHOU J, et al. SiamBAG: band attention grouping-based siamese object tracking network for hyperspectral videos[J]. IEEE transactions on geoscience and remote sensing, 2023, 61: 1-12.

[责任编辑:杜忆忱]