arXiv:2303.01849v1 [eess.AS] 3 Mar 2023

一项调查扩散模型在适应能力方面的表现

Haolin Chen，Philip N. Garner

Idiap研究所，马蒂尼，瑞士

摘要

鉴于扩散技术近期在生成具有自然度的合成语音方面取得了成功，我们研究了如何将扩散用于说话人自适应文本转语音。借鉴更为传统的自适应方法，我们展示了如何通过结合步嵌入的条件层归一化，将自适应功能纳入扩散流程。然而，实验结果表明，尽管该方法具有一定价值，但仅凭这种自适应方式仍无法达到基于Transformer的技术的性能。在第二项实验中，我们展示了如何将扩散与Transformer进行最优结合：由后者承担大部分自适应任务，而前者则有助于提升自然度。

索引词— 文本转语音、说话人自适应、扩散模型、条件层归一化

1. 引言

近年来，自适应文本转语音（TTS） [1, 2, 3, 4]已成功应用于为目标说话人合成个性化语音。在典型的自适应文本转语音场景中，源声学模型通常在大型多说话人语料库上训练，再利用少量自适应数据进行自适应，以合成所需的语音。与此同时，在声学建模这一更广泛的领域中，深度生成模型（DGMs） [5, 6, 7] 已展现出相较于其他方案的优势，能够实现高质量、快速的合成。尤其是近来的扩散模型 [7, 8, 9] ，在可懂度和自然度方面已成为该领域的主导方法。

当前自适应 TTS 的研究主要聚焦于：1）增强源模型对各种声学条件和风格的泛化能力；以及 2）提升自适应的数据和参数效率。前者还可进一步分为：1）采用可插拔的参考编码器，在不同语义层级上生成声学信息和风格的表示 [1, 2, 10]；以及 2）针对性设计模型结构，以控制目标特征 [2, 3]。此外，这类自适应技术应建立在合成质量较高的架构之上；在这一方面，基于扩散的声学模型已经超越了基于流的前代模型 [5, 11], , 同时在网络设计上具有更大的灵活性。自扩散模型首次应用于 TTS 以来，许多研究 [9, 12, 13] 已证明如何大幅加速其生成过程，使其速度接近同类模型中的最快者，同时不会显著降低合成质量。

总的来说，我们关注基于扩散的声学模型的参数高效自适应技术，以增强其泛化能力。尽管扩散模型已在通用声学建模领域得到充分探索，但鲜有研究将其用于自适应文本转语音系统。据我们所知，Guided-TTS 2 [14], 是唯一一种基于扩散的自适应文本转语音系统，它利用带分类器引导的扩散来适应不同的语音。然而，该方法缺乏针对各个目标对象的参数效率，因为自适应过程中会微调扩散模型的所有参数；此外，它也不属于典型的编码器-解码器框架。鉴于其他架构（如 Transformer） [2], 已有参数高效的自适应技术，且扩散模型具有出色的合成质量，社区将非常关注适用于扩散模型的此类方法，从而实现兼具参数效率与高质量的自适应。

基于上述分析，我们研究了基于扩散的声学模型的适应能力，尤其关注参数高效的解决方案。具体而言，我们的实验基于一种典型的扩散去噪网络架构，即双向扩张卷积神经网络。受基于 HMM 的自适应研究发现启发，我们提出将条件层归一化（CLN）引入去噪网络。初步实验结果表明，尽管自适应扩散解码器是可行的，但仅依靠扩散自适应不足以实现高质量的自适应；结果还显示，该去噪器的泛化能力和适应能力较差。我们进一步将自适应 Transformer 层作为解码器的一部分，并考察了添加不同数量此类层对适应质量的影响。结果表明，虽然去噪网络中的 CLN 有助于提升语音质量和说话人相似度，但必须与自适应 Transformer 层结合使用，才能达到可用的适应质量。我们得出结论，对于这种特定类型的扩散模型，在自适应文本转语音系统中的最佳用途是充当后处理网络，帮助细化 Transformer 解码器生成的梅尔频谱图细节。

2. 自适应扩散解码器

2.1. 基于扩散的声学模型

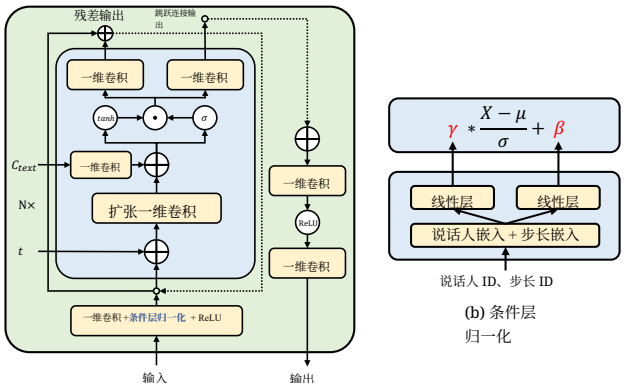
原则上，扩散模型通过扩散过程将先验分布中的样本逐步去噪，从而生成真实数据。尽管采用的方法各不相同，扩散模型的学习问题都可以表述为学习一个去噪网络，使其能够预测每个扩散步骤中的噪声。扩散声学模型的主流架构由基于 Transformer 的音素编码器和扩散去噪解码器组成。本文主要关注去噪器的网络设计。

去噪网络最常用的结构是双向扩张卷积网络 [7, 8, 9, 13]；其他选择包括 U-Net [14, 15]。如图 1a 所示，去噪器将前一步的样本作为输入，在编码音素序列 C_{text} 和步嵌入 t的条件下，预测反向扩散过程中的噪声。该网络主要由一个输入卷积层和 N 个带有残差连接和跳跃输出的卷积块组成，随后将跳跃输出累加，并通过输出卷积层生成最终预测。

2.2. 去噪器的条件层归一化

先前的研究 [2, 3]发现，Transformer 中的层归一化可通过轻量级、可自适应的缩放向量 γ 和偏置向量 β显著影响输出： $LN(x) = \frac{x-\mu}{\sigma} * \gamma + \beta$，其中 μ 和 σ 分别是输入向量 x的均值和方差。此外，这两个向量可以由一个以说话人嵌入为条件的小型神经网络生成。适配新声音时，可以对该网络进行微调，从而在保持适应质量的同时，显著减少需要调整的参数数量。遵循 [2], ，我们将此模块称为条件层归一化（CLN）。

具体而言，我们希望将条件层归一化集成到去噪器中。考虑到条件层归一化在 Transformer 中的应用，其操作作用于整个隐藏表示，并沿着 Transformer 块逐步改变预测。回到去噪网络，最终预测通过累加所有卷积块的跳跃输出得到，这一设计受到多层级融合特征以生成精细细节这一理念的启发。在这些块内部或之间添加条件层归一化，会使其仅作用于部分特征，从而阻碍模型
学习此类层级信息。此外，卷积块的数量 N 通常很多（≥ 12），如果在每个卷积块中都放置条件层归一化（CLN），就会降低参数效率。经过初步的临时实验验证上述假设后，我们将条件层归一化（CLN）放置在输入处的一维卷积层之后，如图 1a 所示。这种放置方式也受到基于隐马尔可夫模型的自适应方法的启发，例如



(a) 去噪网络

(b) 条件层归一化

图 1：去噪网络和条件层归一化（CLN）示意图。

[16], 其中，归一化是在整个特征的频域（或倒谱域）中进行的。

扩散模型的一个特性是，在所有扩散步骤中使用同一个去噪器，并通过步嵌入对其进行条件控制；这也使其成为一种参数高效的解决方案。条件层归一化（CLN）并非只以说话人嵌入为条件，而是将说话人嵌入和步嵌入拼接起来，以控制操作强度，如图 1b 所示。由于该模块在所有扩散步骤中共享，而且归一化只应在样本得到进一步细化时才开始生效，而不是在一开始就生效，因此这一机制使去噪器能够在整个反向扩散过程中自动学习何时开始发挥作用以及作用强度。该方法也受到[17], 的启发，其中基于 Transformer 的去噪器通过条件层归一化（CLN）以步嵌入为条件。

以下各节将介绍相关实验，以验证迄今所述的系统。

2.3. 实验设置

实现细节。 实验中使用的扩散解码器模型架构基于 PriorGrad [9]。作为首个基于扩散的声学模型 Diff-TTS [7], 的改进版本，PriorGrad 利用数据依赖先验，其均值和方差是从数据集中提取的音素级统计量。与使用标准高斯先验的模型相比，PriorGrad 能提供更好的合成质量、更高的参数效率和更快的收敛速度。对于扩散解码器的架构，我们采用了 AdaSpeech [2], 的架构，包括音素编码器、声学条件建模模块和方差适配器。我们的实现基于这两个模型的开源软件 [1] [2] 。我们使用

[1]NATSpeech： https://github.com/NATSpeech/NATSpeech [2] NeuralSpeech： https://github.com/microsoft/NeuralSpeech

文献 [9]提出的扩散解码器，其包含 12 个卷积块、128 个残差通道和 3.5M 个参数。除非另有说明，其他模型配置均遵循 AdaSpeech 和 PriorGrad 的相应部分。扩散解码器需要微调的参数总数为 0.131M，而 AdaSpeech 为 1.184M。

数据。 我们在 LibriTTS 数据集的两个干净子集 train-clean-100 和 train-clean-360 上训练源模型 [18]，。该多说话人文本转语音语料库共包含 1151 位说话人和 245 小时的语音。评估时，我们按照 [4]中的做法，从 VCTK [19]中选取 11 位口音各异的说话人（7 位女性和 4 位男性）。每位说话人的测试集均随机选取 10 条话语，且所有说话人的转录文本相同。语音和文本数据的预处理遵循 AdaSpeech，采样率则设为 22,050 Hz。

训练、自适应与推理。 按照 AdaSpeech 的方法，训练过程分为两个阶段，步数分别为 200K 和 100K。模型在一块 NVIDIA RTX 3090 GPU 上训练，批次大小为 50,000 个语音帧。对于扩散解码器，训练和推理均采用包含 400 步的贝塔调度。我们使用在整个训练集上计算得到的说话人无关先验。除非另有说明，其他超参数均遵循 [9]。自适应时，使用目标说话人的 10 条话语对模型进行微调，训练 2000 步，学习率固定为 2×10^{-4}，且仅优化说话人嵌入和条件层归一化（CLN）。推理过程中，使用在 VCTK 上训练的 HiFi-GAN 声码器 [20] ，将生成的频谱图合成为波形。

2.4. 客观评估

在初步评估中，我们采用 MOSNet [21]，——一种基于神经网络的语音质量客观评估工具，可生成机器评分的平均意见分（MOS）；此外，还使用了 SpeechBrain [22] 提供的预训练说话人验证模型，该模型会计算生成样本与参考样本的说话人嵌入之间的余弦相似度（CS）。余弦相似度的取值范围为 0 到 1；分数越高，表示与参考样本的说话人相似度越高。我们发现，这两种自动评估工具的评分与我们的主观判断一致，并将在最终设置下进行人工评估。

我们比较以下几种设置的性能：1) 真实值梅尔频谱 +声码器，使用真实值梅尔频谱图和 HiFi-GAN 声码器合成波形；2) AdaSpeech，一种基于 Transformer 的自适应声学模型，在解码器中应用了条件层归一化；3) Enc+DiffDec（解码器），一种去噪器中不含条件层归一化的基线系统，在自适应过程中对整个解码器进行微调，作为性能上界；4) Enc + DiffDec（说话人嵌入），架构与前一种方法相同，但仅微调说话人嵌入，作为性能下界；5) Enc +AdaDiffDec，我们提出的系统，在去噪器中加入条件层归一化，并在自适应过程中与说话人嵌入一起进行微调。

表 1：MOSNet 和余弦相似度（CS）评分。

#	模型	MOSNet（↑）	CS（↑）
1	真实值梅尔频谱 + 声码器	4.10	0.96
2	*AdaSpeech*	3.78	0.52
3	*Enc + DiffDec*（解码器）	3.80	0.50
4	*Enc + DiffDec*（说话人嵌入）	3.42	0.20
5	*Enc + AdaDiffDec*	3.58	0.22

2.5. 结果与分析

MOSNet 和余弦相似度的结果如表 1 所示。可以观察到：1）对整个扩散解码器进行自适应（#3）在所有设置中取得了最佳的语音质量和说话人相似度，性能与 AdaSpeech（#2）相近；2）仅微调说话人嵌入（#4）会导致性能较差；3）我们提出的方法（#5）仅略优于基线（#4），但仍远逊于微调整个解码器（#3）和 AdaSpeech（#2）。

结果表明，仅依靠对去噪器中的条件层归一化进行自适应，不足以获得合理的适应质量。此外，我们的测试聆听结果表明，三个基于扩散的系统（#3–5）合成的部分样本难以辨认，这解释了它们为何在客观测试中不如 AdaSpeech。这也说明扩散解码器对域外输入较为敏感，因此泛化能力和适应能力较差。这一点非常有趣，因为作为通用声学模型，它能够合成质量极高且自然的语音。这一现象表明，还需要进一步努力提升该系统的适应能力。

3. 自适应扩散：是个好选择吗?

我们的初步实验结果表明，先前提出的系统无法达到实用的自适应质量，这说明仅对扩散解码器进行自适应可能不是一个好选择；需要在系统中引入其他组件，以提升自适应性能，同时利用扩散解码器的高质量合成能力。

鉴于基于 Transformer 的自适应 TTS 系统已取得不错的适应质量，我们考虑在扩散解码器之前添加带有条件层归一化（CLN）的 Transformer 层，以构建混合架构的解码器。该方法的灵感来自一种常见做法：在声学模型中使用深度生成模型（DGMs）作为后处理网络（post-net），以 [6, 23]

改进 VAE（变分自编码器）或 Transformer 生成的过度平滑样本。如果扩散解码器作为后处理网络来改进 AdaSpeech 的输出，预计适应质量将大幅提升。我们尤其关注额外的 Transformer 层能带来多少性能提升，以及同时对 Transformer 解码器和扩散解码器进行自适应，与仅对 Transformer 解码器进行自适应之间有何差异。

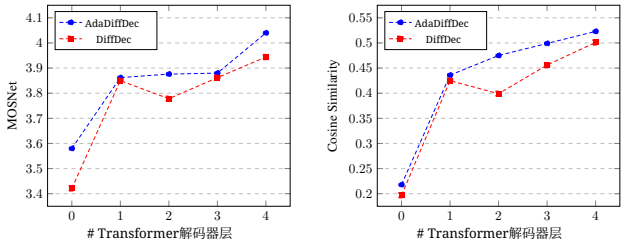
3.1. 实验与评估

此设置中的模型配置采用网格搜索，组合以下两个因素：1）额外 Transformer 解码器层的数量，从 0 到 4，其中 4 对应完整的 Transformer 解码器；2）是否在扩散去噪器中使用条件层归一化。由于需要比较的系统超过 10 个，我们首先像之前的实验一样进行客观评估。随后，我们进一步开展主观听测，以验证客观测试结果所反映的发现。我们预计，增加 Transformer 层数会提升语音质量和说话人相似度。然而，在此类设置下，更重要的是观察扩散中的条件层归一化对性能的影响。

在主观听测中，我们邀请了 10 名评定者，对每个系统的 10 个样本进行自然度平均意见分（MOS）和相似度 SMOS 评分。测试话语从客观测试所用的话语中随机选取，涵盖大多数说话人或口音。请注意，所有系统使用的测试话语都相同。我们选取了两种具有 4 个 Transformer 解码器层的设置，它们相当于在 AdaSpeech 的基础上将扩散解码器用作后处理网络，预计性能最佳；并将其与声码器合成的真实语音和 AdaSpeech 进行比较。

3.2. 结果

客观评估的结果如图 2 所示，其中两项指标的结果分别以独立图表展示。主要观察结果包括：1）与仅使用扩散解码器相比，额外的 Transformer 层显著提升了性能；2）总体而言，随着 Transformer 层数的增加，语音质量和说话人相似度均有所提升；3）在去噪器中加入条件层归一化后，在所有设置下两项指标的表现均有所改善，但当 Transformer 层数较高时，说话人相似度的差异会缩小。主观测试结果如表 2 所示，其中“AS”代表 AdaSpeech。可以看出：1）在 Transformer 解码器之上加入扩散解码器（#3、#4），与仅使用 Transformer 解码器的 AdaSpeech（#2）相比，显著提升了感知语音质量和说话人相似度；2）扩散解码器中的条件层归一化进一步提升了这两项得分，使得 #3 在所有系统中表现最佳；3）语音质量的提升幅度



(a) MOSNet (b) 余弦相似度

图 2：不同 Transformer 解码器层数量设置的 MOSNet 分数和余弦相似度分数。

表 2：MOS 和 SMOS 分数及其 95% 置信区间。

#	模型	平均意见分（↑）	SMOS（↑）
1	真实值梅尔频谱 +声码器	4.77 ± 0.10	4.99 ± 0.02
2	*AdaSpeech*	2.38 ± 0.19	2.94 ± 0.20
3	作为 *+AdaDiffDec*	3.02 ± 0.19	3.24 ± 0.20
4	作为 *+DiffDec*	2.84 ± 0.19	3.16 ± 0.21

大于说话人相似度的提升幅度，后者仅表现出轻微差异。

总体而言，研究表明，尽管去噪网络中的条件层归一化有助于提高自适应质量，但必须与自适应 Transformer 层结合使用，才能达到可用的性能。模型的适应能力主要依赖于自适应 Transformer 层，这表明与 Transformer 相比，扩散去噪器的泛化能力和适应能力较弱。

4. 结论

本文研究了典型的基于扩散的声学模型在参数高效设置下的适应能力。我们为去噪网络提出了条件层归一化，并测试了其在说话人自适应中的效果。研究表明，尽管可以通过这种方法对扩散解码器进行自适应，但必须将其与自适应 Transformer 层结合使用，才能达到可用的自适应质量。这表明，扩散模型的泛化能力和适应能力不如 Transformer。未来的研究应着重改进扩散模型在上述方面的表现，或者将扩散模型用作后处理网络，以细化其他可自适应组件生成的梅尔频谱图。

5. 致谢

本项目获得了瑞士国家科学基金会资助，项目名称为 NAST：语音技术神经架构，资助编号 185010。

6. 参考文献

[1] Yuxuan Wang 等人，“风格令牌：端到端语音合成中的无监督风格建模、控制与迁移”，发表于第 *35* 届国际机器学习大会（*ICML*），2018年，第 80 卷，第 5167–5176 页。

[2] Mingjian Chen 等，“AdaSpeech：用于定制语音的自适应文本转语音”，载于第九届国际学习表征大会，*ICLR*，2021年。

[3]Dongchan Min、Dong Bok Lee、Eunho Yang 和 Sung Ju Hwang，“Meta-StyleSpeech：多说话人自适应文本转语音生成”，载于第*38*届国际机器学习大会论文集，*ICML*，2021年，第139卷，第7748–7759页。

[4] Edresson Casanova 等，“YourTTS：面向所有人的零样本多说话人文本转语音和零样本语音转换”，载于国际机器学习大会，*ICML 2022*，2022年，第162卷，第2709–2720页。

[5] Jaehyeon Kim、Sungwon Kim、Jungil Kong 和 Sungroh Yoon，“Glow-TTS：通过单调对齐搜索实现文本转语音的生成流”，载于*2020*年神经信息处理系统进展，2020年。

[6] Jaehyeon Kim、Jungil Kong 和 Juhee Son，《采用对抗学习的条件变分自编码器实现端到端文本转语音》，载于第*38*届国际机器学习会议，*ICML*，2021年，第139卷，第5530–5540页。

[7] Myeonghun Jeong、Hyeongju Kim、Sung Jun Cheon、Byoung Jin Choi 和 Nam Soo Kim，《Diff-tts：用于文本转语音的去噪扩散模型》，载于*Inter-speech2021*，2021年，第3605–3609页。

[8] Jinglin Liu、Chengxi Li、Yi Ren、Feiyang Chen 和 Zhou Zhao，《Diffsinger：通过浅层扩散机制实现歌声合成》，载于第*36*届美国人工智能学会人工智能会议，2022年，第11020–11028页。

[9] Sang gil Lee 等人，“Priorgrad：通过数据依赖的自适应先验改进条件去噪扩散模型”，发表于国际学习表征会议，2022年。

[10] Rongjie Huang、Yi Ren、Jinglin Liu、Chenye Cui 和 Zhou Zhao，“GenerSpeech：面向可泛化域外文本转语音合成的风格迁移”，*CoRR*，第 abs/2205.07211 卷，2022年。

[11] Rafael Valle、Kevin J. Shih、Ryan Prenger 和 Bryan Catanzaro，《Flowtron：一种用于文本转语音合成的自回归流式生成网络》，载于第九届国际学习表征会议，*ICLR*，2021年。

[12] Songxiang Liu、Dan Su 和 Dong Yu，《Diffgan-tts：利用去噪扩散生成对抗网络实现高保真、高效的文本转语音》，*CoRR*，第 abs/2201.11972 卷，2022年。[13] Rongjie Huang、Zhou Zhao、Huadai Liu、Jinglin Liu、Chenye Cui 和 Yi Ren，《Prodiff：用于高质量文本转语音的渐进式快速扩散模型》，载于*MM ’22*：第*30*届 *ACM* 多媒体国际会议，2022年，第2595–2605页。[14]Sungwon Kim、Heeseung Kim 和 Sungroh Yoon，《Guided-tts 2：一种利用未转录数据实现高质量自适应文本转语音的扩散模型》，*CoRR*，第 abs/2205.15370 卷，2022年。

[15] Vadim Popov、Ivan Vovk、Vladimir Gogoryan、Tasnima Sadekova 和 Mikhail A. Kudinov，“Grad-tts：一种用于文本转语音的扩散概率模型”，发表于第*38*届国际机器学习会议，*ICML*，2021年，第139卷，第8599–8608页。

[16]Christopher J. Leggetter 和 Philip C. Woodland，“使用线性回归进行隐马尔可夫模型的说话人自适应”，技术报告 CUED/F-INFENG/TR. 181，剑桥大学工程系，Trumpington Street，剑桥，CB2 1PZ，英格兰，1994年6月。

[17] Shuyang Gu 等，“用于文本到图像合成的矢量量化扩散模型”，载于 *IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2022*，2022年，第 10686–10696 页。

[18] Heiga Zen 等，“Libritts：源自 LibriSpeech 的文本转语音语料库”，载于 *Interspeech2019*，2019年，第 1526–1530 页。

[19] Junichi Yamagishi、Christophe Veaux 和 Kirsten Mac- Donald，“CSTR VCTK语料库：用于 CSTR语音克隆工具包的英语多说话人语料库（0.92版）”，2019年。

[20] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，“HiFi-GAN：用于高效且高保真语音合成的生成对抗网络”，载于 *2020*年神经信息处理系统进展，2020年。

[21] Chen-Chou Lo 等人，“MOSNet：一种基于深度学习的语音转换客观评估方法”，载于 *Interspeech2019*，2019年，第 1541–1545 页。

[22]Mirco Ravanelli 等人，“SpeechBrain：一款通用语音工具包”，2021年，arXiv:2106.04624。

[23] Yi Ren、Jinglin Liu 和 Zhou Zhao，“Portaspeech：便携且高质量的生成式文本转语音”，载于神经信息处理系统进展 *2021*，2021年，第 13963–13974 页。