

Bert-FlowTTS: 基于 Flow 的语音合成模型

李旺超, 孙 俊

(江南大学人工智能与计算机学院, 无锡 214026)

摘要: 基于标准流 (Normalizing Flow) 的语音合成方法获得广泛的普及, 但是使用标准流模型合成的音频在韵律上存在不自然的问题, 同时语音合成模型需要“单调的硬对齐”, 才能确保合成的音频不会出现漏字和重读现象。针对以上问题, 首先提出增加一个额外的预训练模型来加强神经网络模型对韵律的建模, 使得模型能够挖掘出语音合成输入文本中潜在的句法和语义信息, 其次在上述基础上采用单调对齐搜索策略来实现文本到音频帧的硬对齐, 同时我们使用生成对抗网络 (GAN) 来训练随机时长预测器, 使得合成的音频具有不同的韵律。最后探索并验证了预训练模型对语音合成模型性能的提升, 并总结了详细的模型训练流程。

关键词: 语音合成; 标准流; 预训练模型; 单调对齐搜索; 生成对抗网络
中图分类号: TP391

Bert-FlowTTS: Flow-based speech synthesis model

LI Wangchao, Sun Jun

(School of Artificial Intelligence and Computer Science, JiangNan University, Wuxi 214026)

Abstract: Recently, A widely adopted approach for speech synthesis is based on Normalizing Flow. but text-to-speech using Flow models often suffers from unnatural rhythm. Additionally, text-to-speech models require “monotonic hard alignment” to ensure that the synthesized audio does not have missing or repeated words. To solve the above problems, this paper proposed the neural network model with an additional pretrained model to enhance the modeling of rhythm. This allows the model to find potential syntactic and semantic information from the input text. Secondly, we utilize a Monotonic Alignment Search strategy to achieve hard alignment between the input text and audio frames. Moreover, we incorporate a Generative Adversarial Network (GAN) to train a Stochastic Duration Predictor, enabling the synthesized audio to have different rhythms. And the final we explore and validate the performance improvement of the pretrained model on speech synthesis models and provide a detailed model training procedure.

Key words: text-to-speech; normalizing flow; pretraining model; monotonic alignment search; GAN

0 引言

语音合成 (Text to Speech, TTS) 是一种由给定文本信息生成语音的技术, 是实现人机交互的重要技术, 可以让人与机器间更流畅地交谈。目前, 随着语音合成技术的快速发展, 语音合成技术在虚拟客服, 智能设备, 车载系统, 机器人等领域广泛应用。传统的语音合成方法有拼接法^[1]、参数法^[2]、统计参数法^[3], 但上述方法构建的语音合成系统过于耗时费力, 且隐马尔可夫模型的建模能力有限, 合成的音频质量通常较差。

近年来, 随着深度学习技术的迅速发展, 传统语音合成方法已经被基于深度神经网络的

作者简介: 李旺超 (1999-), 男, 主要研究方向: 深度学习, 语音合成

通信联系人: 孙俊 (1971-), 男, 博导, 主要研究方向为人工智能、计算智能、机器学习、大数据分析等. E-mail: junsun@jiangnan.edu.cn

方法取代，基于深度神经网络的生成式模型在合成语音的自然度，流畅度上提升巨大，现已被广泛用于工业界，例如 Tacotron^[4]，Tacotron2^[5]，FastSpeech^[6]，FastSpeech2^[7]，GlowTTS^[8]，VITS^[9]等。基于神经网络的语音合成通常有两阶段和端到端之分，两阶段的语音合成方法通常先利用声学模型根据输入的文本生成对应的 Mel 声谱图（Mel Spectrogram），第二步使用声码器（Vocoder），将 Mel 声谱图转换为对应的语音波形。端到端的语音合成方法可以直接通过声学模型生成语音波形，少了中间特征 Mel 声谱图的输出。现在的主流趋势是端到端的语音合成方法。

在当前基于神经网络的语音合成模型中，以自回归的方式生成 Mel 谱图，其特别明显的缺点是合成音频的速度过慢，并且随着待合成文本长度的增加，推理时间也是线性增长的。但是很多现实场景对语音合成的实时性有着高要求，所以自回归的语音合成方法也慢慢被非自回归所取代。

本文使用基于流^[10, 11, 12]（Flow）的语音合成模型实现可并行的生成 Mel 谱图，在合成音频的速度上有着巨大的提升。但是，在生成的音频存在着韵律不自然等问题，为了解决这些问题，本文在韵律建模，预训练模型方面进行了尝试，并和主流的非自回归模型和非自回归模型进行了对比。实验结果表明，本文提出的方法在流畅度、自然度上有一定的提升。

1 相关研究

目前主流的语音合成方法都是基于深度神经网络的方法。Wang 等人提出了 Tacotron 系列模型，它由编码器、解码器和声码器组成，它通过自回归的方式来产生 Mel 频谱特征，其缺点也很明显，合成语音的速度慢，同时自回归模型缺乏鲁棒性，例如，当输入文本包含重复单词时，自回归模型通常会产生严重的注意力错误。Ren 等人提出了 FastSpeech 系列模型，FastSpeech 采用的是自注意力机制和一维卷积为模型基础，以非自回归的形式得到梅尔特征输出，这有效的解决了合成语音速度慢的问题，同时 FastSpeech 引入了一个时长预测模块，这个模块是基于 Teacher-Student 的知识蒸馏^[13]框架得到的，但需要先训练一个自回归的语音合成 Teacher 模型，时长预测模块用于解决模型输入和输出长度及不匹配的问题，不过 FastSpeech 合成的语音质量很大程度上依赖 Teacher 模型。随着 Flow 模型在图像生成领域取得的成功，Kim 等人将 Flow 模型用于语音合成领域，提出了 GlowTTS，这是一个基于标准流的生成式模型，GlowTTS 不需要借助额外的模型产生对齐，它可以自己产生文本特征和 Mel 特征的单调对齐，增强了语音合成的鲁棒性，并且合成效果和基于自回归的 Tacotron2 相当。随后 Kim 等人又提出了 VITS，它将变分自编码器^[14]（Variational AutoEncoder, VAE）、Flow 和生成对抗网络^[15]（Generative Adversarial Network, GAN）模型结合起来，实现了一个可以直接从文本到语音波形的端到端模型，并且合成的音频质量逼近真人。

虽然现在先进的神经网络模型实现了端到端的语音合成，使得人们不再去对输入数据进行过多的处理，但是，这些语音合成模型对韵律的建模不够精确，通常只是建模一个均一化

的韵律，这导致了合成语音的不自然。

本文采用基于 Flow 模型作为语音合成的基本架构，针对当前非自回归语音合成模型存在的问题，进行了一些尝试和改进。主要的贡献如下：

从 PnG BERT^[16]得到启发，本文引入 Bert^[17]（Bidirectional Encoder Representations from Transformers, BERT）预训练模型到文本编码器中，通过预训练模型提供的深层的语义特征表示，有效缓解非自回归语音合成存在的韵律不自然的现象，也缓解了语音的数据量有限的问题。此外，我们对合成字符的声调也进行建模，使得模型能学习到每个音素的重读、轻度情况。

引入单调对齐搜索^[8]（Monotonic Alignment Search, MAS）模块，解决了文本特征和语音特征之间的对齐问题，并结合随机时长预测器模块，有效的解决了从文本到波形之间的一对多的映射关系。

采用生成对抗网络（GAN）来训练随机时长预测器（Stochastic Duration Predictor, SDP），使得文本和 Mel 频谱帧的对齐更加符合实际的发音情况。

2 基于 Flow 的生成式语音合成

本文在 GlowTTS 模型的基础上应用了两个改进点，其一是引入 Bert 来捕获句子中潜在的语法语义信息以此来合成更具表现力的语音，其二是利用 GAN 网络来得到一个最符合实际发音情况的时长预测器模型，使得生成的语音具有良好的韵律表现。整体的网络架构如下图所示：

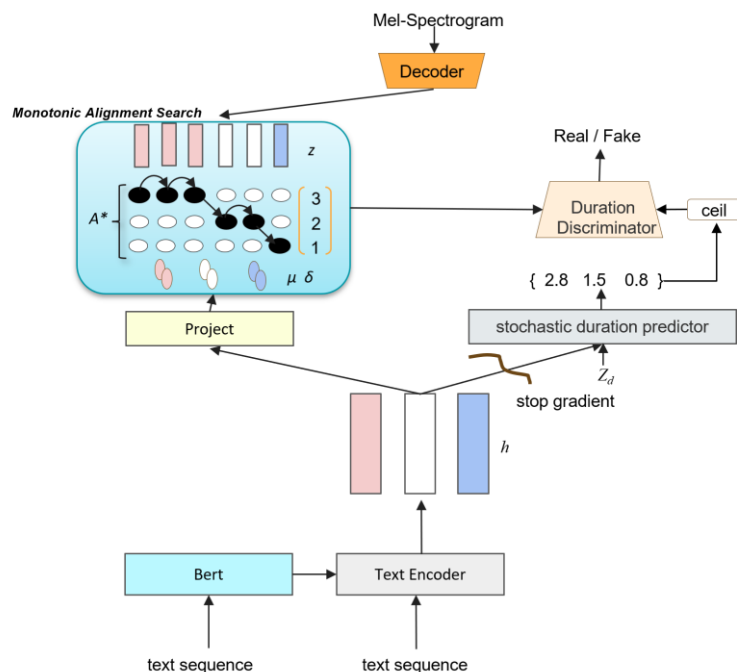


图1 Bert-FlowTTS 的训练流程图

Fig. 1 Training flow chart of Bert-FlowTTS

基于 Flow 的语音合成可以直接通过直接最大化似然函数来进行训练，因为基于 Flow 的语音合成模型可以写出确切的似然函数。如公式(1)所示：

$$\log P_X(x|c; \theta, A) = \log P_Z(z; c, \theta, A) + \log \det \frac{\partial f_{dec}(x)}{\partial z} \quad (1)$$

其中 x 表示音频的 Mel 频谱, c 表示音频中的文本内容, θ 表示文本编码器的参数, A 表示单调对齐矩阵, f_{dec} 是语音合成的解码器模块。我们的训练目标就是要最大化该对数似然。

模型的训练流程如下: 从一段音频中得到 Mel 频谱特征 x , 通过基于 Flow 的 Decoder 模块 f_{dec} 将 x 转换为隐变量 z , 这一步不需要任何的文本信息, 并且我们让隐变量 z 满足各向同性的高斯分布 P_z , 然后, 文本编码器 (Text Encoder) 将文本序列 c 通过一系列处理得到文本的高级表示 h , 并通过一个线性层得到高斯分布的统计量 μ 和 δ , 然后让每一个隐变量 z 都遵循文本编码器预测的这些分布之一。我们需要将 z 和 h 对应起来, 这种对应关系定义为对齐, 如果隐变量 z_j 遵循第 i 个文本标记的预测分布 $N(z_j; \mu_i, \delta_i)$, 则我们定义 $A(j) = i$ 。这里的对齐必须是单调的硬对齐, 因此, 给定一个对齐 A , 我们必须满足公式(2):

$$\log P_Z(z; c, \theta, A) = \sum_{j=1}^{T_{mel}} \log N(z_j; \mu_{A(j)}, \sigma_{A(j)}) \quad (2)$$

$$\max_{\theta, A} L(\theta, A) = \log P_X(x|c; \theta, A) \quad (3)$$

由目标函数(3)可以得到我们需要优化的参数为 θ 和对齐矩阵 A , 但若是同时优化 θ 和对齐矩阵 A , 这在计算上是无法实现的, 因为目标参数的解空间太大了, 从计算上是不可行的。

为了解决训练问题, 我们可以把公式(3)改写为公式(4):

$$\max_{\theta} \max_A L(\theta, A) = \log P_X(x|c; \theta, A) \quad (4)$$

这相当于把训练阶段可以分解为两个连续的问题, 在给定参数 θ 的情况下, 使用单调对齐搜索算法找到一个最可能的对齐 A^* , 然后在满足这个对齐矩阵的条件下使用梯度下降算法更新参数 θ 。虽然我们修改后的目标不能保证得到公式(3)的全局最优解, 但它仍然提供了全局解的一个很好的下界。

2.1 文本编码器

文本编码器的作用是将输入的文本转换成文本的高级表示 h , 再经过一个线性映射层就可以得到每一个 h 对应高斯分布的统计量 μ 和 δ 。通常情况下语音合成的数据集都在 10000 条左右, 数据量较小, 想合成流畅自然的语音是一项具有挑战性的任务, 且数据量少的情况下, 将会限制文本编码器的学习能力, 此外, 文本编码器通常需要对文本的含义和语法结构进行建模, 以生成合适的声学特征。当数据量较小时, 模型可能无法获得足够的语言知识来生成高质量的音频。

针对以上问题, 本文将 Bert 模型引入到文本编码器中, 利用 Bert 提取文本序列的深层语义信息, 将 Bert 输出的文本表征和音素嵌入、音调嵌入进行加性融合, 得到最终的嵌入向量, 由于 Bert 的输出是以字为基本单位 (token), 而文本编码器的输入是以音素为 token, 因此在这个层面上两者维度不能, 本文采用重复 (repeat) 操作, 将 Bert 的输出按照每个字对应的音素数量做 repeat, 使维度统一。之后将嵌入向量输入到 6 层 Transformer^[18] 的 Encoder 架构中, 超参数配置同 Transformer-TTS^[19] 的 Encoder 架构, 只是去掉了 Transformer-TTS 的 Encoder 中的预处理网络 (Prenet), Transformer 中的注意力机制允许模型在编码阶段有效地捕捉长距离的依赖关系。对于语音合成任务而言, 理解文本中词与词之间的关系对于生成流畅、自然的语音非常重要。较深层的 Transformer 编码器能更好地处理长句子, 捕捉 token 之间的上下文相关性, 有助于提升语音合成的质量。此外注意力机制还可以计算每个 token 与其他所有 token 之间的注意力权重, 从而生成上下文相关的表示。在语音合成中, 音素在不同上下文中可能有不同的发音或语义, 因此需要能够对其进行上下文相关建模。多层

Transformer 可以产生更丰富、更准确的上下文相关表示，有助于更好地捕捉语义信息。通过使用多层 Transformer 的 Encoder，文本编码器可以进行多次的特征抽取过程。每一层 Encoder 都会逐渐提取出越来越抽象、语义更丰富的特征表示。这种逐层的特征抽取有助于将文本的底层特征与高层语义信息相结合，提供更全面、多样的信息供语音合成模型使用，文本编码器采用架构如图 2 所示：

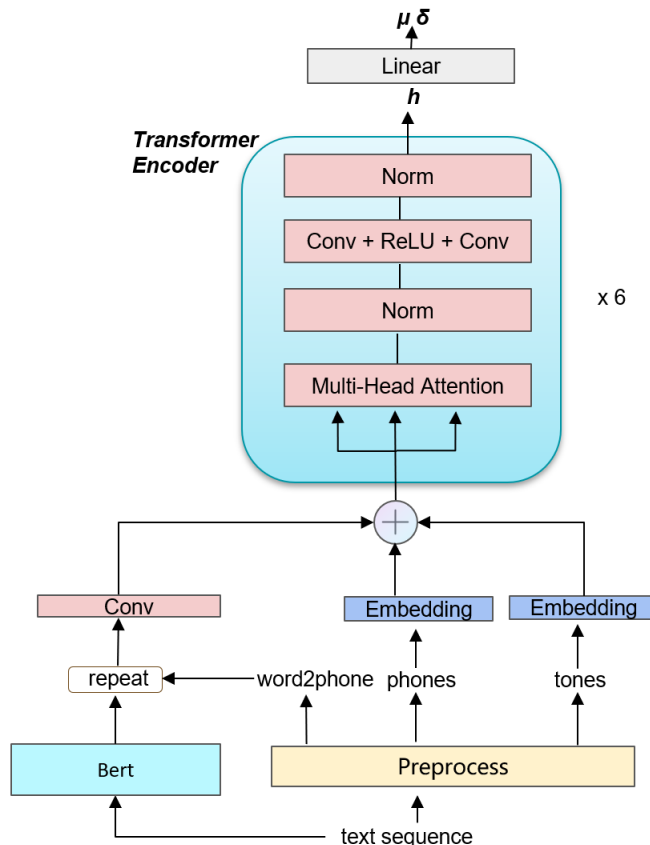


图 2 文本编码器结构图

Fig. 2 Text encoder structure diagram

此外，我们加入了一个额外的声调 embedding 层，通过对待合成文本预处理出来的声调信息作为额外的输入，我们期望模型能学习到不同音素组合的发音区别，使得模型在发音上更为精准。

2.2 基于 Flow 的解码器

解码器是基于 Flow 的非自回归网络，Flow 的基本思想是通过一系列可逆的变换将输入数据映射到一个潜在空间中，并学习该空间中数据的分布，从而实现生成新数据样本的目的。在训练过程中，我们需要高效地将 Mel 频谱帧 y 转换为隐变量 z ，以进行最大似然估计和单调对齐搜索。而在推理过程中，需要将先验分布得到的隐变量 z 高效地反演成 Mel 频谱图分布，以进行并行解码。因此，我们使用 Flow 模型作为解码器，可以在并行中执行前向和反向变换，这有效的提升了语音合成的速度。本文中使用的 Flow 结构是由激活归一化 (Actnorm)、可逆 1×1 卷积 (Invertible 1×1 Convolution) 和仿射耦合层^[12] (Affine Coupling Layer) 组成的块 (Block) 堆叠 12 层而成，结构如图 3 所示：

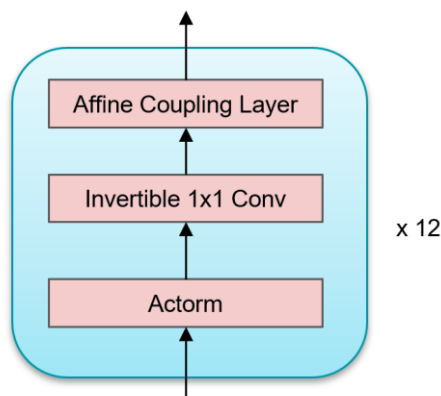


图3 解码器结构图

Fig. 3 Decoder structure diagram

其中仿射耦合层使用的是 WaveGlow^[20]中的仿射耦合层架构，但是去掉了其中的局部条件。使用 Flow 模型作为解码器，能够对复杂的概率分布进行建模，它可以准确的估计样本的概率密度，并且可以提供精确的概率计算，可以方便地与其他语音合成模块进行组合，如文本编码器、声码器等。这种灵活性使得基于 Flow 模型的语音合成系统可以自由地进行模块替换、定制化调整和模型组合，以适应不同的应用需求。

2.3 GAN 训练随机时长预测器

从 VITS 中的随机时长预测器得到启发，基于 Flow 的随机时长预测器在提高合成语音的自然度方面比确定性的时长预测器更为有效，因为对于同一句话，在不同的说话人或是不同的语境下说出，语气音调可能相差较大。然而基于 Flow 的随机时长预测器需要更大的计算量和一些复杂的技术。本文使用 GAN 去训练随机时长预测器，能得到更自然的合成语音。我们使用文本序列的隐变量 h 和高斯噪声 z_d 作为生成器 G 的输入，对于鉴别器 D ，一般的鉴别器的输入都是固定长度的输入，但是在语音合成任务中，输入序列的长度是变化的，为了正确区分可变长度的输入，我们采用了一种基于时间步长的鉴别器，该鉴别器可以鉴别每个 token 的持续时间，这里生成器和鉴别器输入的隐变量 h 都是使用 stop gradient 操作后的 h ，这可以用于限制梯度传播的范围，以避免影响最大似然目标。并且我们使用了两种损失，最小二乘损失和均方误差损失，损失函数如(5) (6) (7)式所示：

$$L_{adv}(D) = E_{(d, z_d, h)} [(D(d, h) - 1)^2 + (D(G(z_d, h), h))^2] \quad (5)$$

$$L_{adv}(G) = E_{(z_d, h)} [(D(G(z_d, h)) - 1)^2] \quad (6)$$

$$L_{mse} = MSE(G(z_d, h), d) \quad (7)$$

由于随机时长预测器的输入要通过上取整 (ceil) 操作才能送入鉴别器，但是 ceil 操作是不可导的，为了能使用梯度下降算法，本文使用了直通估计器 (straight-through estimator, STE) 处理不可导问题。

2.4 声码器

我们使用 HiFi-GAN^[21]作为声码器，与传统的声码器相比，HiFi-GAN 能够生成更加真实、逼真的音频、具有更高的保真度和更快的合成速度，可以说 HiFi-GAN 的出现使得两阶段的语音合成系统的效果登上了新的高峰。它能够捕捉到音频的细微结构、音色和表达，并以高品质还原。HiFi-GAN 还具有较好的语音风格控制能力。通过调整模型参数或输入特征，可以实现对不同的语音风格进行控制，如男声、女声、儿童声等。这使得 HiFi-GAN 成为一种灵活且可定制的声码器，适用于多种语音合成场景和应用需求。且 HiFi-GAN 具有一定的零

样本学习能力。在训练阶段，HiFi-GAN 可以通过大量无标签的音频数据学习到音频的生成规律，而在运行时，它可以根据给定的 Mel 频谱图生成相应的高质量音频，此外，HiFi-GAN 在未见过的说话人上也有着不错的表现。

3 实验和结果分析

3.1 数据准备

本文的实验都基于 LJSpeech-1.1 语音数据集，该数据集为大约 24h 的语音数据，由一位发音人阅读 7 本非小说类书籍的 13100 个短音剪辑组成，音频的采样率为 22050HZ，采样位深为 16bit。使用汉宁窗（Hamming）处理，帧长为 50ms，帧移为 12.5ms，对数幅度 Mel 频谱图被提取为目标语音的表示形式。本文将数据集随机划分成训练集、验证集、测试集，分别为 12500、100、500 条语音片段。

3.2 训练细节

声学模型采用 GlowTTS 作为主体模型结构，Bert 模型使用的是 24 层 Transforme-Encoder 架构且隐藏状态为 1024 维的 DeBERTa V3^[22]，声码器部分则使用的是 HiFi-GAN V1 模型，且训练的数据集同样是采用 LJ Speech-1.1。其中声码器模型和声学模型分开进行训练。模型各个部分的超参数的设置如表 1 所示：

表 1 超参数配置
Tab. 1 Hyperparameter configuration

超参数	Bert-FlowTTS
Bert Hidden Dimension	1024
Embedding Dimension	192
Encoder Blocks	6
Encoder Multi-Head Attention Hidden Dimension	192
Encoder Multi-Head Attention Heads	2
Encoder Conv Kernel Size	3
Encoder Conv Filter Size	768
Encoder Dropout	0.1
SDP Kernel Size	3
SDP Filter Size	192
SDP Dropout	0.5
Decoder Blocks	12
Decoder Invertible 1×1 Conv Groups	80
Decoder Affine Coupling Dilation	1
Decoder Affine Coupling Layers	4
Decoder Affine Coupling Kernel Size	5
Decoder Affine Coupling Filter Size	192
Decoder Dropout	0.05

本文实验采用单 GPU（1 个 4090， 24GB）进行训练，并且使用了混合精度训练的方法加速模型的训练，本文的训练包括语音合成模型，以及随机时长预测器。其中 DeBERTa 模型冻结参数，不参与梯度更新。优化器使用 Adam，其中 $\beta_1=0.9$ ， $\beta_2=0.98$ ， $\varepsilon=1e-9$ ，并且采用预热（warmup）策略更新学习率，warmup steps 设置为 4000 次迭代，预设的学习率为 $1e-1$ ，batch_size 取 32。

3.3 Mel 谱图生成情况

图 4 给出了测试集句子“the resistance to arrest and the attempted shooting of another police officer by the man (Lee Harvey Oswald) subsequently accused of assassinating President Kennedy.”的 Mel 谱图对比。通过对比图 4 左真实的 Mel 谱图和图 4 右模型预测的 Mel 谱图，我们可以发现由模型生成的 Mel 谱图与原始音频的非常相似，高频成分饱满。且可以听出合成的语音感情色彩丰富。这表明我们的模型在处理频谱细节方面有着优秀的表现。

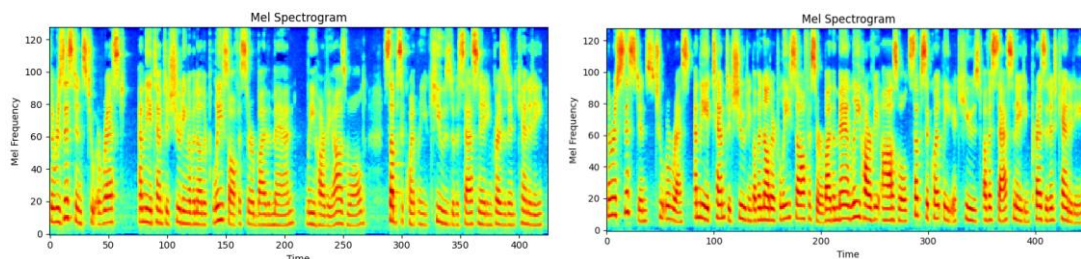


图 4 真实的 Mel 谱图和生成的 Mel 谱图

Fig. 4 Text Real Mel spectrum and generated Mel spectrum

图 5 展示了真实音频（Ground Truth,GT）的音高曲线和合成音频的音高曲线对比，从总体轮廓上可以看出，合成音频的音高曲线基本接近 GT，但是一些细节方处与 GT 有差异，通过对合成音频的分析，发现本文提出来的方案在合成的音频上更加昂扬顿挫，重读特别明显，推测是输入时的音调信息在模型中起到了一定的作用。

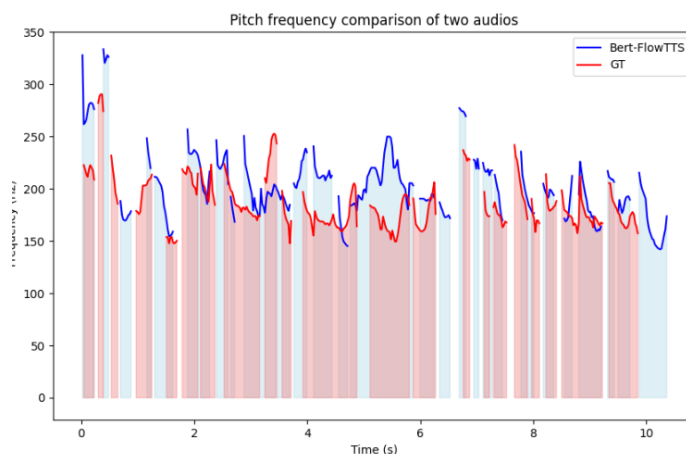


图 5 真实的音高曲线和合成音频的音高曲线

Fig. 4 The fundamental frequency (F0) contours of synthesized speech samples

3.4 音频评估

本文从测试集中挑选了 20 句不同长度的音频作为评估集，比较不同的模型和真实音频的平均意见得分（Mean Option Score, MOS），它的取值范围为 1-5，1 分表示合成的音频质量极差，完全听不懂。5 分表示音频质量高，听的很清晰，和真人发音完全一样。参与打分

的主要人员有 10 人，主要是本校英语专业的学生。实验的详细结果如表 2 所示。

表 2 不同模型间的主观意见得分对比

Tab. 2 MOS comparison between different models

模型	MOS
Tacotron2 + HiFi-GAN	3.77 (± 0.08)
Bert-FlowTTS + HiFi-GAN	4.32 (± 0.07)
GlowTTS + HiFi-GAN	4.14 (± 0.07)
VITS	4.43 (± 0.06)
Ground Truth	4.46 (± 0.06)

5 为了测试加入 Bert 和 GAN 训练随机时长预测器后对语音合成结果的影响，我们进行了消融实验。声码器统一使用的是 HiFi-GAN 模型，同样进行的是 MOS 测评，测评结果如表 3 所示：

表 3 消融实验结果

Tab. 2 Result of ablation experiment

模型	MOS
Bert-FlowTTS(DP)	4.26 (± 0.07)
Bert-FlowTTS(SDP)	4.32 (± 0.07)
FlowTTS (DP)	4.10 (± 0.05)
FlowTTS (SDP)	4.14 (± 0.06)
Ground Truth	4.46 (± 0.06)

10 从表 3 中可以看出，引入 Bert 模型对于语音合成系统的提升较为明显，能有效辅助语音合成模型对发音的建模，提高音频的自然性和流畅性。使用 GAN 训练的随机时长预测器的引入比起固定时长预测器来说提升不大，推测是因为本文的语音合成是针对单个说话人的模型，对于单个说话人来说，随机时长预测器反映的文本特征和频谱帧的对映关系比较单一，不能发挥出随机时长预测器的优势。

15 4 结论

本文通过对 GlowTTS 的文本编码器进行修改，通过嵌入额外的音调信息和 Bert 生成的文本高级表示，使得模型可以挖掘出输入文本中的发音信息以及潜在的句法语义信息来生成表达性语音，然后是使用 GAN 训练的随机时长预测器去代替固定的时长预测器，去解决非自回归语音合成方法的韵律较差问题，最后本文通过主观的方式去验证本文所提出的基于 Flow 的语音合成模型的可行性，提出的方法有效的提高了合成语音的流畅性和自然性。

[参考文献] (References)

- 25 [1] HUNT A J, BLACK A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, 1996, 1: 373-376.
- [2] BLACK A W, TAYLOR P A. Automatically clustering similar units for unit selection in speech synthesis[J]. 1997.

- [3] TOKUDA K, YOSHIMURA T, MASUKO T, et al. Speech parameter generation algorithms for HMM-based speech synthesis[C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). IEEE, 2000, 3: 1315-1318.
- [4] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.
- [5] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4779-4783.
- [6] REN Y, RUAN Y, TAN X, et al. FastSpeech: Fast, robust and controllable text to speech[J]. arXiv preprint arXiv:1905.09263, 2019.
- [7] REN, YI et al. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech[J]. ArXiv abs/2006.04558 (2020): n. pag.
- [8] KIM J, KIM S, KONG J, et al. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search[J]. 2020.
- [9] KIM, JAEHYEON et al. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech[J]. ArXiv abs/2106.06103 (2021): n. pag.
- [10] DINH, LAURENT, et al. NICE: Non-linear Independent Components Estimation[J]. CoRR abs/1410.8516 (2014): n. pag.
- [11] DINH, LAURENT, et al. Density estimation using Real NVP[J]. ArXiv abs/1605.08803 (2016): n. pag.
- [12] KINGMA D P, DHARIWAL P. Glow: Generative Flow with Invertible 1x1 Convolutions[J]. 2018.
- [13] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
- [14] KINGMA, DIEDERIK P. and Max Welling. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114 (2013).
- [15] CRESWELL, ANTONIA, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine 35.1 (2018): 53-65.
- [16] JIA, YE, et al. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS[J]. arXiv preprint arXiv:2103.15060 (2021).
- [17] DEVLIN, JACOB, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. arXiv preprint arXiv:1810.04805 (2018).
- [18] VASWANI, ASHISH, et al. Attention is all you need[C]. Advances in neural information processing systems 30 (2017).
- [19] LI, NAIHAN, et al. Neural speech synthesis with transformer network[C]. Proceedings of the AAAI conference on artificial intelligence. Vol. 33, No. 01. 2019.
- [20] PRENGER R, VALLE R, CATANZARO B. WaveGlow: A Flow-based Generative Network for Speech Synthesis[C]. 2018.
- [21] SU, JIAQI, et al. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks[C]. ArXiv abs/2006.05694 (2020): n. pag.
- [22] HE, PENGCHENG, et al. Deberta: Decoding-enhanced bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654 (2020).