doi:10.3969/j.issn.1672-1292.2023.04.005

轻量且基频可预测的端到端语音合成系统

梁　婷[1],艾斯卡尔·艾木都拉[1],刘　煌[2],徐　颖[2]

(1.新疆大学信息科学与工程学院,新疆 乌鲁木齐 830046)
(2.上海格子互动信息技术有限公司,上海 200000)

[**摘要**]　提出了一种轻量级的基频可控的完全端到端的语音合成模型. 该模型基于目前最流行的完全的端到端的语音合成模型 VITS 做出了三处改进,使得合成的语音韵律感更强,从而提高语音合成的自然度和表现力,同时提高发音的准确性和推理速度. 首先,引入帧先验网络得到细粒度的均值方差表示,且引入音素预测器和 CTC loss 以提高发音的稳定性. 其次,在模型中使用音素真实时长对齐文本和音频帧,并且加入 F0 预测器,增强语音的韵律感. 另外,用多频带和短时傅立叶变换替换原始模型中的 Decoder,有效提高了模型的推理速度. 最后,使用 MOS 测试和 RTF 作为实验主观和客观的评判标准. 实验证明,模型在音频自然度和表现力方面提高了至少 5%,且相比原始 VITS 推理速度提高了 3 倍.

[**关键词**]　端到端语音合成,韵律预测,逆快速傅立叶变换,变分字编码器,流,多频带
[**中图分类号**]TP391.1　[**文献标志码**]A　[**文章编号**]1672-1292(2023)04-0037-06

A Lightweight End-to-End Speech Synthesis System with Pitch Prediction

Liang Ting[1], Askar Hamdulla[1], Liu Huang[2], Xu Ying[2]

(1.School of Information Science and Engineering, Xinjiang University, Wulumuqi 830046, China)
(2.Shanghai GERZZ Interactive Information Technology Co., Ltd, Shanghai 200000, China)

Abstract: This paper proposes a lightweight end-to-end speech synthesis model with pitch prediction. The model in this paper is based on VITS, an end-to-end speech generation model which adopts VAE-based posterior encoder augmented with normalizing flow based prior encoder and adversarial decoder, and three improvements are made to make the synthesized speech more rhythmical and more stable in a more efficient way. To be more specific. Firstly, to improve the accuracy of pronunciation and naturalness of speech, we introduce a length regulator and a frame prior network to get the frame-level mean and variance on acoustic features, modeling the rich acoustic variation in speech, and phone predictor and CTC loss are introduced to improve the stability of pronunciation. Secondly, the ground truth duration of phonemes is used for alignment of text and frame in the model, and F0 predictor is added to enhance the sense of rhythm of speech. Thirdly, the decoder in the original VITS model with multi-band generation and inverse short-time Fourier transform, which effectively improves the inference speed of the model. Experiments show that the proposed model greatly improves the naturalness and expressiveness by 5% from the MOS(mean opinion score) value and improves the inference speed by 3 times from RTF(real-time factor) compared with the original VITS.

Key words: end-to-end speech synthesis, prosodic prediction, ISTFT, VAE, flow, sub-band

随着机器学习的迅猛发展,基于深度学习的语音合成逐渐流行. 端到端的语音合成技术已经极大地简化了语音合成的模型的复杂度和流程,并且基于端到端的语音合成技术在某些领域已经达到了真人水平. 起初端到端的语音合成模型都分为两个阶段(如 FastSpeech[1],Tacotron 系列[2-3]),先是声学模型由文本预测出频谱图,再由声码器将频谱图转换为对应的采样点,但两个阶段的模型增加了模型训练的复杂度,并且由声学模型的预测可能和真实值会有一定的失真,导致声码器合成的音频和真实值之间的误差较大. 所以完全的端到端的模型逐渐流行(FastSpeech2s[4],EATS[5],Glow-WaveGAN[6]和 VITS). 其中 VITS 是首个自然度超过两阶段模型的完全的端到端的语音合成模型. VITS 把 Flow[7]和 VAE[8]应用于端到端的语音合成模型,训练非常简单,且多说话人模型也可以达到较高的自然度. 但 VITS 的模型训练收敛较

收稿日期:2023-04-24.
通讯作者:艾斯卡尔·艾木都拉,博士,教授,主要研究方向:语音合成,自然语言处理和语音识别等. E-mail:askar@xju.edu.cn

慢,且全局信息的学习能力较弱,如因为采用了随机的时长预测器,导致韵律和风格的合成较平淡且模型在合成时会有明显错音和变调的问题. 因此本文针对这些问题作出了 3 个方面的改进.

具体而言,首先,为了提高发音的准确性,我们引入帧先验网络得到细粒度的均值方差表示,且引入音素预测器和 CTC loss[9]来提高发音的稳定性. 其次,在模型中使用音素真实时长对齐文本和音频帧,并且加入 F0 预测器,增强语音的韵律感. 另外,用多频带生成和逆短时傅立叶变换替换原始模型中的 Decoder,有效提高了模型的推理速度. 实验证明,本文提出的模型极大减少了发音的错误率,提高了 5%的表现力和自然度,且相比原始 VITS 推理速度提高了 4 倍.

1　相关工作

基于端到端的学习方式给语音合成的发展带来了越来越大的影响力,引起了学术界和工业界的极高的重视. 基于端到端的语音合成系统,可以直接从文本层面预测到接近语音层面的特征,而不再需要人手工进行复杂的特征设计. 起初端到端的语音合成模型都是分为两个阶段,先由声学模型将文本映射到频谱,再由声码器将频谱转化为音频采样点. 但这种两个阶段的语音合成系统,训练流程复杂,并且由声学模型的预测可能和真实值会有一定的失真,导致声码器合成的音频和真实值之间的误差较大. 2021 年 KAKAO 公司和 KAIST 公司提出了单阶段的完全端到端的语音合成模型 VITS,是首个自然度超过两阶段架构的完全的端到端语音合成模型. 该模型得益于图像领域中把 Flow 引入 VAE 提升生成效果的研究,成功把 Flow-VAE 应用到了完全的端到端语音合成任务中. 它的高清音质和训练简单的特点,使 VITS 在学术届和工业界广受欢迎. 但是 VITS 的训练推理速度较慢,且全局信息的学习能力较弱,如韵律和风格的合成较平淡,时长较弱. 针对 VITS 的推理速度问题,已有实验提出应用多频带逆快速傅立叶变换解决[9-11],这种方法利用神经网络的稀疏性和单一的共享网络来生成所有的子频带信号,有效地提高了推理速度. 对于增强模型的韵律感,可以在模型中加入 F0 预测器[12-13]. 本文将从推理速度,韵律增强和发音稳定 3 个方面对 VITS 做了改进,提出了一种高效的且可以进行韵律预测的端到端语音合成系统.

2　高速且带基频预测的完全端到端语音合成系统

本文提出的方法:先验编码器,后验编码器和解码器. 训练和推理阶段的总体框架如图 1 和图 2 所示. 和 VITS 一样,本文模型也是以文本作为输入且以波形作为输出的完全的端到端的模型. 然而两个模型还是有一定的差异性. VITS 使用动态对齐搜索(monotic alignment search,MAS)作文本和频谱之间的对齐训练随机时长预测器,而在实验中,直接使用真实的时长作为时长预测器的训练目标,真实时长由 MFA

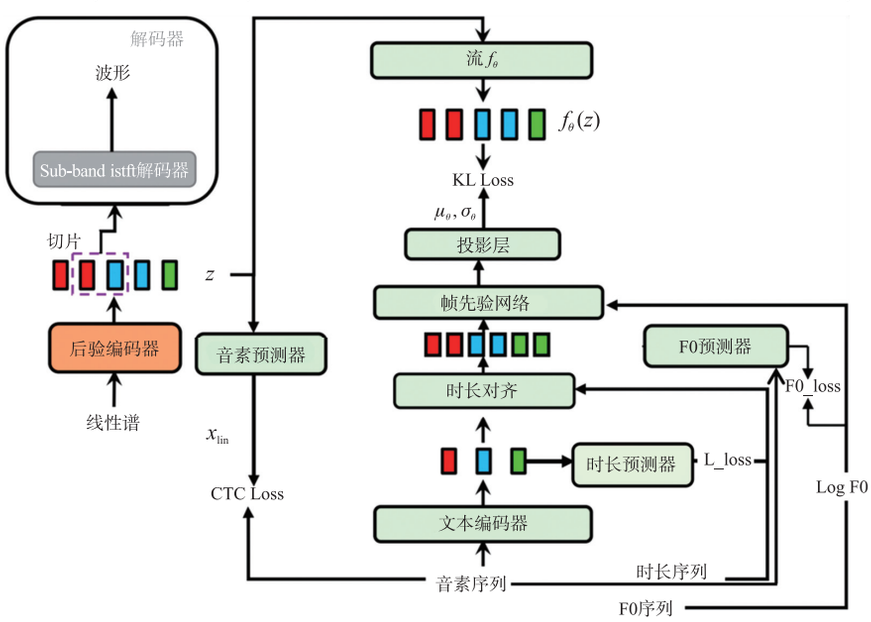

图 1　训练时模型结构

Fig. 1　Model architecture in training

(montreal forced alignment)提取,并且加入帧先验网络得到帧级别的均值方差. 第二,VITS 的 Decoder 结构基于 Hifi-GAN 的 Vocoder,该结构利用转置卷机网络对输入的声学特征进行上采样,该 Vocoder 部分占据了 96%以上的推理时间[14],所以文本采用简单的 ISTFT 网络替换原有声码器,提高模型的推理速度. 第三,我们在模型中引入 F0 预测器,且使用真实的 F0 作为训练目标,提高合成语音的韵律感.

该模型中的后验编码器只在训练阶段使用,推理阶段不用. 对于多说话人模型,说话人信息被添加到文本编码器,长度预测器,F0 预测器,解码器模块中,以使模型更好地学习到说话人的发音特征.

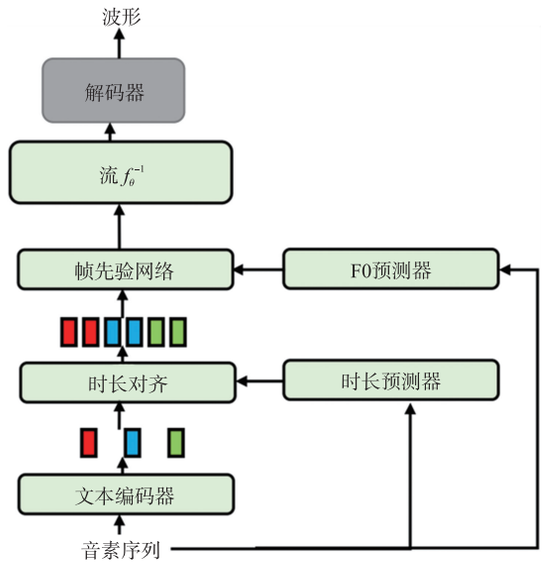

图 2　推理时模型结构

Fig. 2　Model architecture in inference

2.1　先验编码器

先验编码器从音素序列中预测先验分布. 先验编码器中除了原始 VITS 中有的文本编码器,还更改帧序列和音素序列对齐的方法,使用音素真实的时长对齐而不是像原始 VITS 中采用 MSA 机制进行对齐,另外增加了 F0 预测器提高语音的表达力,而且为了提高发音的稳定性,改善 VITS 变调错音的问题,加入了帧先验网络和 CTC(connectionist temporal classification)loss.

2.1.1　文本编码器

和原始 VITS 一样,文本编码器模块输入是文本,使用基于自注意力机制的模块生成音素级别的表征,每个音素被编码为 192 维度的向量.

2.1.2　时长预测器

时长预测器在推理阶段提供每个音素的时长. 由两层 256 单元的双向 LSTM 组成. 原始 VITS 在训练随机时长预测器时,使用 MAS 得到每个音素的目标时长. 由于 VITS 的时长的预测是随机的,且由 MAS 对齐得到的音素的时长与真实时长有一定的差距. 所以本文使用一个两层 256 单元的双向 LSTM 模块作为时长预测器,且用每个音素真实的时长作为时长预测器的训练目标. 在训练时,会计算真实时长 d_{GT}和长预测器预测的时长 $\hat{d}$ 的均方差损失 L_{dur},其计算公式如式(1)所示. 真实时长由 MFA 获取,并且添加时长对齐(length regulator,LR)模块把音素级别的表征扩展到帧级别. 该时长对齐模块的输入在训练时来自每个音素的真实时长,在推理阶段来自时长预测器的预测输出.

$$L_{dur} = \| d_{GT} - \hat{d} \|_2. \tag{1}$$

2.1.3　F0 预测器和帧先验网络

在 VITS 的训练阶段,文本编码器提取音素级别的文本信息,作为提取潜变量 z 的先验知识. 但是每个音素的声学特征的变化是非常丰富的,所以仅仅用音素级别的均值方差表示帧序列远远不够. 所以文本参考 VISinger[14],在模型中添加帧先验网络,在扩帧之后计算均值和方差. 帧先验网络对帧序列后处理得到帧级别的均值方差,帧先验网络由一系列一维卷基层组成. 另外,VITS 在韵律方面合成较平,缺少抑扬顿挫感. 所以在模型中添加 F0 预测器. 和 VISinger 类似,使用 F0 预测器指导帧先验网络的学习. F0 预测器由多个 FFT 块组成. 帧先验网络的输入包括经时长扩帧后的帧序列和基频值. 训练阶段将真实的基频值送入帧先验网络,同时真实基频值也用来指导 F0 预测器的学习. 推理阶段帧先验网络的基频值来自 F0 预测器的预测输出. 真实基频值的提取参考 Fastpitch 的方法. 帧先验网络的输入来自音素序列,在训练阶段 F0 预测器的输出 $\hat{f}0$ 和真实基频 $f0_{GT}$计算均方差损失 L_{f0}如式(2)所示. 由帧先验网络得到帧级别的均值和方差,再采样出隐变量 z. 然后再使用标准化流 f 将 z 变成一个更复杂的分布. 标准化流和 VITS 类似.

$$L_{f0} = \| f0_{GT} - f0 \|_2. \tag{2}$$

2.2　后验编码器

与 VITS 类似,后验编码器以线性谱作为输入,提取后验分布 $p(z|y)$ 的均值和方差,得到潜变量 z.

2.3 Decoder

Decoder 根据提取的中间变量 z 生成语音波形. VITS 使用 Hifi-GAN 结构作为解码器. 由文献[14]可知,VITS 的 Decoder 模块占据了 VITS 模型的 96%的推理速度,所以为了提高模型的推理速度,参考文献[14]使用多频带的 ISTFT 替换 VITS 中 Decoder. 具体而言,先对 VAE 的隐变量进行 S 倍的上采样,然后投影到 N 个子带信号的幅值和相位变量. 接着对幅值和相位变量进行 ISTFT 生成每个子信号,这些子带信号再经过上采样到原始信号的采样率,然后使用一个可训练的合成滤波器组将这 N 个子带信号整合为全带信号. 可训练的合成滤波器可以让模型以数据驱动的方式分解语音波形,提高推理速度的同时提高合成语音的质量. Decoder 的结构如图 3 所示.

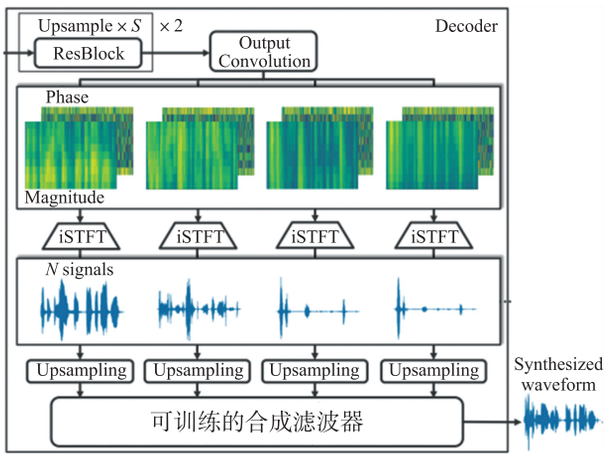

图 3　多频带 ISTFT 解码器

Fig. 3　Sub-band ISTFT decoder

2.4 模型最终 loss

最终训练 loss 可以表示为:

$$L=\lambda_{\text{recon}} * L_{\text{recon}}+L_{\text{kl}}+L_{\text{adv}}(G)+\lambda_{fm} * L_{fm}(G)+L_{f0}+L_{\text{ctc}}+L_{\text{dur}}. \tag{3}$$

式中,L_{kl}是 Kl 散度和 VITS 类似,L_{recon}代表真实的 mel 谱和生成的 mel 谱之间的 $L1$ loss,此时 λ_{recon}为 45. $L_{\text{adv}}(G)$是 GAN 的生成器 loss. L_{fm}是特征匹配 loss,用于提高训练的稳定性,这里的 λ_{fm}为 2. L_{f0}是 F0 预测器生成的 $f0$ 值和真实 $f0$ 之间的均方误差(mean squared error,MSE) loss. L_{dur}是时长预测器的输出和真实时长之间的 MSE loss. L_{ctc}是音素的 CTC loss.

3 实验与分析

3.1 数据集

使用两个男生两个女生一共大约 20 h 的中文数据训练了多说话人的模型. 为了便于训练,将这些音频切分成 12 s 左右的短句子,大约 6 千条,其中的一百条用于验证,两百条用于测试. 实验中所有的音频被采样到 24 kHz 和量化到 16 bit,音素的真实时长由 MFA 获取.

3.2 模型配置

共训练了以下模型作对比.

- V1:原始 VITS.
- V2:在 VITS 的基础上增加了帧先验网络和音素时长对齐模块(length regular).
- V3:在 V2 的基础上增加了 F0 预测器.
- V4:在 V3 的基础上增加了 CTC loss.
- V5:文本提出的模型,在 V4 的基础上替换了 Decoder 使用子频带 ISTFT 解码器.

在提出的模型中,时长预测器包含两层 256 个单元的双向 LSTM 网络,F0 预测器包含 6 个 FFT block,帧先验网络包含 4 个 FFT block. 子频带 ISTFT 解码器中把全频带分为 4 个子频带(sub-band),生成器中两个残差网络再分别进行 4 倍的上采样,且基于卷机网络的可训练的合成滤波器的 kernel size 设置为 63. 所有模型都用一样的数据集,其中带 LR 的模型,其音素的真实时长来自 MFA,带 F0 预测器的模型,其 F0 的真实值通过 parselmouth 提取. 训练时,VITS 用到的线性谱由 STFT 获取,并作为后验编码器的输入. 每个实验都是在单张 V100 上训练的,batch 设置为 32. 所有的模型都训练 600k.

3.3 实验结果

为了测试这些模型的性能,语音的音质、表现力、自然度以及音色相似度的 MOS 测试作为评判标准. 用上文提到的 200 句测试集中随机挑选了 100 句分别测试上述 6 个模型中的 4 个发音人. 并邀请 30 位测评者在语音的音质、表现力、自然度以及音色相似度 4 个维度上面对音频进行 1 到 5 的打分. 这 30 位

测评者中,有 10 位是高级语音算法工程师,10 位是专业的配音演员,其余 10 位是语言学专业的同学. 另外,为了使模型得到客观的比较,还比较了这 5 个模型的参数量,并针对这 100 句测试语句,在 CPU 上计算了这 5 个模型的 RTF(real-time factor)来比较模型的推理速度,RTF 定义为一句话的合成时间和这句话时长的比值,这里取 100 句的平均值,结果如表 1 所示.

表 1　测试结果

Table 1　The results of test

模型	音质	表现力	自然度	音色相似度	params	RTF
Recording	4.29	4.36	4.50	—	—	—
V0	4.07	4.10	4.16	4.12	35.60M	0.221
V1	4.16	4.13	4.10	4.27	39.10M	0.216
V2	4.13	4.10	4.12	4.25	47.52M	0.223
$V2_{100}$	4.11	4.06	4.06	4.20	47.52M	0.225
V3	4.20	4.25	4.20	4.30	53.80M	0.235
V4	4.23	4.35	4.39	4.32	55.96M	0.232
V5	4.20	4.32	4.38	4.32	54.37M	0.078

表 1 中的 $V2_{100}$是 V2 模型训练了 100k 的测试结果,可以看到和 600k 的结果相差不多,说明用真实时长对齐可以加快模型的收敛速度. 另外,加了 F0 后的模型在表现力和自然度上面有较大的提高,说明 F0 预测器可以有效提高模型的韵律感. 另外,使用了 CTC loss 的模型的自然度最高,且并没有降低模型的表现力和音质,说明 CTC loss 有助于提高模型的发音稳定性. 从 RTF 指标可以看出,在使用了 multi-band 的 Decoder 后,可以将模型的推理速度提高三倍多,且并不会影响模型的音质和表现力等其他指标.

4　结论

本文提出了一个完全的端到端的高速且基频可控语音合成模型,结合完全端到端的框架,融入时长预测器和 F0 预测器,以提高合成的语音的韵律感,并且实验结果显示引入的 CTC loss 可以有效提高发音的稳定性以提高模型的自然度,另外使用了多频带逆快速傅立叶变换 Decoder,在不降低表现力等指标的基础上将模型的推理速度提高了 3 倍多. 但是模型相比较原始的 VITS,参数量较大,后续会继续研究轻量级的高表现力的端到端语音合成模型.

[参考文献](References)

[1] REN Y,RUAN Y J,TAN X,et al. Fastspeech:Fast,robust and controllable text to speech[C]//33rd Conference on Neural Information Processing Systems. Vancouver,Canada,2019.

[2] WANG Y,SKERRY-RYAN R J,STANTON D,et al. Tacotron:Towards end-to-end speech synthesis[J/OL]. arXiv Preprint arXiv:1703.10135,2017.

[3] 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). Calgary, AB, Canada: IEEE,2018.

[4] REN Y,HU C X,TAN X,et al. FastSpeech 2:Fast and high-quality end-to-end text to speech[J/OL]. arXiv Preprint arXiv: 2006.04558,2020.

[5] JEFF D, SANDER D, MIKOŁAJ B, et al. End-to-end adversarial text-to-speech[J/OL]. arXiv Preprint arXiv: 2006.03575,2020.

[6] CONG J,YANG S,XIE L,et al. Glow-wavegan:Learning speech representations from gan based variational auto-encoder for high fifidelity flflow-based speech synthesis[J/OL]. arXiv Preprint arXiv:2016.10831,2021.

[7] REZENDE D J,MOHAMED S. Variational inference with normalizing flflows[J/OL]. arXiv Preprint arXiv:1505.05770, 2015.

[8] KINGMA D P,WELLING M. Auto-encoding variational bayes[J/OL]. arXiv Preprint arXiv:1312.6114,2013.

[9] YANG G,YANG S,LIU K,et al. Multi-band MelGAN:Faster waveform generation for high-quality text-to-speech[J/OL]. arXiv Preprint arXiv:2005.051006,2021.

[10] YU C,LU H,HU N,et al. DurIAN:Duration informed attention network for speech synthesis[J/OL]. arXiv Preprint arXiv:1909.01700,2019.

[11] CUI Y,WANG X,HE L,et al. An effificient sub-band linear prediction for LPCNet-based neural synthesis[C]//Interspeech 2020. Shanghai,China,2022:3555-3559.

[12] ZHANG Y M,CONG J,XUE H Y,et al. VISinger:Variational inference with adversarial learning for end-to-end singing voice synthesis[J/OL]. arXiv Preprint arXiv:2110.08813,2021.

[13] JU Y,KIM I,YANG H,et al. TriniTTS:Pitch-controllable end-to-end TTS without external aligner[C]//Interspeech 2022. Incheon,Korea,2022:16-20.

[14] KAWAMURA M,SHIRAHATA Y,YAMAMOTO R,et al. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform[J/OL]. arXiv Preprint arXiv:2210.15975,2022.

[责任编辑:陈 庆]