ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 2,Mar. 2024,pp. 406－415
DOI:10. 16337/j. 1004-9037. 2024. 02. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种融合激励和颤音建模的端到端歌唱合成方法

周　骁[1,2]，胡亚军[1]，潘　嘉[1]，胡国平[1]，凌震华[2]

(1. 科大讯飞股份有限公司,合肥 230088;2. 中国科学技术大学信息科学技术学院,合肥 230026)

摘　要：近年来,歌唱合成技术快速发展,基于变分推理和流模型的端到端歌唱合成(VISinger)成为主流,但其在效果上和真人仍有一定差距,主要体现在合成歌声中的音高听感不连续、颤音合成不佳及发音不稳定等。为此,本文针对性地提出了一系列改进方法:针对基频稳定性问题,提出在解码器中增加激励模块,将基频信息以激励信号的形式显式提供给解码器;针对颤音合成不自然问题,增加颤音预测模块,通过流式模型和变分数据增强,显式对歌声中的颤音进行建模;进一步在先验网络中增加ReZero策略。实验结果显示,增加激励信号能提升合成基频的稳定性,颤音建模对颤音的恢复有显著提升作用,ReZero策略对训练速度和发音稳定性有一定提升。主观测听中,本文提出的模型在歌唱合成自然度上相比VISinger有显著优势,平均意见分(Mean opinion score, MOS)达到3.95,对比两阶段建模方法DiffSinger＋HiFiGAN也有明显优势,证明了本文所提方法的有效性。

关键词：端到端歌唱合成;神经网络;颤音建模;归一化流;变分数据增强

中图分类号：TP391　　**文献标志码**:A

An End-to-End Singing Voice Synthesis Method with Excitation and Vibrato Modeling

ZHOU Xiao[1,2], HU Yajun[1], PAN Jia[1], HU Guoping[1], LING Zhenhua[2]

(1. iFLYTEK Co. Ltd., Hefei 230088, China; 2. School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: In recent years, singing voice synthesis technology has developed rapidly, and end-to-end singing voice synthesis (VISinger) based on variational inference and normalizing flow has become mainstream. But there is still a certain gap between its effect and the sound quality of real persons, which is mainly reflected in the discontinuous hearing of pitch, poor synthesis of vibrato, and unstable articulation in the synthesized singing voice. We propose three main improvements. Firstly, to address the problem of fundamental frequency stability, we propose to add an excitation module in the decoder to explicitly provide the fundamental frequency information to the decoder in the form of an excitation signal; secondly, to address the problem of unnatural vibrato synthesis, we add a vibrato prediction module to explicitly model the vibrato in the song using flow with variational data augmentation; thirdly, we further add a ReZero strategy to the frame prior network. Experimental results show that increasing the excitation signal can improve the stability of the synthesized fundamental frequency, the vibrato modeling has a significant enhancement effect on the recovery of vibrato, and the ReZero strategy has a certain improvement on the training speed and articulation stability. Subjective evaluation demonstrates that the proposed model has a

基金项目:科技创新2030——“新一代人工智能”重大项目(2020AAA0103600)。
收稿日期:2022-12-13;**修订日期**:2023-01-20

significant advantage over VISinger in the naturalness of singing voice synthesis, with mean opinion score (MOS) reaching 3.95, and also has a significant advantage over the two-stage modeling method DiffSinger+HiFiGAN, proving the effectiveness of the proposed method.

Key words: end-to-end singing voice synthesis; neural networks; vibrato modeling; normalizing flow; variational data augmentation

引　言

歌唱合成(Singing voice synthesis, SVS)是将人工智能与音乐相结合的技术,旨在通过歌词和乐谱生成高质量的歌声,实现机器像人一样歌唱,并最终实现情感和表现力。该技术可以用于娱乐交互,创造虚拟歌手。类比于文本到语音合成(Text to speech, TTS)技术,歌唱合成也是合成技术的一个分支,且两者基础概念和方法有许多相似之处。TTS技术是将输入的文本转换为音频,而SVS则是将输入的歌词文本和曲谱的旋律转换为音频。歌唱合成中的歌词部分需要保证输出音频中含有歌词信息,而旋律部分需要保证输出音频中含有与曲谱匹配的音符音高和音符时值,同时需要具有表现力以体现歌曲的情感。早期的SVS模型基于隐马尔可夫模型(Hidden Markov model, HMM)[1],然而基于HMM的系统会出现过度平滑,从而降低合成歌声的自然度。后来基于深度神经网络(Deep neural networks, DNNs)[2]的SVS系统也被提出并证明了它们优于基于HMM的系统。最近,一些神经网络模型如HiFiSinger[3]、ByteSing[4]、Sinsy[5]、DiffSinger[6]被提出并显著提高了合成歌唱的自然度。HiFiSinger将覆盖更宽频带的48 kHz采样率引入SVS,并将梅尔频谱分成多个子带,每个子带使用一个基于生成对抗网络(Generative adversarial network, GAN)的判别器进行真假判别,可合成出高保真的歌声;ByteSing将基于神经网络的声学模型、时长模型和声码器组合成一个系统,合成出的韵律可媲美录音;Sinsy在声学建模中考虑了颤音,使用对音高鲁棒的声码器以及自动音高校正技术,即使训练数据有音高标记错误,也可以用正确的音高合成歌唱;DiffSinger使用了基于去噪扩散概率模型(Denoising diffusion probabilistic model, DDPM)的声学模型,这种声学模型具有强大的建模能力,其效果超过了基于GAN的声学模型,在其他领域也得到了广泛的应用。不过由于它们合成的音质与真人演唱依然有一定差距,一体化模型VISinger[7]被提出用于SVS,它使用变分自编码器(Variational autoencoder, VAE)将声学模型和声码器合并到统一的框架中,并展示了相对于两阶段模型在合成歌唱上的巨大优势。然而VISinger依然存在基频(Fundamental frequency, F0)不连续、颤音合成不自然的问题。一些工作试图显式地在声码器中引入基频作为额外输入以预测波形,例如神经源滤波器(Neural source filter, NSF)模型中的激励模块[8]生成具有指定基频的正弦激励,这有助于生成具有谐波结构的波形。另一些工作则试图将颤音(Vibrato)引入歌唱的建模中[5,9-10],因为良好的颤音与歌唱的高表现力密切相关,例如基于正弦的参数建模(预测颤音的幅度和颤音的频率)[5]、基于基频差分的颤音建模[5]或者基于基频差分谱的颤音建模[9]等。

本文对VISinger进行改进,名为VISinging,创新点有

(1) 针对VISinger合成音频中浊音段基频和高次谐波不稳定、有间断的问题,通过激励模块生成激励信号让解码器生成具有谐波结构的波形;

(2) 针对VISinger没有显式建模颤音的问题,提出了颤音预测模块用以生成叠加在音高上的颤音;

(3) 进一步在帧级先验网络的残差模块中使用ReZero策略[11]。

在一个普通话女声歌唱音库上的实验表明,所提出的方法在自然度平均意见分(Mean opinion score, MOS)上显著优于一体化模型VISinger以及两阶段模型DiffSinger+HiFiGAN。实验发现,该方

法可进一步降低先验分布和后验分布存在的KL(Kullback-Leibler)损失、提升训练速度和减少发音错误。

1 端到端歌唱合成方法

1.1 背景介绍

使用基于对抗学习与变分推断的一体化模型是一种合成音频的新颖方法[7,12],它利用条件变分自编码器(Conditional VAE, CVAE)在潜空间中采样,该空间的任何点都可以被重建为波形。一种基于CVAE的SVS模型VISinger由3部分组成,包含1个后验编码器、1个先验编码器和1个解码器,如图1所示。

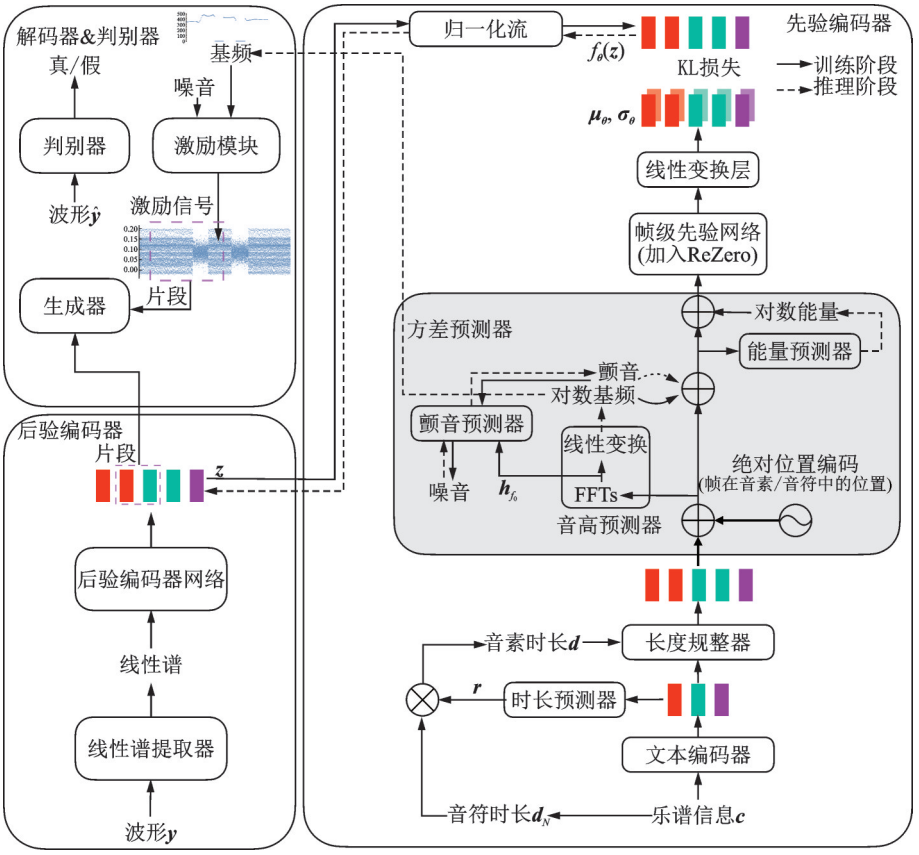

图1 VISinging模型的整体框图

Fig.1 Overall framework of VISinging model

对于CVAE来说,后验编码器从波形$\boldsymbol{y}$中提取隐变量$\boldsymbol{z}$,解码器根据$\boldsymbol{z}$重建波形$\hat{\boldsymbol{y}}$,即

$$\boldsymbol{z}=\mathrm{Enc}(\boldsymbol{y})\sim q(\boldsymbol{z}|\boldsymbol{y}) \tag{1}$$

$$\hat{\boldsymbol{y}}=\mathrm{Dec}(\boldsymbol{z})\sim p(\boldsymbol{y}|\boldsymbol{z}) \tag{2}$$

在CVAE中,给定音乐乐谱$\boldsymbol{c}$作为条件,先验编码器生成隐变量$\boldsymbol{z}$的先验分布$p(\boldsymbol{z}|\boldsymbol{c})$。为了使CVAE的证据下限$\lg_{p_\theta}(\boldsymbol{y}|\boldsymbol{c})$最大化,$\theta$为模型参数,需要最小化训练损失$L_{\mathrm{VAE}}$,它包括以下两种损失

$$L_{\mathrm{VAE}}=L_{\mathrm{recon}}+D_{\mathrm{KL}} \tag{3}$$

式中:L_{recon}是重建损失$-\lg\left(\boldsymbol{y}|\boldsymbol{z}\right)$,使用真实波形与预测波形的梅尔频谱的$L_1$范数来实现;$D_{\text{KL}}$通过计算后验分布$q(\boldsymbol{z}|\boldsymbol{y})$与先验分布$p(\boldsymbol{z}|\boldsymbol{c})$之间的KL散度得到,其中$\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})$,$\boldsymbol{\phi}$为模型参数。后验编码器将波形$\boldsymbol{y}$编码为隐变量$\boldsymbol{z}$,其中波形的线性谱被用作后验编码器网络的输入,以预测后验高斯分布$p(\boldsymbol{z}|\boldsymbol{y})$的均值和方差,接着只要通过使用重参数化技巧从$p(\boldsymbol{z}|\boldsymbol{y})$中采样,即可得到隐变量$\boldsymbol{z}$。

给定乐谱$\boldsymbol{c}$作为条件,先验编码器模块用于产生CVAE中的先验分布$p(\boldsymbol{z}|\boldsymbol{c})$。乐谱输入$\boldsymbol{c}$包括音素类别、音符音高、音符时长、位置和延长音等嵌入向量的相加,其中音符时长嵌入向量使用绝对位置编码得到,音符的音高和音符的时长被上采样为音素级的特征,上采样过程是简单的音素重复,上采样倍数是当前音符中的音素数量。位置向量是一个布尔向量,表示音素在音符中的位置。延长音向量也是布尔向量,表示当前音素是否是前一个音素的重复。文本编码器将乐谱$\boldsymbol{c}$作为输入并产生一个音素级表征,接着时长预测器使用音符归一化[7]方式根据音符时长预测音素时长,再使用长度规整器将音素级表征上采样到帧级表征$\boldsymbol{h}_{\text{text}}$,接着采用方差适配器(Variance adapter)[13]对输入增加方差信息,例如使用基频或能量,然后帧级先验网络(Frame prior network)[7]将帧级表征进一步处理成帧级先验高斯分布,其均值为$\boldsymbol{\mu}_\theta$,方差为$\boldsymbol{\sigma}_\theta$,再使用归一化流对其进行转化,进一步提高先验分布的表达能力,与后验分布更加适配。本文提出了3个创新模块,如图2所示。

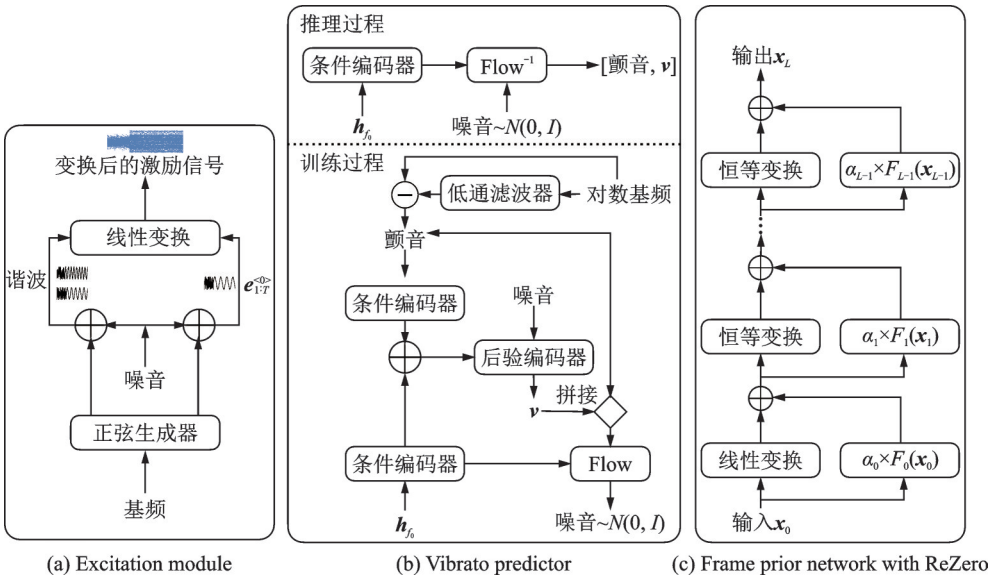

图2　本文提出的3个创新模块

Fig.2　Three innovative modules proposed in this paper

1.2　融合激励模块的解码器

解码器由基于HiFi-GAN的生成器、判别器和激励模块组成。方差适配器中音高预测器的输出是对数基频,该基频及其谐波作为激励模块[8]的输入。激励模块的框图如图2(a)所示,它会生成的浊音段是正弦与噪声的叠加、清音段是噪声的激励信号,再将经线性变换后的激励信号提供给生成器预测波形。具体来说,激励模块会根据基频$\boldsymbol{f}_{1:T}$及其谐波频率产生一个正弦激励信号$\boldsymbol{e}_{1:T}$。假设第t帧的瞬时频率为$f_t\in\mathbf{R}^+$,对于清音帧$f_t=0$。激励信号为$\boldsymbol{e}_{1:T}^{\mathrm{T}}=\left\{\boldsymbol{e}_{1:T}^{\langle 0\rangle},\boldsymbol{e}_{1:T}^{\langle 1\rangle},\cdots,\boldsymbol{e}_{1:T}^{\langle h\rangle},\cdots,\boldsymbol{e}_{1:T}^{\langle H\rangle}\right\}$(直到$H$次谐波),其中第$h$次谐波对应的激励信号$\boldsymbol{e}_{1:T}^{\langle h\rangle}$为

$$e_t^{\langle h\rangle}=\begin{cases}\alpha\sin\left(\sum_{k=1}^{t}2\pi\dfrac{(h+1)\times f_k}{f_s}+\phi\right)+n_t & f_t>0\\ \dfrac{1}{3\sigma}n_t & f_t=0\end{cases}\tag{4}$$

式中：$n_t\sim\mathcal{N}(0,\sigma^2)$为高斯噪声，$\phi\in[-\pi,\pi]$为在训练和推理阶段随机产生的初始相位，$f_s$为波形的采样率。对应于基频和$H$个谐波频率的激励信号（共$H+1$个）通过线性变换层后，使用tanh函数作激活函数得到变换后的激励信号。超参数被设置为$\sigma=0.003$，$\alpha=0.1$，$H=8$。

与VITS生成器相同，VISinging的生成器也只关注输入的一个切片序列。生成器使用HiFi-GAN V1结构[14]，它由上采样层与多感受野融合（Multi-receptive field fusion，MRF）模块多次堆叠而成，其中MRF由交替使用带洞卷积和普通卷积的网络构成，目的是提取不同长度数据中包含的模式。生成器的输入为z片段和变换后的激励信号片段。令波形长度为N，第1个输入的分辨率是帧级（序列长度为$N/256$），它分别经4次上采样（上采样倍数分别是[8, 8, 2, 2]），分别得到长度是[$N/32$, $N/4$, $N/2$, N]的序列。第2个输入的分辨率是波形点级（序列长度为N），它分别经3次独立的下采样$\left(\left[\frac{1}{32},\frac{1}{4},\frac{1}{2}\right]\right)$和一次$1\times1$卷积，得到长度为[$N/32$, $N/4$, $N/2$, N]的序列。第1个输入的上采样和第2个输入的下采样每次的分辨率相同即可相加作为每层MRF网络的输入，最后多层MRF网络的输出相加，再预测波形。判别器有两个，分别是HiFi-GAN中的多周期判别器（Multi-period discriminator，MPD）和多尺度判别器（Multi-scale discriminator，MSD）。

1.3 基于Flow和变分数据增强的颤音建模

如图1所示，方差适配器由音高预测器、颤音预测器和能量预测器组成。首先，在进入方差适配器前，帧级表征$\boldsymbol{h}_{\text{text}}$会与位置嵌入相加，即将帧在音素中的位置、帧在音符中的位置通过绝对位置编码的形式加入。接着将这个加入了位置编码的$\boldsymbol{h}_{\text{text}}$与音符音高嵌入向量的和作为音高预测器的输入，音高预测器由多个FFT块堆叠（最后一个FFT块的输出记为$\boldsymbol{h}_{f_0}$），再经过一个线性变换层映射到通道数为2的时间序列，然后是一个sigmoid和恒等函数，分别用来预测清浊音（Unvoiced voiced，U/V）和对数基频。训练时音高损失为真实对数基频与预测对数基频的L_2范数，清浊音采用二元交叉熵作为损失函数。

颤音是指歌声呈波浪式的活动，其振动频率大致在4.5～6.5 Hz之间，是一种可以增加歌唱表现力的歌唱技巧。当气息通过声带发出声音，歌手有意识地将之震动，便会形成颤音。颤音的计算方法可以分为两种，第1种是基于正弦的颤音计算，即取一段浊音段对应的对数基频并假设其为正弦信号，衡量其振动幅度和振动频率；第2种方法是基于差分的颤音计算，其好处在于不假定颤音为正弦，可生成具有更复杂形状的颤音，由于音乐中音高的单位音分是基频的对数，因此基频作对数运算之前会对基频加入一个微小的正实数e以防对数运算错误。基于差分的颤音具体生成方法是对数基频的每个浊音段做低通（Low-pass，LP）滤波，获得平滑后的对数基频包络$\log_2\boldsymbol{f}_{0_{\text{LP}}}$，再将它们的差值作为浊音段的颤音$\boldsymbol{V}$（通常颤音的振动幅度有限，需要做限幅处理，即按照设定的最大颤音幅度对颤音波幅进行削平），对于清音段的颤音$\boldsymbol{V}$令其为$\boldsymbol{0}$，公式为

$$\log_2\boldsymbol{f}_{0_{\text{LP}}}=\begin{cases}\text{LP}(\log_2 f_0) & f_0>\text{e}\\ \log_2\text{e} & f_0=\text{e}\end{cases}\tag{5}$$

$$\boldsymbol{V}=\log_2\boldsymbol{f}_0-\log_2\boldsymbol{f}_{0_{\text{LP}}}\tag{6}$$

由于颤音具有准周期性且振动幅度稳定，因此本文将颤音序列视为一种分布，文本采用基于差分的颤音计算，并基于Flow的生成式模型依据最大似然准则进行训练。Flow模型是通过一系列可逆变换函数$\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_L$将输入的$\boldsymbol{x}$映射到随机变量$\epsilon$，它将输入的分布$\lg p(\boldsymbol{x};\boldsymbol{\theta})$根据变量变换

(Change-of-variables)定理与输出的分布 $\lg p_\epsilon(\epsilon)$ 联系,即

$$\lg p(\boldsymbol{x};\boldsymbol{\theta})=\lg p_\epsilon(\epsilon)+\lg\left|\frac{\partial\epsilon}{\partial\boldsymbol{x}}\right| \tag{7}$$

最后一项是 $\boldsymbol{f}$ 函数的雅可比行列式。建立输入输出的联系后,可根据分布 p_ϵ 进行采样得到 ϵ,再通过可逆变换函数的逆运算得到 $\boldsymbol{x}$。一般的生成式模型都具有中间隐层的维度高于输入的特点,然而Flow模型不具备因此限制了其表达能力,成为维度瓶颈问题。变分数据增强[15]通过引入随机变量 $\boldsymbol{v}$ 提升Flow输入变量的维度,进而提升中间隐层的维度,提高流模型的表达能力。具体而言,变分数据增强优化的目标为

$$\max_{\theta,\phi} E_{\hat{p}(\boldsymbol{x})q(\boldsymbol{v}|\boldsymbol{x};\boldsymbol{\phi})}[\lg p(\boldsymbol{x},\boldsymbol{v};\boldsymbol{\theta})-\lg q(\boldsymbol{v}|\boldsymbol{x};\boldsymbol{\phi})] \tag{8}$$

第1项将 $\boldsymbol{x}$、$\boldsymbol{v}$ 当作Flow的输入并且是普通Flow模型训练对应的目标;第2项是增强数据分布 $\lg q(\boldsymbol{v}|\boldsymbol{x};\boldsymbol{\phi})$,它使用另一个条件流进行建模,该条件流定义了可逆变换

$$\boldsymbol{v}=g^{-1}\left(\epsilon_q;\boldsymbol{x},\boldsymbol{\phi}\right),\lg q(\boldsymbol{v}|\boldsymbol{x};\boldsymbol{\phi})=\lg p_\epsilon\left(\epsilon_q\right)-\lg\left|\frac{\partial\boldsymbol{v}}{\partial\epsilon_q}\right| \tag{9}$$

推理阶段,Flow将随机噪声 ϵ 当作输入,通过一系列逆变换后得到输出 $\boldsymbol{x}$ 与 $\boldsymbol{v}$,因此需要丢弃 $\boldsymbol{v}$ 保留 $\boldsymbol{x}$。颤音预测器的框图见图2(b),该模型结构与VITS[12]的随机时长预测器类似,其主网络是基于带孔深度可分离卷积(Dilated and depth-wise separable convolutional, DDSConv)残差块的堆叠,并在耦合层中使用神经样条流(Neural spline flows)[16]与变分数据增强来提高流模型的表达能力。由于颤音包络是一个一维连续序列,直接使用流模型会引起维度瓶颈问题,因此引入了额外的随机变量 $\boldsymbol{v}$ 与颤音序列进行拼接得到高维的隐层表征以解决维度瓶颈的问题。随机变量 $\boldsymbol{v}$ 通过近似的后验分布进行采样获得,训练时颤音损失 L_{vib} 为负的变分下界。合成阶段,将基频预测器中的隐层变量 $\boldsymbol{h}_{f_0}$ 作为条件,高斯噪声 ϵ 通过Flow模型的逆变换得到颤音 $\boldsymbol{V}$ 与随机变量 $\boldsymbol{v}$,取第一项。训练阶段采取强制教学方式即真实基频会量化到256个基频值,推理阶段预测的基频和预测的颤音会相加作为新的预测的基频再量化到256个基频值,量化的基频索引再通过嵌入层转换为音高表征,并与帧级表征 $\boldsymbol{h}_{\text{text}}$ 相加作为能量预测器的输入,另一方面基频及其谐波也会作为解码器中激励模块的输入。

能量预测器由多个一维卷积组成,最后接一个线性层用以预测帧级能量。其中能量由音频的功率谱逐帧计算得到,为了模拟人耳对能量的感知,对功率谱不同频率系数的值进行加权。训练时能量损失 L_{Ene} 为真实对数能量与预测对数能量的 L_2 范数。最后,能量也会被量化到256个能量值,然后量化的能量索引通过嵌入层转换为能量表征,然后再与能量预测器的输入相加得到 $\boldsymbol{h}'_{\text{text}}$。

1.4　增加ReZero结构的先验网络

帧级先验网络将方差适配器的输出 $\boldsymbol{h}'_{\text{text}}$ 作为输入,并通过残差块堆叠一维卷积得到更细粒度的帧级均值 $\boldsymbol{\mu}_\theta$ 和方差 $\boldsymbol{\sigma}_\theta$。在低资源数据(如歌唱合成数据集)的训练过程中,发现若帧级先验网络全部使用基于一维卷积的残差网络,KL损失较大甚至可能会发散到无穷大,这可能是因为先验分布和后验分布之间存在巨大差距。因此,一个简单ReZero策略[11]被引入到残差块中,见图2(c)。在训练的初始阶段,对于输入 $\boldsymbol{x}$,拥有ReZero($\alpha=0$)的残差块等同于恒等变换即输出 $\boldsymbol{y}=\boldsymbol{x}$,随着训练的进行会逐渐增加残差连接 $F(\boldsymbol{x})$ 的比重,比重由可学习的参数 α 控制,即

$$\boldsymbol{y}=\boldsymbol{x}+\alpha F(\boldsymbol{x}) \tag{10}$$

式中 α 在训练开始时全为0,并在随后的训练中动态地将其调整为合适的值。因此,在初始阶段,帧级先验网络等同于一个几乎完全透明的恒等变换层,在反向传播时几乎所有参数都不存在梯度(除了帧级先验网络中用于变换维度的线性变换)。使用ReZero有两方面好处,一方面,可以使用更深的残差网络

来得到更细粒度的帧级统计向量;另一方面,也有利于信号在深度神经网络中的传播,使得网络的训练更加稳定。

2 实 验

歌曲选自中国流行歌曲,录音采样率为48 kHz,波形点采用16比特量化,实验中将它们下采样至24 kHz。歌曲首先被以音频的形式导入到旋律转录软件Tony中,由音乐标注人员进行音符边界的微调以及音高的校正,导出为MIDI后再使用Synthesizer V软件给MIDI加入歌词。这些MIDI将被转为TextGrid格式(其中歌词信息经过前端转为了音素信息),由于音素边界更为精细,语音标注人员在Praat软件中微调音素的边界,音符的边界以音素边界为准,所以可以看出音素和音调的时长是由专业标注人员根据实际演唱手动标注的。为了便于模型训练,音频被分成平均11.9 s的片段,数据集一共6.01 h的音频数据,具有1 820个片段,平均分成1 676个片段的训练集、1个片段的验证集和143个片段的测试集。训练集用于训练SVS模型,验证集用于调整超参数和可视化训练时的中间结果。实验中,帧长为42.67 ms,帧位移为10.67 ms。乐谱中的歌词信息不是以汉字的形式进行输入,而是用音素类别(在实验中为了简化起见,将汉语的声母和韵母视为音素)。音素类别、音符音高、音符时长、位置和延长音的嵌入向量维度都为192。隐变量z的维度是256维。文本编码器包含6个FFT,使用2头的自注意力机制。时长预测器由3个一维卷积网络组成,并且在时长预测器中采用了音符归一化[6-7]。音高预测器也包含6个FFT,使用2头的自注意力机制。颤音预测器中的流模型采用6层DDSConv残差块的堆叠,后验流模型采用4层DDSConv残差块的堆叠,基频平滑所用的低通滤波器截止频率是3 Hz(假定颤音频率至少3 Hz)。能量预测器包含12层的一维卷积层,每两层使用LayerNorm做归一化。帧级先验网络由6个有ReZero的一维卷积残差网络组成。激励信号中谐波数设置为$H=8$,因此最后的线性变换是一个9个节点输入,1个节点输出的全连接层,得到变换后的1维激励信号。其余的超参数设置与VISinger中的参数一致。使用AdamW优化器[17]训练网络,$\beta_1=0.8$,$\beta_2=0.99$,权重衰减$\lambda=0.01$。在每轮训练中,学习率衰减按$0.999^{1/8}$为倍率进行衰减,初始学习率为2×10^{-4}。所有模型都在V100上训练到了16万步骤,批处理大小设置为24。

2.1 实验配置

实验建立了以下5个系统用于比较。

(1)VISinging:该模型是按照第1节的介绍构建的。

(2)VISinging-Vib:在VISinging的方差适配器中去掉了颤音预测模块,损失函数也去掉了L_{vib}。

(3)VISinging-Vib-Ext:在VISinging-Vib的解码器中去掉了激励模块。

(4)VISinger:基于一体化建模的基线系统,将VISinging-Vib-Ext的帧级先验网络从带ReZero的残差一维卷积网络退化成了普通一维卷积网络,即带有能量预测器的VISinger[7]①。注意这里是没有加入音素预测模块和对应的音素识别损失的②。

(5)DiffSinger+HiFiGAN:基于两阶段建模的基线系统,声学模型是基于去噪扩散概率模型的DiffSinger[6],声码器是多人歌唱数据(大约70 h)上训练的通用声码器HiFiGAN[14]。

2.2 实验结果

2.2.1 整体评价

首先,在测试集上进行MOS的评估,以测量若干系统与歌唱录音的音质和自然度,其中自然度侧

①https://arxiv.org/abs/2110.08813v1.

②https://arxiv.org/abs/2110.08813v2.

重于在韵律方面进行评估。歌唱合成系统包括VISinging、VISinger、DiffSinger+HifiGAN。录音和3个系统各合成20个测试片段并由14位汉语母语者进行测听评价,要求测听人员使用从1(非常不自然)到5(非常自然)的分数以0.5的间隔对每个合成歌唱进行评分。表1显示了不同系统的MOS分结果。可以看到VISinging在自然度上的表现优于VISinger与DiffSinger+HifiGAN,VISinging在音质上的表现优于VISinger,与DiffSinger+HifiGAN基本持平。这表明本文提出的多个方法联合发挥了作用,即融合激励模块的解码器、基于归一化流的颤音预测模块,以及加入ReZero策略的帧级先验网络。

表1　多个系统的MOS分比较

Table 1　MOS comparison for different systems

模型	音质	自然度
VISinging	3.91	3.95
VISinger	3.61	3.53
DiffSinger+HifiGAN	3.92	3.88
录音	4.11	4.03

2.2.2　消融实验

接着评估了4种一体化模型VISinging、VISinging-Vib、VISinging-Vib-Ext和VISinger在测试集上预测的声学特征的客观指标[18]。指标包括梅尔倒谱失真(Mel-cepstral distortion, MCD)、基频的均方根误差(F0 RMSE)、基频的相关系数(F0 CORR)和U/V误差百分比(UV)。

(1)MCD:先计算合成语音中提取的每一帧梅尔倒谱相比于自然语音中提取的每一帧梅尔倒谱的失真,再计算所有帧梅尔倒谱失真的平均值。对于每一帧,梅尔倒谱失真定义为

$$\mathrm{MCD}=\frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}\left(\boldsymbol{c}_{\mathrm{r}}(m)-\boldsymbol{c}_{\mathrm{s}}(m)\right)^{2}} \tag{11}$$

式中:$\boldsymbol{c}_{\mathrm{r}}$和$\boldsymbol{c}_{\mathrm{s}}$分别表示自然语音和合成语音中提取的梅尔倒谱,$M$表示梅尔倒谱的阶数。梅尔倒谱失真反映了梅尔倒谱层面的失真程度,值越小失真越小,反映在歌唱上是音质越接近于原始音频。

(2)F0 RMSE:先计算合成语音中提取的每一帧基频与自然语音中提取的每一帧基频之间的均方根误差,再计算所有帧基频均方根误差的平均值。对于每一帧,基频以音分定义的均方根误差定义为

$$\mathrm{RMSE}=1\,200\sqrt{\left(\log_2 F_{\mathrm{r}}-\log_2 F_{\mathrm{s}}\right)^{2}} \tag{12}$$

式中:F_{r}和F_{s}分别表示自然语音和合成语音中提取的基频(单位为Hz),音分(cent)作为RMSE单位是因为音分与半音的关系以及人耳对音乐的感知。RMSE反映了基频层面的失真程度,值越低失真越小,反映在歌唱上是唱的准不跑调。

(3)F0 CORR:定义为合成语音与自然语音的浊音段的基频相关系数,值越高失真越小,反映在歌唱上是基频包络的起伏符合原始音频基频包络的起伏。

(4)UV:定义为合成语音与自然语音清浊不匹配的帧数与总帧数的比值,值越低失真越小,反映在歌唱是有声段、非有声段(例如呼吸音和静音)的基频合理性。

为了计算这4个指标,通过STRAIGHT[19]分析,以5 ms为帧移位,对合成波形提取了12维梅尔倒谱系数(Mel-cepstral coefficients, MCC)和F0。歌唱合成中预测的声学特征可直接与参考特征对齐,以计算MCD、RMSE、CORR和UV指标,结果如表2所示。从表2可以看到,VISinging、VISinging-Vib、VISinging-Vib-Ext具有相似的客观指标。VISinging系列模型与VISinger相比,具有可比的客观指标(更低的MCD和更高的F0 RMSE)。

表2　4种一体化模型在预测测试集声学特征上的客观指标对比

Table 2　Comparison of objective indicators of four integrated models in predicting acoustic features of test sets

模型	MCD/dB	RMSE(Cent)	CORR	UV/%
VISinging	4.19	64.12	0.988 8	10.21
VISinging-Vib	4.19	65.17	0.990 2	9.84
VISinging-Vib-Ext	4.20	64.62	0.989 6	9.81
VISinger	4.26	58.87	0.990 0	9.69

进一步将具有颤音的测试片段进行频谱可视化，图3显示了一个示例，可以明显地看到VISinging的谐波清晰并且具有颤音，这源于显式地引入了颤音的预测，并且值得注意的是颤音形状并不是完全的正弦形状，这符合基于差分的颤音计算而不是基于正弦的颤音计算的假设。相比之下，虽然VISinging-Vib的谐波分量依然如录音般清晰，但是基频形状平坦。再去掉激励模块之后，VISinging-Vib-Ext的高次谐波分量已不再清晰存在毛刺，在听觉上具有可察觉的不连续。

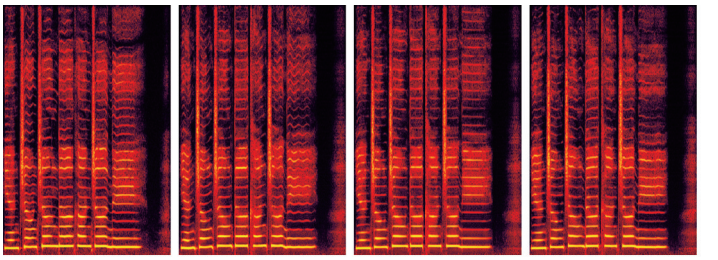

图3 单个测试片段录音与3种模型合成音频的频谱图

Fig.3 Spectrum diagrams of recording and synthesizing audios by three models for single test segment

为了验证ReZero在训练中发挥的作用，可以比较VISinging-Vib-Ext与VISinger在训练阶段KL损失的曲线，结果如图4所示。从图4可以看出，带有ReZero的VISinging-Vib-Ext的KL损失收敛得更快更低，说明使用ReZero的一维卷积残差网络比直接进行一维卷积网络的效果更好。另外由于训练初始阶段KL相对较大无法画在图中，通过观察数值可以发现，在训练初始阶段ReZero对于缓解较大的KL损失具有很大作用。非正式测听显示，VISinger的合成音频存在发音音素错误的问题，而加入ReZero后发音错误问题有所缓解。

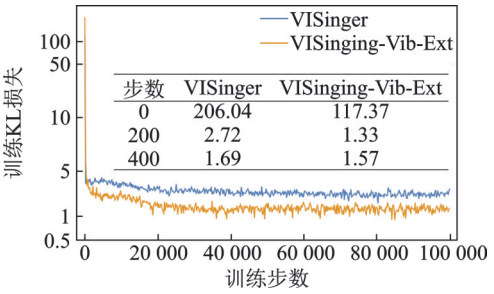

图4 VISinging-Vib-Ext与VISinger训练时的KL损失曲线对比

Fig.4 Comparison of KL loss curves of VISinging-Vib-Ext and VISinger at the training stage

3 结束语

本文提出了端到端歌唱合成系统VISinging，以解决VISinger音高听感不连续、颤音合成不佳及发音不稳定的问题。为了提升歌声合成效果，提出了3个创新点，包括融合激励模块的解码器、增加颤音预测模块以及在帧级先验网络中使用具有ReZero策略。实验结果显示，融合激励模块的解码器提升了合成音频中基频和谐波的稳定性，颤音建模对颤音的恢复有显著提升作用，ReZero策略对KL损失的收敛、训练速度和发音稳定性有一定提升。在主观测听上，VISinging显著优于一体化建模方法VISinger和两阶段建模方法DiffSinger+HiFiGAN。以后将继续在端到端框架上进行歌唱合成的研究，包括多唱歌技巧、多声部、可兼容说唱与独白的歌唱合成等。

参考文献：

[1] OURA K, MASE A, YAMADA T, et al. Recent development of the HMM-based singing voice synthesis system—Sinsy [C]//Proceedings of the 7th ISCA Workshop on Speech Synthesis(SSW 7). Kyoto, Japan: ISCA, 2010: 211-216.

[2] NISHIMURA M, HASHIMOTO K, OURA K, et al. Singing voice synthesis based on deep neural networks[C]// Proceedings of Terspeech. San Francisco, USA: ISCA, 2016: 2478-2482.

[3] CHEN J, TAN X, LUAN J, et al. HiFiSinger: Towards high-fidelity neural singing voice synthesis[EB/OL]. (2020-09-03). https://doi.org/10.48550/arXiv.2009.01776.

[4] GU Y, YIN X, RAO Y, et al. ByteSing: A Chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and WaveRNN vocoders[C]//Proceedings of the 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). HongKong, China: IEEE, 2021: 1-5.

[5] HONO Y, HASHIMOTO K, OURA K, et al. Sinsy: A deep neural network-based singing voice synthesis system[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2803-2815.

[6] LIU Jinglin, LI Chengxi, REN Yi, et al. DiffSinger: Singing voice synthesis via shallow diffusion mechanism[EB/OL]. (2021-05-06). https://doi.org/10.48550/arXiv.2105.02446.

[7] ZHANG Yongmao, CONG Jian, XUE Heyang, et al. VISinger: Variational inference with adversarial learning for end-to-end singing voice synthesis[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7237-7241.

[8] WANG X, TAKAKI S, YAMAGISHI J. Neural source-filter-based waveform model for statistical parametric speech synthesis[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l]: IEEE, 2019: 5916-5920.

[9] LIU R, WEN X, LU C, et al. Vibrato learning in multi-singer singing voice synthesis[C]//Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). [S.l]: IEEE, 2021: 773-779.

[10] SONG Yingjie, SONG Wei, ZHANG Wei, et al. Singing voice synthesis with vibrato modeling and latent energy representation[C]//Proceedings of 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). [S.l]: IEEE, 2022: 1-6.

[11] BACHLECHNER T, MAJUMDER B P, MAO H H, et al. ReZero is all you need: Fast convergence at large depth[EB/OL]. (2020-03-10). https://doi.org/10.48550/arXiv.2003.04887.

[12] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]//Proceedings of the 38th International Conference on Machine Learning. Hawaii, USA: PMLR, 2021.

[13] REN Yi, HU Chenxu, TAN Xu, et al. FastSpeech 2: Fast and high-quality end-to-end text to speech[C]//Proceedings of the 9th International Conference on Learning Representations, ICLR 2021. Virtual Event, Austria: [s.n.], 2021.

[14] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[EB/OL]. (2020-10-23). https://doi.org/10.48550/arXiv.2010.05646.

[15] CHEN Jianfei, LU Cheng, CHENLI Biqi, et al. VFlow: More expressive generative flows with variational data augmentation [EB/OL]. (2020-02-22). https://doi.org/10.48550/arXiv.2002.09741.

[16] DURKAN C, BEKASOV A, MURRAY I, et al. Neural spline flows[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems.Vancouver, Canada: [s.n.], 2019: 7509-7520.

[17] LOSHCHILOV Iiya, HUTTER Frank. Decoupled weight decay regularization[C]//Proceedings of International Conference on Learning Representations. New Orleans, USA: [s.n.], 2019.

[18] HAYASHI T, TAMAMORI A, KOBAYASHI K, et al. An investigation of multi-speaker training for wavenet vocoder[C]//Proceedings of 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).[S.l]: IEEE, 2017: 712-718.

[19] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNE A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3/4): 187-207.

作者简介:

周骁(1993-),男,博士,研究方向:语音合成、歌唱合成等,E-mail: xiaozhou3@iflytek.com。

胡亚军(1990-),男,博士,研究方向:语音合成、声音转换、歌唱合成等。

潘嘉(1985-),男,博士,高级工程师,研究方向:语音识别、说话人识别、语音合成、深度学习等。

胡国平(1977-),男,正高级工程师,研究方向:智能语音、认知智能技术的研发与落地工作。

凌震华(1979-),**通信作者**,男,教授,博士生导师,研究方向:语音信号处理与自然语言处理,E-mail: zhling@ustc.edu.cn。

(编辑:王静)