文章编号:1674-2974(2025)04-0114-10 DOI:10.16339/j.cnki.hdxbzkb.2025271

基于音素级韵律建模的自回归零样本语音合成

岳焕景 ,王嘉玮 ,杨敬钰 †
(天津大学 电气自动化与信息工程学院, 天津 300072)

摘 要:为了提升合成韵律的自然度和稳定性,提出了基于音素级韵律建模的自回归语音合成模型.该模型从词级别停顿和音素时长两方面改进韵律建模.为了提升词级别停顿的多样性和准确性,在文本前端提出了停顿预测模块.该模块基于原始文本来预测多类停顿标签,从而为语音合成提供停顿时长建模的准确参考.为了提升音素时长的自然度,提出了时长预测模块.该模块预测每个音素的混合高斯分布,并通过随机采样来获得多样化的音素时长.为了提升自回归模型中的音素时长建模的稳定性,提出了注意力判别模块.该模块应用于自回归的每个时间步中,并通过注意力和判断机制来避免对齐紊乱现象.实验结果表明,所提三种模块可有效提升韵律建模的自然度和稳定性,从而提升语音合成的效果.

关键词:语音合成;韵律建模;停顿预测
中图分类号:TP37 **文献标志码**:A

Autoregressive Zero-shot Speech Synthesis Based on Phoneme-level Prosody Modeling

YUE Huanjing, WANG Jiawei, YANG Jingyu†
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: To improve the naturalness and robustness of synthesized prosody, a autoregressive speech synthesis model based on phoneme-level prosody modeling is proposed. This model enhances prosody modeling from two aspects: inter-word pauses and phoneme durations. To enhance the diversity and accuracy of inter-word pauses, a pause prediction module is proposed at the text frontend. This module predicts multiple pause labels based on the original text, thereby providing accurate references for pause duration modeling in speech synthesis. To enhance the naturalness of phoneme durations, a duration prediction module is proposed. This module predicts a mixture Gaussian distribution for each phoneme and obtains diversified phoneme durations through random sampling. To stabilize phoneme duration modeling in the autoregressive model, an attention-based discrimination module is proposed. This module is applied at each time step of the autoregressive process and avoids alignment disorder through attention and discrimination mechanisms. Experimental results demonstrate that the three proposed modules effectively enhance the naturalness and robustness of prosody modeling, thereby improving the quality of speech synthesis.

Key words: speech synthesis; prosody modeling; pause prediction

---

* **收稿日期**:2024-02-04
基金项目:国家自然科学基金资助项目(61672378), National Natural Science Foundation of China(61672378)
作者简介:岳焕景(1987—),女,山东济宁人,天津大学副教授,博士
† 通信联系人,E-mail: yjy@tju.edu.cn

语音合成通过模拟人类的语音生成过程,将文本信息转化为语音信息.传统的多人语音合成需要大量的语音进行微调训练,限制了其应用范围;而零样本语音合成旨在给定少数参考语音的情况下,合成未见过的目标语音[1],无需额外样本微调训练.本文工作专注于零样本合成场景.

主流的基于深度学习的语音合成方法可以分为自回归方法和非自回归方法.两种方法各有优缺点:自回归方法合成质量优秀,但合成速度慢,且存在对齐紊乱(错误累计)现象[2];非自回归方法合成速度快,但合成质量比较平庸[3].本文专注于通过韵律建模提升合成质量,因此采取自回归方法.

近年来,许多研究在韵律建模方面有了显著成果.Shen 等人[4]提出了基于位置敏感注意力机制的对齐方式,将自回归模型先前时间步的累计注意力作为当前时间步的条件特征,减少了跳词和复读等对齐紊乱现象,提升了自回归模型合成韵律的稳定性.Kim 等人[5]提出了随机音素时长预测器,使用流生成模型来模拟时长分布,提升了音素时长的多样性.Yang 等人[6]基于双向编码器表征模型[7](bidirectional encoder representation from transformers, BERT)提出了停顿预测模型,通过预测断句停顿标签,提升了合成韵律的自然度.

在零样本语音合成中,由于参考语音较短,所提供的信息量较少,因此韵律建模更加困难.为了提升零样本场景下韵律建模的自然度和稳定性,本文提出了基于音素级韵律建模的自回归零样本语音合成.对于停顿,本文提出使用词级别的停顿预测模块来实现自然的停顿插入.对于音素时长,本文提出使用基于混合高斯分布的时长预测模块来提升音素时长的自然度.对于自回归模型的韵律稳定性,本文提出使用基于注意力[8]的判别模块来避免对齐紊乱现象.本文提供了合成音频的展示网页以供直观评估:phoswjw.github.io/prosody_zs.

1　模型结构和原理

1.1　模型整体结构

零样本语音合成的目标是测试时只需输入一小段从未见过的说话人的参考语音 $\boldsymbol{r}$(通常为3~15 s),模型即可由输入文本 $\boldsymbol{x}$ 合成该说话人的语音 $\hat{\boldsymbol{y}}$,无需额外的样本进行微调训练,该过程可表示为:

$$\hat{\boldsymbol{y}} = M(\boldsymbol{x}, \boldsymbol{r}) \tag{1}$$

其中,$\boldsymbol{r}$ 的语义内容与输入文本不同.

本文所提模型的整体结构如图 1 所示,该模型的主体结构可分为文本编码器、自回归 Transformer 解码器、声学解码器、韵律建模四部分.其中韵律建模部分主要包括停顿预测模块、时长预测模块,以及注意力判别模块.

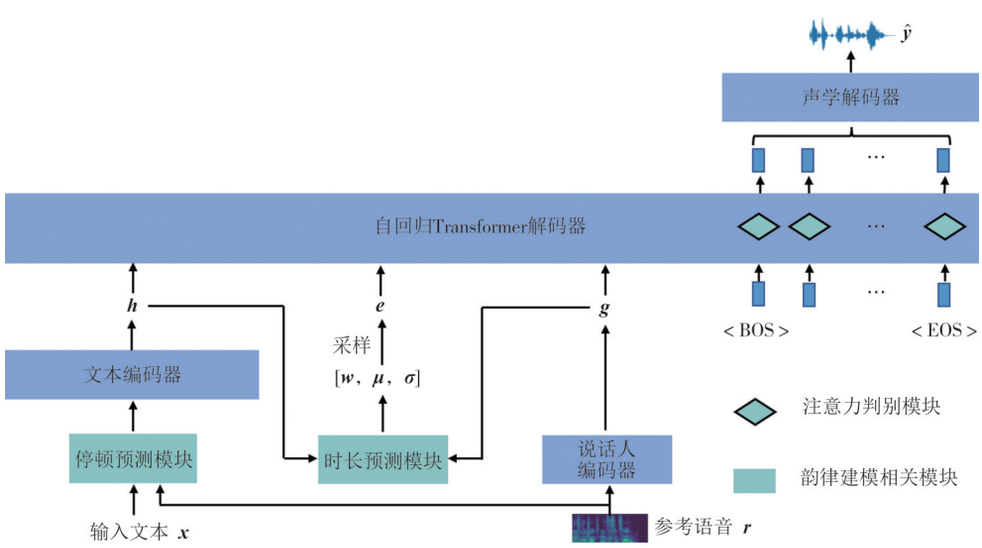


图 1　模型整体结构

Fig.1　Architecture of the proposed model

对于 $\boldsymbol{x}$,首先使用停顿预测模块来预测词间停顿,并将输出的停顿标签插入 $\boldsymbol{x}$,然后使用预训练的词转音素模型将插入停顿后的文本转换为音素序列,随后将其输入基于 Transformer 的文本编码器[9],得到文本向量 $\boldsymbol{h}$.对于 $\boldsymbol{r}$,将其作为预训练的说话人编码器[10]的输入,得到声纹向量 $\boldsymbol{g}$.然后,将 $\boldsymbol{h}$ 和 $\boldsymbol{g}$ 作为

时长预测模块的输入,输出时长分布[$\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\sigma}$],随后采样[$\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\sigma}$],得到时长向量$\boldsymbol{e}$.最后,将$\boldsymbol{h}$、$\boldsymbol{g}$以及$\boldsymbol{e}$输入基于Transformer的自回归解码器[11-12]中,该解码器自回归地生成声音表征序列,并将其作为声学解码器[12]的输入,最终生成目标语音$\hat{\boldsymbol{y}}$.需要注意的是,在图1自回归生成的过程中,<BOS>和<EOS>分别代表开始标记和结束标记,且注意力判别模块被应用在每一个时间步之中.下面将分别介绍韵律建模部分的三个模块.

1.2 停顿预测模块

在语音合成中,模型通常会按照输入文本序列中的标点符号来建模停顿,即符号停顿(punctuation-indicated pauses, PIPs).然而,这种方法一方面忽略了非符号停顿(respiratory pauses, RPs);另一方面缺乏对停顿时长的分类[13].基于以上观察,本文按照不同的时长划分停顿类别,并在词级别上分别预测PIPs和RPs的多类标签.本文将停顿标签按时长分为以下5类:0(不停顿),1(<200 ms),2(200~400 ms),3(400~600 ms),4(>600 ms).

本文提出的停顿预测模块如图2所示.该模块整体是一个分类语言模型,$\boldsymbol{x}$为由单词和标点符号组成的原始文本.由于停顿预测是基于词级别的,训练数据无法涵盖所有的单词,因此需要使用分词工具[14]对$\boldsymbol{x}$进行处理,将$\boldsymbol{x}$中的复杂词汇拆分并用子词替代,例如bookshelf被分为book和shelf,然后shelf将代替bookshelf.通过分词操作,可以大大降低对训练数据的词汇覆盖率要求.为了提高模型理解语言上下文信息的能力,本文使用预训练的BERT模型和双向门控循环单元[15](gated recurrent unit, GRU)对文本序列进行处理,二者之间会插入一个说话人调制模块[12],其目的是使用当前说话人的参考语音$\boldsymbol{r}$为模型提供目标说话人的个性化习惯信息.双向GRU的输出经过Softmax,得到最终的分类概率向量$\boldsymbol{p}$.

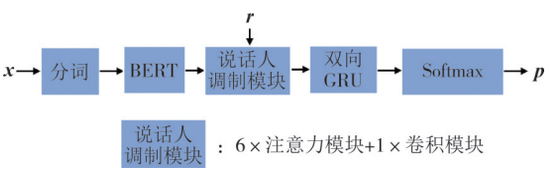


图2 停顿预测模块

Fig.2 Pause prediction module

在训练过程中,$\boldsymbol{p}$与真实停顿标签之间会计算损失函数.本文使用蒙特利尔强制对齐模型(Montreal forced aligner, MFA)[16]获取真实音频中所有的静音位置和时长,并按所述标准划分为不同的类别,从而得到真实停顿标签.考虑到在数据集中,不同类别停顿的数量不平衡,本文采用加权交叉熵损失函数(weighted cross-entropy loss, WCE),根据训练过程中每个类别的样本数量动态调整不同类别的权重,避免模型忽略少数样本的类别.WCE表示如下:

$$L_{WCE}(\boldsymbol{u},\hat{\boldsymbol{u}}) = -\sum_{0}^{C-1} w_k u_k \ln(\hat{u}_k) \tag{2}$$

式中:$\boldsymbol{u}$是真实标签向量;$\hat{\boldsymbol{u}}$是预测的概率向量;w_k是权重向量;C是向量维度.

1.3 时长预测模块

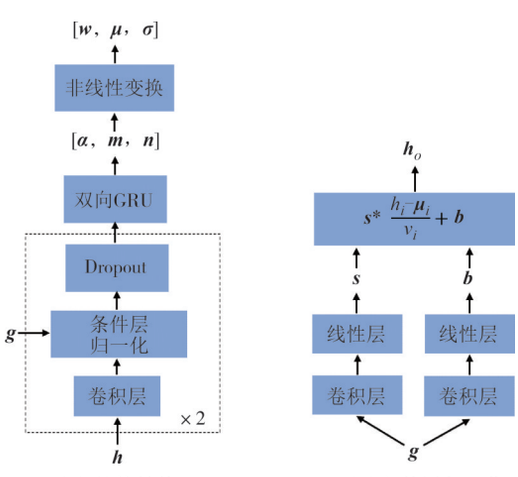
1.3.1 模型结构

时长预测通常是语音合成的子任务,其目标是根据输入的音素序列,预测每个音素的持续时长.传统的时长预测模块[17]通常会预测一组确定的音素时长,这大大限制了韵律模式的多样性.基于以上观察,本文通过预测音素时长的混合高斯分布来提升时长建模的多样性.

本文所提的时长预测模块如图3(a)所示,文本向量$\boldsymbol{h}$首先被输入两层特征提取网络中,每一层由卷积层(一维卷积+LeakyReLU激活)、条件层归一化以及Dropout组成.其中,条件层归一化用于引入声纹向量$\boldsymbol{g}$,利用说话人的信息引导模型预测出更符合声纹习惯的个性化时长分布.条件层归一化的结构如图3(b)所示,$\boldsymbol{g}$通过卷积层和线性层分别得到条件层归一化的权重$\boldsymbol{s}$和偏置$\boldsymbol{b}$,并通过$\boldsymbol{s}$和$\boldsymbol{b}$来调制当前的特征$\boldsymbol{h}_i$,得到输出特征$\boldsymbol{h}_o$:

$$\boldsymbol{h}_o = \boldsymbol{s} * \frac{\boldsymbol{h}_i - \boldsymbol{u}_i}{\boldsymbol{v}_i} + \boldsymbol{b} \tag{3}$$

式中:u_i和v_i分别代表当前特征的均值和标准差.



(a) 整体结构 (b)条件层归一化

图3 时长预测模块

Fig.3 Duration prediction module

为了更好地处理序列的上下文依赖关系,使用双向 GRU 对先前网络的输出进一步处理,得到向量组[$\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{n}$].由于混合高斯分布的各分量权重之和需为 1,且各分量方差需为正值,因此需要对[$\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{n}$]进行非线性变换,基于 Softmax 函数和指数函数的特性,非线性变换表示如下:

$$\mu_i = m_i \tag{4}$$

$$\sigma_i^{\,2} = \exp(n_i) \tag{5}$$

$$w_i = \frac{\exp(\alpha_i)}{\sum_{k=1}^{G}\exp(\alpha_k)} \tag{6}$$

式中:w_i、μ_i和$\sigma_i^{\,2}$分别为第i个高斯分量对应的权重、均值和方差;G为高斯分量的数目.

在推理阶段,如图 1 所示,模型会从预测的混合高斯分布[$\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}$]中采样得到时长向量$\boldsymbol{e}$,为自回归生成提供时长参考.

1.3.2　训练

在训练阶段,时长预测模块会与外部的对齐信息进行有监督训练,如图 4 所示.首先使用预训练的 MFA 模型来处理真实语音$\boldsymbol{y}$,得到音素时长序列d.d中的元素均为标量,为了满足监督训练[$\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}$]的要求,需要使用变分数据增广来对$d$进行升维[5].本文使用 4 层 Wavenet[18]作为变分数据增广的模型结构,该模型将d作为条件输入,从高斯噪声N中得到高维向量组$\boldsymbol{v}$(d与$\boldsymbol{v}$具有相同的时间分辨率),然后将d与$\boldsymbol{v}$在特征维度上拼接,得到时长向量$\boldsymbol{e}$.

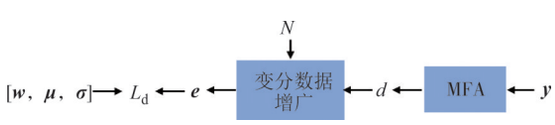


图 4　时长预测训练示意图

Fig.4　Training schematic of duration prediction

本文使用负对数似然损失函数来约束时长预测模块的训练,表示如下:

$$L_{\mathrm{d}} = \sum_{i=1}^{K} -\ln\left(\sum_{k=1}^{G} w_{k,i}\mathcal{N}\left(e_i;\ \mu_{k,i}, \sigma_{k,i}^2\right)\right) \tag{7}$$

式中:$w_{k,i}$、$\mu_{k,i}$和$\sigma_{k,i}^2$代表第i个音素的第k个高斯分量所对应的权重、均值和方差;e_i表示第i个随机变量.

1.4　注意力判别模块

自回归语音合成模型依靠每个时间步的输入表征与文本(音素)向量之间的关联来进行韵律对齐,这种方法能使自回归模型更好地关注序列的上下文内容依赖关系.然而,实际上训练很难让模型达到完美的泛化能力,这意味着在推理过程中可能出现对齐紊乱问题,即某个时间步对齐模块未能获取正确的加权音素向量,导致给到合成模块的内容信息错误,从而出现跳词、复读、模糊发音的现象.基于以上观察,本文通过使用注意力判别模块来解决对齐紊乱问题.

本文所提的注意力判别模块如图 5 所示,图 5 展示了自回归 Transformer 解码器中单个时间步的结构,包括注意力判别模块(虚线框内部)和 Transformer 解码模块.注意力判别模块有三个输入:第k个时间步的输入表征$\boldsymbol{i}_k \in \mathbb{R}^{C_1}$,文本向量$\boldsymbol{h} \in \mathbb{R}^{T\times C_2}$,以及时长向量$\boldsymbol{e} \in \mathbb{R}^{T\times C_3}$.

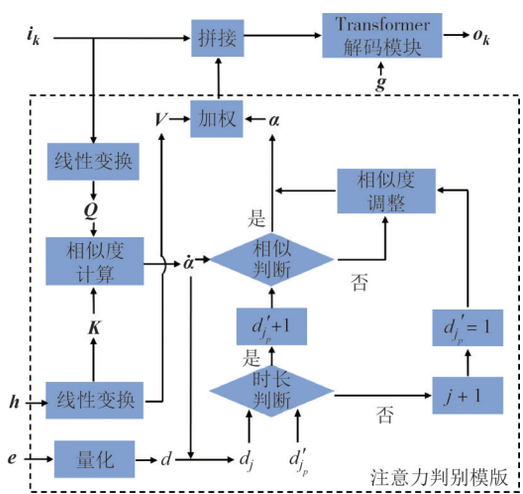


图 5　自回归 Transformer 解码器

Fig.5　Autoregressive Transformer decoder

首先,使用线性变换将$\boldsymbol{i}_k$和$\boldsymbol{h}$转换到三个不同的特征域:

$$\boldsymbol{Q} = \boldsymbol{i}_k W_q \tag{8}$$

$$\boldsymbol{K} = \boldsymbol{h} W_k \tag{9}$$

$$\boldsymbol{V} = \boldsymbol{h} W_v \tag{10}$$

然后计算相似度,得到$\boldsymbol{i}_k$对$\boldsymbol{h}$的注意力权重$\dot{\boldsymbol{\alpha}}$:

$$\dot{\boldsymbol{\alpha}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C'}}\right) \tag{11}$$

其中,C'为特征维度.

再次,注意力判别模块会进行时长判断.由混合高斯时长分布采样得来的时长向量$\boldsymbol{e}$经过量化[19]后得到音素时长序列d(每个元素均为标量),然后依据$\dot{\boldsymbol{\alpha}}$中最大相似度的音素索引$j$,从$d$中选择$d_j$,$d_j$即为当前时间步的参考时长.此外,模型会持续维护一个累计时长d'_{j_p},代表先前的最相似音素j_p的累计持续时长.时长判断的标准有两方面:一是判断当前时间步的最相似音素j与先前时间步的最相似音素j_p是否

一致;二是判断 d'_{j_p} 是否小于 d_j.只有当 $j = j_p$ 且 $d'_{j_p} < d_j$ 时,时长判断通过,因为这意味着当前时间步的内容信息正确且满足不超过参考时长的要求(若超过则有可能同一个音素发音会持续预测,出现复读现象),通过后执行 $d'_{j_p} + 1$ 并继续进行相似判断;若时长判断不通过,将会更新最相似音素($j_p + 1$)并将 d'_{j_p} 重置为1($d'_{j_p} = 1$),而后进行相似度的调整.

通过以上方式,时长判断控制每个时间步最相似音素的选择,即控制每个时间步最应该合成的内容信息,避免跳词和复读现象.

相似判断只有在时长判断通过的前提下才会执行.本文设置了最大相似度的最低阈值 β 来约束自回归生成,而相似判断标准是:判断 $\dot{\boldsymbol{\alpha}}$ 中的最大相似度 $\dot{\boldsymbol{\alpha}}_j$ 是否大于 β.若 $\dot{\boldsymbol{\alpha}}_j > \beta$,判断通过,$\dot{\boldsymbol{\alpha}}$ 会成为当前时间步的音素向量权重 $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \dot{\boldsymbol{\alpha}} \tag{12}$$

若相似判断不通过,将会进行相似度调整.相似度调整依赖于 β,表示如下:

$$\alpha_j = \beta \tag{13}$$

$$\alpha_k = \frac{\dot{\alpha}_k \dot{\alpha}_j}{\beta}, k \neq j \tag{14}$$

通过引入阈值 β 和相似度调整,相似判断将最相似音素的权重控制在较高数值,避免出现发音模糊的现象.此外还保留了其他音素的相对权重关系,保持上下文依赖关系.

在得到 $\boldsymbol{\alpha}$ 后,$\boldsymbol{\alpha}$ 与 $\boldsymbol{V}$ 加权,得到当前时间步的加权音素向量,该向量作为自回归生成的内容信息参考,与 $\boldsymbol{i}_k$ 拼接,作为 Transformer 解码模块的输入,最终生成当前时间步的输出表征 $\boldsymbol{o}_k$.

1.5 语音合成基本结构

本文的主要工作为韵律建模,语音合成部分采用了现有的方法.为了便于理解,本节简要介绍语音合成部分的基本结构.

1.5.1 文本编码器

本文采用了文献[9]所提出的文本编码器,其结构如图6所示,$\boldsymbol{x}_p$ 代表插入停顿标签后的文本序列,首先将其转换为音素序列,然后输入6层Transformer中,每一层由多头自注意力、归一化、卷积层以及跳跃连接构成,最终输出文本向量 $\boldsymbol{h}$.

1.5.2 Transformer 解码模块

自回归解码器单个时间步的结构由注意力判别模块和Transformer解码模块组成(图5所示),其中Transformer解码模块采用了文献[11]所提的结构,

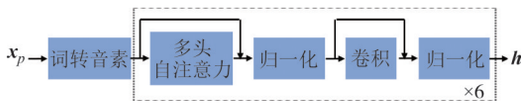


图6 文本编码器

Fig.6 Text encoder

如图7所示,主要由12层Transformer(虚线框内)组成.

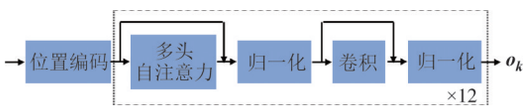


图7 Transformer 解码模块

Fig.7 Transformer decoder module

1.5.3 声学解码器

本文采用了文献[12]所提的声学解码器,用于将自回归解码器的输出 $\boldsymbol{y}_o$ 转换为波形 $\hat{\boldsymbol{y}}$,如图8所示,其结构由4层多感受野层(虚线框内)组成,每层由3个残差膨胀卷积层、反卷积层及ReLU激活函数组成.

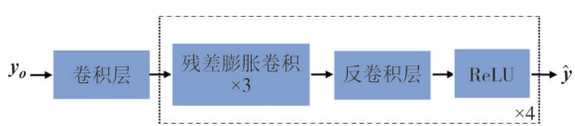


图8 声学解码器

Fig.8 Vocoder

1.6 损失函数

在训练阶段,除了式(2)和式(7)所表示的损失函数外,本文模型主体结构还使用了重建损失和生成对抗损失.

对于合成语音 $\hat{\boldsymbol{y}}$ 和真实语音 $\boldsymbol{y}$,使用相同参数设置将其分别转换为 $\hat{\boldsymbol{y}}_m$ 和 $\boldsymbol{y}_m$,重建损失计算二者之间的距离,表示为:

$$L_{\text{re}} = \left|\left| \hat{\boldsymbol{y}}_m - \boldsymbol{y}_m \right|\right|^2 \tag{15}$$

对于对抗训练的损失,本文参照Kong等人[24]的建议,采取最小二乘损失函数来避免梯度消失现象:

$$L_{\text{adv}}(G) = \mathbb{E}_z\left[\left(D\left(G(z)\right) - 1\right)^2\right] \tag{16}$$

$$L_{\text{adv}}(D) = \mathbb{E}_{y,z}\left[\left(D\left(G(z)\right)\right)^2 + \left(D(y) - 1\right)^2\right] \tag{17}$$

2 数据集、实验设置与评估指标

2.1 数据集

本文使用多个数据集,包括:VCTK[20],LibriTTS[21]以及MLS子集[22].VCTK数据集包含109个

英文母语说话人的录音和人为文本标注.语音数据包含多种自然的风格、口音和韵律音调,采样率为48 kHz.每位说话人贡献了约400句语料,时长总计超过44 h.本文实验选取9个不同口音的说话人用作测试集,其余用作训练集和验证集.LibriTTS数据集来源自Librispeech[23],包含2 456个英文母语说话人的录音和转录文本,语音的采样率为24 kHz,总时长超过585 h.在实验中,80%的数据用于训练集,10%用于验证集,10%用于测试集.MLS是一个跨语言的语音数据集,涵盖8种语言,包含超过50 000 h的音频和相应的转录文本,数据来源自Librivox开源项目,其原始采样率为48 kHz.其中英文部分包含超过44 600 h的数据,包含2 742个男性说话人和2 748个女性说话人.本文从MLS英文数据集中选取100位说话人作为实验数据集,语音时长为504 h.MLS数据集90%用于训练集,10%用于验证集.

VCTK,LibriTTS以及MLS的训练集共同用于网络训练.测试阶段,使用VCTK与LibriTTS的测试集共20位说话人进行混合测试,其中9位来自VCTK,11位来自LibriTTS.

2.2　实验设置

实验所使用的音频均被重采样至24 kHz.计算梅尔谱图需要用到短时傅里叶变换(short time Fourier transform, STFT),实验使用Hann窗口来计算STFT,FFT点数设置为1 024,窗口长度为1 024个采样点,帧移为256个采样点.对于梅尔频谱,采用80通道的梅尔滤波器组将STFT频谱映射到梅尔频谱.

在训练过程中,批次大小被设置为32,采用AdamW优化器,初始学习率被设置为2×10^{-4},指数衰减率被设置为0.999.实验基于PyTorch深度学习框架,采用DeepSpeed加速框架,在NVIDIA A100上训练60个周期.

2.3　评估指标

本文采用多种指标综合评估模型性能,包括平均主观意见分[1-6](mean opinion score, MOS)、相似度平均主观意见分[1](similarity mean opinion score, Sim-MOS)、MUSHRA分数[25](MUSHRA score, MUSHRA)、梅尔倒谱失真[26](Mel-cepstral distortion, MCD)、词错误率[27](word error rate, WER)、F_1分数[6].

MOS用于评估语音整体的质量,评测者聆听多组数据,根据主观意见打分,评分范围为1到5,评分越高,语音质量越好.最后,按照一定的置信水平计算均值,得到MOS分数.Sim-MOS是MOS的扩展,其评分标准与MOS基本一致,不同在于,Sim-MOS侧重于合成音频与真实音频整体的相似程度.MUSHRA与MOS都属于国际电信联盟(ITU)制定的音频主观评价指标,用于评价音频整体或者某个特定属性的表现.与MOS不同的是,MUSHRA可以对更细微的差异进行评级.MUSHRA打分范围为0~100,分数越高越好.

在本文实验中,MUSHRA被定义为专注于韵律自然度的评分.所有的主观评估均在20位评测者20组音频的条件下进行,并在95%置信水平下得到最终的主观分数.

MCD是一种客观评价指标,用于衡量两个梅尔倒谱系数(Mel frequency cepstral coefficients, MFCC)序列之间的距离.在本文实验中,MCD被应用于真实音频和合成音频之间,MCD越低,代表合成音频越接近人类发音的自然度.

WER是语音识别领域常用的客观评价指标,基于编辑距离算法计算错词率,用于衡量识别文本的正确程度,WER越低,识别文本越准确.在本文实验中,WER主要用于衡量韵律对齐的稳定性,通过人工转录合成音频来得到对应的文本,然后与真实文本计算WER.在计算WER的人工转录步骤中,若单词中出现部分读音明显缺读或无法听清读音的情况,均视为错词.

F_1分数是一种用于评估分类模型性能的指标,结合了精确率和召回率两个指标,通过一个调和平均数来计算,以综合考虑分类器的精度和召回能力,F_1越高代表分类性能越好.

3　实验结果与分析

3.1　对比实验

本文与其他三种零样本语音合成方法在五种指标上进行了对比实验.Attentron是一种利用基于循环神经网络的自回归模型,使用粗、细粒度两个编码器分别提取基本声纹信息和韵律等细致风格信息,并作为自回归生成的指导信息.YourTTS是一种基于条件变分自编码器的非自回归模型,使用内部动态规划算法与随机时长预测器联合建模韵律,并联合生成对抗训练实现端到端生成.Pause-based TTS是一种基于语言模型的自回归模型,使用预训练的语言模型来建模韵律停顿,并基于Transformer结构实现端到端生成.对比实验的结果如表1所示,本文在五个指标均取得了领先.

表1 对比实验

Tab.1 Comparison experiment

模型	MOS	Sim-MOS	MUSHRA	MCD	WER/%
真实语音	4.36±0.06	4.32±0.06	91.7±1.3	0	0
Attentron[28]	3.87±0.07	3.74±0.08	64.3±3.6	6.83	5.5
YourTTS[1]	4.08±0.06	4.04±0.07	74.9±2.9	4.60	1.1
Pause-based TTS[6]	4.20±0.06	4.11±0.06	81.7±2.5	3.45	2.3
本文完整模型	**4.23±0.06**	**4.18±0.06**	**85.8±2.2**	**3.19**	**0.7**

对于MOS,本文方法相较于其他三种方法分别提升了0.36、0.15、0.03,与Pause-based TTS较为相近,并领先其他两种方法较多.本文方法与Pause-based TTS均采用了基于Transformer语言模型的语音合成架构,而Attentron和YourTTS分别采用基于循环神经网络和基于条件变分自编码器的语音合成架构.此外,本文方法与Pause-based TTS均采用了BERT语言模型作为停顿预测的主要结构.实验结果表明语言模型的自回归生成与注意力机制对于提升序列数据的上下文理解能力有很大作用,能够提升合成音频序列的发音连贯性以及停顿预测的准确度,进而提升合成语音的整体自然度.

对于Sim-MOS,本文方法相较于其他三种方法分别提升了0.44、0.14、0.07;对于MCD,则分别降低了3.64、1.41、0.26;对于MUSHRA,本文方法相较于其他三种方法分别提升了21.5、10.9、4.1.Attentron的表现与本文方法差距最大,这是因为其对韵律的建模完全依赖于先前时间步序列的信息,一方面无法实现自然的停顿插入;另一方面未对自回归生成进行约束,导致出现错误累计现象的概率相对较高,严重影响了模型的平均表现.YourTTS与本文方法的差距也相对较大,这是因为非自回归生成虽然不会产生错误累计现象,但YourTTS的随机持续时长预测器是并行预测,没有上文信息的引导,导致其预测结果存在过平滑现象,无法模拟更加自然的音素时长(例如自然状态下存在的很长或很短的音素时长);同时,YourTTS没有单独的停顿建模,导致其无法模拟自然的发音停顿习惯.Pause-based TTS与本文方法均采用了基于BERT的停顿建模,能够更好地模拟真实人声的发音停顿习惯,因此Pause-based TTS的表现与本文方法差距最小.三种指标的实验结果共同表明本文的韵律建模能够增强对目标说话人韵律信息的捕捉能力,进而提升合成音频与人声的相似度.

对于WER,本文方法的实验结果为0.7%,低于非自回归YourTTS的1.1%,而其他两种自回归方法分别为5.5%和2.3%,这表明本文所提的注意力判别模块能够有效解决自回归生成中的对齐紊乱问题,进而提升韵律建模的稳定性.

3.2 消融实验

为了验证所提韵律建模方法的有效性,本文对基于音素级韵律建模的自回归零样本语音合成模型进行了消融实验,分别对停顿预测模块、时长预测模块、注意力判别模块进行了消融.对于时长预测模块,本文选择静态时长预测模块[29]来代替,对于其他两种模块则直接消去.消融实验的结果如表2所示.

表2 消融实验

Tab.2 Ablation study

模型配置	MOS	Sim-MOS	MUSHRA	MCD	WER/%
真实语音	4.34±0.06	4.28±0.06	91.1±1.3	0	0
停顿预测	4.02±0.07	3.99±0.07	72.1±3.8	4.38	0.8
时长预测	4.15±0.06	4.10±0.06	80.7±2.6	3.31	0.8
注意力判别	4.10±0.06	4.03±0.07	79.3±2.9	3.95	2.6
本文完整模型	**4.20±0.06**	**4.18±0.06**	**86.1±1.9**	**3.19**	**0.7**

消融停顿预测模块后,模型的MOS、Sim-MOS、MUSHRA指标分别下降了0.18、0.19、14.0,MCD上涨了1.19,模型性能下降较大幅度,这说明预测的多类停顿对于引导模型合成自然的韵律十分重要.此外,WER仅上升了0.1个百分点,说明停顿预测模块与韵律稳定性无关.

消融时长预测模块后,模型的MOS、Sim-MOS、MUSHRA指标分别下降了0.05、0.08、5.4,MCD上涨了0.12,模型性能下降,但幅度相对较小,说明预测的混合高斯时长分布能够提升韵律模式的多样性,从而提升模型性能,但总体上影响力不如停顿预测模块.WER同样仅上升了0.1个百分点,表明更换时长预测模块不会对韵律稳定性产生影响.

消融注意力判别模块后,模型的MOS、Sim-MOS、MUSHRA指标分别下降了0.10、0.15、6.8,MCD上涨了0.76,WER上升了1.9个百分点,表明注意力判别模块能够有效解决自回归生成中的对齐紊乱问题,提升韵律稳定性.

3.3 停顿预测对比实验

为了评估停顿预测模块的性能,本文与其他两种停顿预测模型在F_1分数上进行了对比实验.BLSTM-CRF基于长短时记忆网络(long short-term memory, LSTM)预测停顿;Pause-based TTS基于BERT-base预测停顿.

实验结果如表3所示,本文模型的F_1分数相较于BLSTM-CRF和Pause-based TTS分别提升了8.87个百分点和1.57个百分点.本文与Pause-based TTS均使用了BERT,而本文效果更好,说明本文的说话人调制方法能够有效为模型提供目标说话人的个性化信息,增强对多说话人不同条件的适应能力;而本文相较于BLSTM-CRF提升较大,说明BERT对于处理文本的上下文依赖关系有着重要作用.

表3　停顿预测对比实验

Tab.3　Comparison experiment of pause prediction

模型	F_1/%
BLSTM-CRF[30]	80.24
Pause-based TTS[6]	87.54
本文模型	**89.11**

3.4 时长预测对比实验

为了评估时长预测模块的性能,本文在维持语音合成总体框架不变的前提下,对使用不同的时长预测的模型所合成的音频进行了MUSHRA评估.Fastspeech2的时长预测器基于卷积结构,实现静态时长预测;VITS的时长预测器基于变分推理,实现随机时长预测.

实验结果如表4所示,相较于使用Fastspeech2和VITS中的时长预测模型,使用本文模型的MUSHRA分数分别上涨5.7、3.4,表明本文时长预测模块能够通过建模混合高斯时长分布,有效提升韵律模式的多样性,进而提升合成语音的自然度.

表4　时长预测对比实验

Tab.4　Comparison experiment of duration prediction

模型	MUSHRA
Fastspeech2[29]	80.5±2.7
VITS[5]	82.8±2.2
本文模型	**86.2±1.9**

3.5 阈值实验

在注意力判别模块中,本文设置了最大相似度阈值β来约束注意力机制.β的数值决定了当前时间步的最大相似音素的最低权重占比,以及其他音素的权重调整比例.

为了评估β的数值设置对韵律稳定性的影响,本文在WER指标上对0.4、0.6、0.8三种不同的β进行了对比实验.实验结果如表5所示,将β从0.8降至0.6,WER出现了0.4个百分点的涨幅;当β从0.6降至0.4时,出现了1.2个百分点的涨幅.这说明β设置为较高数值有利于模型更好地学习到拥有最大相似度的内容信息,从而减少对齐紊乱现象.

表5　不同阈值对比实验

Tab.5　Comparison experiment of different thresholds

β	WER/%
0.4	2.3
0.6	1.1
0.8	**0.7**

3.6 与相关工作的对比分析

目前已有多种基于注意力机制的自回归语音合成工作,其中与本文工作最相关的是Valle[31]与Pause-based TTS[6].然而,Valle未对韵律进行单独建模,而是完全依赖于先前时间步的序列信息,在中小规模训练数据(一千多小时或数百小时)的条件下,会出现停顿不够自然、韵律对齐不够稳定等问题.Pause-based TTS相较于Valle,主要是增加了基于BERT的停顿预测,使得句中停顿更加自然,但由于未对自回归生成进行约束,仍然会出现对齐紊乱现象.本文工作除了停顿预测,还额外新增了时长预测模块和注意力判别模块,注意力判别模块接收时长预测模块输出的音素时长信息作为判断参考依据,注意力权重与判断参考依据共同影响音素时长的建

模,从而避免出现对齐紊乱的现象.

4 结 论

本文基于零样本语音合成的韵律建模不理想以及自回归生成存在韵律对齐紊乱的情况,提出了一种基于音素级韵律建模的自回归零样本语音合成方法,以提升韵律的自然度和稳定性.首先在文本前端预测多类停顿,为韵律建模提供参考;然后预测混合高斯时长分布,并通过采样得到参考时长;最后在自回归生成的每个时间步引入注意力判别模块,通过控制文本向量权重,避免出现对齐紊乱现象.与其他零样本合成方法的对比实验表明了本文方法的优越性;消融实验和各模块自身的对比实验进一步证明了本文所提各模块的有效性.此外,本文所提的韵律建模方法相对于原始语音合成模型属于新增模块,增加了模型复杂度,降低了应用时的推理速度,未来拟在维持合成质量的前提下,进一步探索更加轻量化的韵律建模方式,提升合成的速度.具体将从两方面入手,一是尝试对模型进行流式处理,即将整段音频分块输出,目的是实现推理与输出并行;二是尝试使用深度可分离卷积等操作降低模型复杂度,直接提升推理速度.

参考文献

[1] CASANOVA E, WEBER J, SHULBY C, et al. YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone [C]//International Conference on Machine Learning. PMLR, 2022: 2709-2720.

[2] WANG Y X, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[EB/OL]. 2017: 1703.10135. https://arxiv.org/abs/1703.10135v2.

[3] REN Y, RUAN Y J, TAN X, et al. FastSpeech: fast, robust and controllable text to speech[EB/OL]. 2019: 1905.09263. https://arxiv. org/abs/1905.09263v5.

[4] SHEN J, PANG R M, 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: 4779-4783.

[5] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech [C]// International Conference on Machine Learning. PMLR, 2021: 5530-5540.

[6] YANG D, KORIYAMA T, SAITO Y, et al. Duration-aware pause insertion using pre-trained language model for multi-speaker text-to-speech [C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece. IEEE, 2023: 1-5.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. 2018:1810.04805. https://arxiv. org/abs/1810.04805v2.

[8] 岳焕景,多文昕,杨敬钰. 基于邻域自适应注意力的跨域融合语音增强[J]. 湖南大学学报(自然科学版),2023,50(12): 59-68.
YUE H J, DUO W X, YANG J Y. Neighborhood adaptive attention based cross-domain fusion network for speech enhancement [J]. Journal of Hunan University (Natural Sciences), 2023, 50(12): 59-68.(in Chinese)

[9] BORSOS Z, MARINIER R, VINCENT D, et al. AudioLM: a language modeling approach to audio generation [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2023, 31:2523-2533.

[10] MARY N J M S, UMESH S, KATTA S V. S-vectors and TESA: speaker embeddings and a speaker authenticator based on transformer encoder [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30:404-413.

[11] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[12] BETKER J. Better speech synthesis through scaling[EB/OL]. 2023:2305.07243. https://arxiv. org/abs/2305.07243v2.

[13] PAMISETTY G, SRI RAMA MURTY K. Prosody-TTS: an end-to-end speech synthesis system with prosody control [J]. Circuits, Systems, and Signal Processing, 2023, 42(1):361-384.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. 2013: 1301.3781.https://arxiv.org/abs/1301.3781v3.

[15] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. 2014:1412.3555. https://arxiv. org/abs/1412.3555v1.

[16] MCAULIFFE M, SOCOLOF M, MIHUC S, et al. Montreal forced aligner: trainable text-speech alignment using kaldi [C]// Interspeech 2017. ISCA, 2017: 498-502.

[17] LANCUCKI A. Fastpitch: parallel text-to-speech with pitch prediction [C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada. IEEE, 2021: 6588-6592.

[18] VAN DEN OORD A, DIELEMAN S, ZEN H G, et al. WaveNet: a generative model for raw audio [EB/OL]. 2016: 1609.03499. https://arxiv. org/abs/1609.03499v2.

[19] WANG X, TAKAKI S, YAMAGISHI J, et al. A vector quantized variational autoencoder (VQ-VAE) autoregressive neural F_0

model for statistical parametric speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28:157-170.

[20] VEAUX C, YAMAGISHI J, MACDONALD K. SUPERSEDED-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit[J]. 2016.

[21] ZEN H G, DANG V, CLARK R, et al. LibriTTS: a corpus derived from LibriSpeech for text-to-speech [EB/OL]. 2019: 1904.02882.https://arxiv.org/abs/1904.02882v1.

[22] PRATAP V, XU Q T, SRIRAM A, et al. MLS: a large-scale multilingual dataset for speech research [EB/OL]. 2020: 2012.03411. https://arxiv. org/abs/2012.03411v2.

[23] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015:5206-5210.

[24] 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.

[25] DU C P, YU K. Phone-level prosody modelling with GMM-based MDN for diverse and controllable speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 190-201.

[26] MIAO C F, LIANG S, CHEN M C, et al. Flow-TTS: a non-autoregressive network for text to speech based on flow [C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 7209-7213.

[27] LI B, GULATI A, YU J H, et al. A better and faster end-to-end model for streaming ASR [C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada. IEEE, 2021:5634-5638.

[28] CHOI S, HAN S, KIM D, et al. Attentron: few-shot text-to-speech utilizing attention-based variable-length embedding [EB/OL]. 2020: 2005.08484.https://arxiv.org/abs/2005.08484v2.

[29] REN Y, HU C X, TAN X, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[EB/OL]. 2020: 2006.04558. https://arxiv.org/abs/2006.04558v8.

[30] ZHENG Y B, TAO J H, WEN Z Q, et al. BLSTM-CRF based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end[C]//Interspeech 2018. ISCA, 2018: 47-51.

[31] WANG C Y, CHEN S Y, WU Y, et al. Neural codec language models are zero-shot text to speech synthesizers[EB/OL]. 2023: 2301.02111.https://arxiv.org/abs/2301.02111v1.