计算机科学
COMPUTER SCIENCE

基于层次化Conformer的语音合成

吴克伟, 韩超, 孙永宣, 彭梦昊, 谢昭

引用本文

吴克伟, 韩超, 孙永宣, 彭梦昊, 谢昭. 基于层次化Conformer的语音合成[J]. 计算机科学, 2024, 51(2): 161-171.

WU Kewei, HAN Chao, SUN Yongxuan, PENG Menghao, XIE Zhao. Hierarchical Conformer Based Speech Synthesis [J]. Computer Science, 2024, 51(2): 161-171.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100125

基于层次化 Conformer 的语音合成

吴克伟[1,2,3]　韩　超[3]　孙永宣[1,2,3]　彭梦昊[3]　谢　昭[1,2,3]
1 大数据知识工程教育部重点实验室(合肥工业大学)　合肥 230601
2 情感计算与先进智能机器安徽省重点实验室(合肥工业大学)　合肥 230601
3 合肥工业大学计算机与信息学院　合肥 230601
(wu_kewei1984@163.com)

摘　要　语音合成需要将输入语句的文本转换为包含音素、单词和语句的语音信号。现有语音合成方法将语句看作一个整体,难以准确地合成出不同长度的语音信号。通过分析语音信号中蕴含的层次化关系,分别设计基于 Conformer 的层次化文本编码器和基于 Conformer 的层次化语音编码器,并提出了一种基于层次化文本-语音 Conformer 的语音合成模型。首先,该模型根据输入文本信号的长度,构建层次化文本编码器,包括音素级、单词级、语句级文本编码器 3 个层次,不同层次的文本编码器描述不同长度的文本信息;并使用 Conformer 的注意力机制来学习该长度信号中不同时间特征之间的关系。利用层次化的文本编码器,能够找出语句中不同长度需要强调的信息,有效实现不同长度的文本特征提取,缓解合成的语音信号持续时间长度不确定的问题。其次,层次化语音编码器包括音素级、单词级、语句级语音编码器 3 个层次。每个层次的语音编码器将文本特征作为 Conformer 的查询向量,将语音特征作为 Conformer 的关键字向量和值向量,来提取文本特征和语音特征的匹配关系。利用层次化的语音编码器和文本语音匹配关系,可以缓解不同长度语音信号合成不准确的问题。所提模型的层次化文本-语音编码器可以灵活地嵌入现有的多种解码器中,通过文本和语音之间的互补,提供更为可靠的语音合成结果。在 LJSpeech 和 LibriTTS 两个数据集上进行实验验证,实验结果表明,所提方法的梅尔倒谱失真小于现有语音合成方法。

关键词:语音合成;文本编码器;语音编码器;层次化模型;Conformer
中图分类号　TP391

Hierarchical Conformer Based Speech Synthesis

WU Kewei[1,2,3], HAN Chao[3], SUN Yongxuan[1,2,3], PENG Menghao[3] and XIE Zhao[1,2,3]
1 Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
2 Anhui Provincial Key Laboratory of Emotional Computingand Advanced Intelligent Machine, Hefei University of Technology, Hefei 230601, China
3 School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China

Abstract　Speech synthesis requires synthesizing the input speech text into a speech signal containing phonemes, words and utterances. Existing speech synthesis methods consider utterance as a whole, and it is difficult to synthesize different lengths of speech signals accurately. In this paper, we analyze the hierarchical relationships embedded in speech signals, design a Conformer-based hierarchical text encoder and a Conformer-based hierarchical speech encoder, and propose a speech synthesis model based on the hierarchical text-speech Conformer. First, the model constructs hierarchical text encoders according to the length of the input text signal, including three levels of phoneme level, word level, and utterance level text encoders. Each level of text encoder, describes text information of different lengths and uses Conformer's attention mechanism to learn the relationship between different temporal features in the signal of that length. Using the hierarchical text encoder, we can find out the information that needs to be emphasized at different lengths in the utterance, and effectively achieve the extraction of text features at different lengths to alleviate the problem of uncertainty in the duration of the synthesized speech signal. Second, the hierarchical speech encoder includes three levels: phoneme level, word level, and utterance level speech encoder. For each level of speech encoder, the text features is

到稿日期:2022-11-15　返修日期:2023-04-11
基金项目:安徽省重点研究与开发计划(202004d07020004);安徽省自然科学基金(2108085MF203);中央高校基本科研业务费专项资金(PA2021GDSK0072,JZ2021HGQA0219)
This work was supported by the Key Research and Development Program of Anhui Province(2004d07020004), Natural Science Foundation of Anhui Province (2108085MF203) and Special Funds for Basic Scientific Research Operations of Central Universities (PA2021GDSK0072, JZ2021HGQA0219).
通信作者:谢昭(xiezhao@hfut.edu.cn)

used as the query vector of the Conformer, and the speech features are used as the keyword vector and value vector of the Conformer to extract the matching relationship between text features and speech features. The problem of inaccurate synthesis of different length speech signals can be alleviated by using hierarchical speech encoder and text-to-speech matching relations. The hierarchical text-to-speech encoder modeled in this paper can be flexibly embedded into a variety of existing decoders to provide more reliable speech synthesis results through the complementarity between text and speech. Experimental validation is performed on two datasets, LJSpeech and LibriTTS, and experimental results show that the Mel inversion distortion of the proposed method is smaller than that of existing speech synthesis methods.

Keywords Speech synthesis, Text encoder, Speech encoder, Hierarchical model, Conformer

1 引言

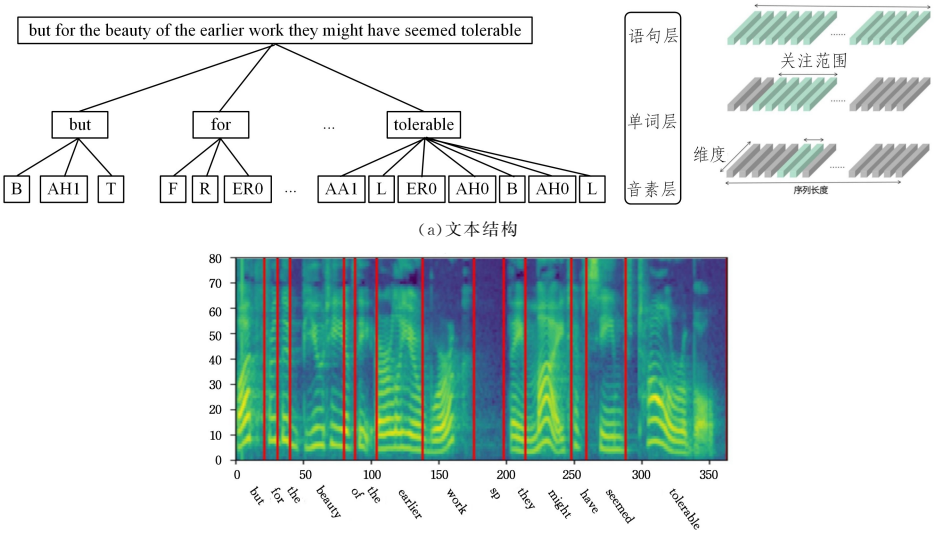
语音合成的目的是合成可理解的自然音频。语音合成作为人机语音交互系统中的核心技术之一，被广泛应用于智能客服、智能车载、智能教育等领域。然而，现有语音合成将文本和语音作为一个整体来看待，忽视了文本和语音中的结构问题。文本和语音的结构由不同层次构成，几个音素组成一个词，多个词组成一句话。因此，语音合成需要提供不同层次的文本和语音信息作为输入，并对这些信息进行建模。

现有语音合成模型[1-7]使用自注意力构建全局依赖关系来改善文本的表征学习。Transformer TTS[2]模型引入Transformer[3]作为编码器，任意两个不同时刻的输入通过自注意机制直接连接，有效地解决了远程依赖问题。但是，在语音合成中，每个单词的发音主要取决于当前的输入单词及其相邻单词。Transformer中多头注意力机制的加权平均操作可能会导致注意力分散，并忽略相邻信号的关系。这种分散可能会导致低估局部信息的重要性，从而导致发音错误。Yang等[4]提出了基于相对位置感知注意力和可学习高斯偏置注意力的两种局部上下文建模方法，以增强自注意力机制的局部性。FastSpeech 1/2[5-6]和FastPitch[7]模型用卷积神经网络替换前馈模块，构建局部依赖关系。上述方法表明多头注意力网络可以从局部建模中受益。

全局特征和局部特征都在语音合成的建模中起到了关键作用。Transformer的自注意力和卷积神经网络各有优势和不足。自注意力机制能够对全局的上下文进行建模，但不擅长提取局部特征；卷积神经网络则相反。Gulati等[8]提出了Conformer结构，其将注意力机制的全局建模能力和CNN的局部建模能力结合起来，同时发挥两者的优势。DelightfulTTS[9]模型的编码器采用了改进的Conformer，以更好地学习模型中的局部和全局相关性。但是，上述方法都是从文本中预测声学条件，而这些因素并不完全包含在文本中。同时，这种预测方式需要从语音中提取声学特征作为标签，现有的算法无法提取全部的声学条件标签。上述模型无法对语音中固有的其他声学条件进行建模，不足以完全预测语音的变化，导致合成语音不够自然。

自然度在很大程度上取决于合成语音的表现力，它由内容、音色、韵律和情感等多种特征决定，这些特征都可以从语音结构中获取。因此，Dai等[10]提出通过一个带有预训练Conformer音频编码器的神经文本-语音模型，从文本-音频数据中自动提取韵律标签。Skerry-Ryan等[11]通过参考编码器从参考音频中提取固定长度的韵律学习表示，将其作为解码器的输入，可以在话语之间传递韵律。AdaSpeech[12]为了处理不同的声学条件，对不同粒度的声学条件进行了韵律建模。然而，上述模型只关注了整体的语音结构，忽略了不同层次对语音结构的影响。同时，上述模型没有考虑不同层次的文本和语音之间的关联，存在文本没有对应正确的声学特征的问题。

图1给出了一句话的文本结构和语音结构。



(a)文本结构

(b)语音结构

图1 语句的文本结构和语音结构

Fig. 1 Text structure and speech structure of sentences

在文本结构中，文本具有多种层次，每种层次具有不同的关注范围。在语音结构中，语音结构和文本结构具有复杂的对应关系。语句中单词对应的 Mel 频谱图用红线分割。“sp”表示说话的停顿，间隔越长停顿越久。图中颜色的深浅表示幅度的大小，颜色越深幅度越大。颜色亮度变化曲线表示语调变化。从图中可以看出：(1)一句话的文本结构是由几个单词组成，单词又由多个音素构成。同样，文本对应的 Mel 频谱图也是由单词和音素一层层构成。因此，语音的表达是多层次的，并且不同层次的注意力关注的范围也不同。(2)图中相同的单词“the”对应着不同的韵律特征。同时单词“work”对应的部分音素被淡化，其中的音素“k”发音被弱化，而更加关注音素“w”和“ER1”。该现象说明需要设计层次化网络，用于关注不同粒度的文本特征，并提取需要强调的文本特征。

针对上述现象，本文设计了一种基于层次化 Conformer 的语音合成网络。该模型针对单词语音持续时间长度变化范围大的问题，设计了层次化文本编码器。该模块使用层次化结构提取语句中单词之间的关系，提供词组级的单词注意力，以及单词中音素之间的关系。该模块使用音素级 Conformer，用于找出单词中不同音素长度的关系；使用单词级 Conformer 来提取词组中单词之间的关系；并提供语句级 Conformer，用于获取语句中整体的文本特征。同时，为了能准确提取语音特征中单词对应的声学特征，该模型设计了层次化语音编码器。该编码器将文本特征和语音特征映射到相同的特征维度，并使用注意力机制进行查询。注意力模块使用文本特征作为查询向量，使用语音特征作为关键字和值向量，来提取文本特征和语音特征的匹配关系。

本文的主要贡献如下：

(1)为了缓解合成的语音信号持续时间长度不确定的问题，本文设计了基于 Conformer 的层次化文本编码器。层次化文本编码器包括音素级、单词级、语句级文本编码器 3 个层次，分别用于描述不同长度的文本信息。每个层次的编码器使用 Conformer 的注意力机制来学习该长度信号中不同时间特征之间的关系，找出其中需要强调的信息，从而准确描述不同持续时间的语音信号。

(2)为了缓解不同长度语音信号合成不准确的问题，本文设计了基于 Conformer 的层次化语音编码器。层次化语音编码器包括音素级、单词级、语句级语音编码器 3 个层次。每个层次的语音编码器将文本特征作为 Conformer 的查询向量，将语音特征作为 Conformer 的关键字向量和值向量，来提取文本特征和语音特征的匹配关系，同时，利用层次化的语音编码器和文本语音匹配关系，来准确实现文本特征查询和语音信号合成。

(3)本文模型的层次化文本-语音编码器可以灵活地嵌入到现有的多种解码器中，通过文本和语音之间的互补，提供更为可靠的语音合成结果。在 LJSpeech 和 LibriTTS 两个数据集上的实验结果表明，本文方法的梅尔倒谱失真小于现有语音合成方法。

2 相关工作

合成自然的语音需要输入文本特征和声学特征。文本特征表示合成语音的内容，声学特征表示如何表达合成的内容。

2.1 文本特征提取

注意力机制被广泛应用于提取文本特征。Tacotron 2[13]引入了位置敏感注意力，使用之前解码处理的累积注意力权重作为一个额外的特征，使得模型在沿着输入序列向前移动时能够保持前后一致，减少了解码过程中潜在的子序列重复或遗漏。He 等[14]提出了逐步单调注意力，其中，在每个解码步骤中，注意力对齐位置最多移动一步，并且不允许跳过任何输入单元。但是，这些工作对注意力进行了约束，限制了注意力本身的全局建模能力。Zheng 等[15]将局部递归神经网络引入 Transformer 中，局部递归神经网络可以对局部结构进行建模，Transformer 可以捕获长期依赖关系。Zhao 等[16]提出了混合轻量级卷积，充分利用了序列的局部结构，并将其与注意力相结合。Liu 等[17]提出了一种具有自动学习的语音表示方式的系统，并联合优化了声学模型和声码器。Morioka 等[18]提出了一种少样本说话人自适应方法，在 Conformer 层之间插入可训练的轻量级模块，称为残余适配器。上述模型改进了多头注意力的局部建模能力，但忽略了文本是由不同层次结构组成的，没有充分考虑不同层次需要关注的范围。Lei 等[19]提出了层次上下文编码器和参考编码器，通过提取有效的上下文信息来预测说话风格，作为解码器输入来提高语音质量。然而，该模型只关注了文本的层次化结构，忽略了不同层次语音结构的影响。

2.2 声学特征提取

上述模型可以从文本中预测声学特征，但是，声学特征不完全包含在文本中。这些模型没有对语音中的声学特征进行建模，存在一对多映射问题，即在持续时间、音高、音量、说话者风格、情感等方面，同一句话有多种对应的语音变化，导致合成语音不够自然。语音合成模型提供声学变化信息作为输入，并对这些变化信息进行建模，能够缓解一对多映射问题。Wang 等[20]引入全局风格标记(Global Style Tokens，GST)，通过对韵律进行解耦，来表达多种样式的控制和传递任务。TP-GST[21]对 GST 进行了两个扩展，不仅可以学习不同的说话风格，且在推理过程中不需要辅助输入就能合成表达性语音。Attentron[22]利用细粒度和粗粒度编码器生成可变尺度嵌入和全局尺度嵌入，来克隆未见的说话人。Parallel Tacotron[23]采用基于变分自编码器的残差编码器，缓解了文本到语音问题的一对多映射的困难，同时允许并行训练和推理。Bae 等[24]提出了一种基于层次的、多尺度变分自编码器的非自回归文本-语音模型来生成具有不同说话风格的自然语音。Chien 等[25]提出了一个层次韵律建模框架，其中音素级韵律预测是基于单词级韵律预测，以结合音素级和单词级韵律预测的优势。然而，韵律建模没有考虑文本特征和声学特征之间的对齐。本文采用层次化语音编码器的交叉注意力对不同层次的声学特征和文本特征进行对齐，找出与当前音素相关的声学特征。

3 层次化语音合成

3.1 模型框架

本文模型的整体设计如图 2 所示。

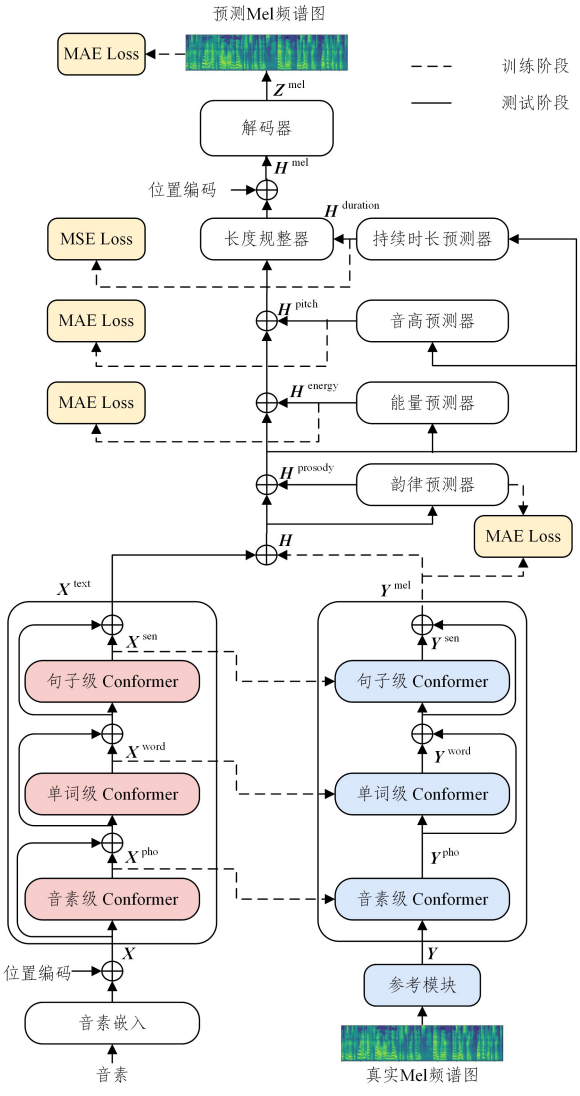


图 2 层次化 Conformer 语音合成网络

Fig. 2 Speech synthesis network with hierarchical Conformer

本文采用 FastSpeech 2 模型作为 Baseline。FastSpeech 2 采用改进后的 Transformer 作为编码器，用卷积神经网络替代前馈网络，这样就丢失了前馈网络将特征映射到高维空间的能力。因此，本文在 FastSpeech 2 中引入 Conformer[8] 结构来构建语音合成模型。本文模型包括层次化文本编码器、层次化语音编码器、韵律预测器、音高预测器、能量预测器、持续时长预测器、长度规整器和解码器。文本编码器将文本序列转换为文本特征。语音编码器将 Mel 频谱图转换为声学特征。韵律预测器预测潜在的声学特征。音高和能量预测器通过音素级特征来预测对应的声学特征。持续时长预测器用来预测每个音素的持续时间。长度规整器对音素级特征序列进行长度调整，对齐 Mel 频谱图的序列长度。解码器通过对齐后的音素级特征序列来预测 Mel 频谱图。

3.2 层次化文本编码器

本文通过固定大小的窗口注意模式来约束编码器的每个自注意层的注意范围。窗口注意力模式使自注意力层关注范围内的数据，并且通过修改窗口大小来控制聚焦范围。层次化文本编码器首先关注文本的局部上下文，随着网络的深入，逐渐关注整体上下文。

如图 3 所示，层次化文本编码器主要由 3 个模块组成，包括音素级 Conformer 模块、单词级 Conformer 模块和语句级 Conformer 模块。文本编码器通过上述 3 个模块学习不同层次的文本特征。3 个模块中的 Conformer 是由前馈网络、多头自注意力和卷积网络组成。前馈模块包括层归一化、两个线性层和非线性激活函数 Swish[26]。多头注意力模块包括层归一化和注意力。多头自注意力将注意力计算限制在一个窗口范围内，同时，该方法可以极大地减少计算量。卷积模块包含两个 1×1 卷积、GLU[27] 激活层、一维深度卷积、批量归一化和 Swish 激活层。

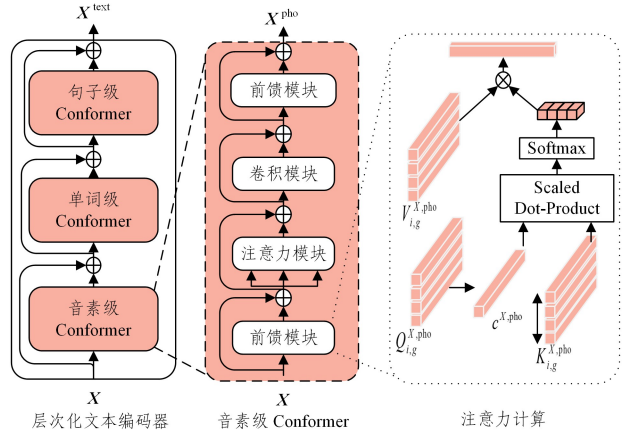


图 3 层次化文本编码器、音素级 Conformer 及其注意力计算

Fig. 3 Hierarchical text encoder, phoneme-level Conformer and its attention calculation

首先通过预处理将文本转换成音素序列，然后使用音素嵌入映射成音素矩阵，再通过位置编码提供句子中音素的位置信息。音素级 Conformer 输入音素矩阵 $\boldsymbol{X}=\{x_1,\cdots,x_L\}$，其中 $\boldsymbol{X}\in R^{L\times d}$，$x_l$ 表示语句中第 l 个音素，L 是语句长度，d 是音素嵌入的维度。然后，通过该模块计算得到音素级文本特征 $\boldsymbol{X}^{\text{pho}}$，计算式如下：

$$\boldsymbol{X}^{\text{pho}}=Conformer_{\text{pho}}(\boldsymbol{X}) \tag{1}$$

其中，$\boldsymbol{X}^{\text{pho}}\in R^{L\times d}$，$Conformer_{\text{pho}}(\cdot)$ 表示音素级 Conformer 模块的计算，其计算式如下：

$$\begin{cases}\boldsymbol{X}^{\text{pho}}=Layernorm\left(\boldsymbol{X}''+\dfrac{1}{2}FFN(\boldsymbol{X}'')\right)\\ \boldsymbol{X}''=\boldsymbol{X}'+Conv(\boldsymbol{X}')\\ \boldsymbol{X}'=MHSA(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{X}})\\ \widetilde{\boldsymbol{X}}=\boldsymbol{X}+\dfrac{1}{2}FFN(\boldsymbol{X})\end{cases} \tag{2}$$

其中，FFN 表示前馈模块，$Conv$ 表示卷积模块，$MHSA$ 表示多头注意力模块。多头注意力模块将输入 $\widetilde{\boldsymbol{X}}$ 划分为 $L/c^{X,\text{pho}}$ 个长度为 $c^{X,\text{pho}}$ 的块，$c^{X,\text{pho}}$ 表示音素级 Conformer 注意力的关注范围。第 g 块 $\boldsymbol{X}_g^{\text{pho}}=\{x_{g1},\cdots,X_{gc^{x,\text{pho}}}\}$，其中 $g\in\{1,\cdots,L/c^{X,\text{pho}}\}$，将其赋值给查询矩阵 $\boldsymbol{Q}_{i,g}^{X,\text{pho}}\in R^{c^{X,\text{pho}}\times d_k}$、关键字矩阵 $\boldsymbol{K}_{i,g}^{X,\text{pho}}\in R^{c^{X,\text{pho}}\times d_k}$ 和值矩阵 $\boldsymbol{V}_{i,g}^{X,\text{pho}}\in R^{c^{X,\text{pho}}\times d_v}$，$i\in\{1,\cdots,h\}$，$h$ 是注意力 head 的数量。d_k 是关键字矩阵和查询矩阵的维度，d_v 是值矩阵的维度。$\boldsymbol{Q}_{i,g}^{X,\text{pho}}$，$\boldsymbol{K}_{i,g}^{X,\text{pho}}$ 和 $\boldsymbol{V}_{i,g}^{X,\text{pho}}$ 用来获取第 g 块内的注意力 $\boldsymbol{head}_{i,g}^{X,\text{pho}}\in R^{c^{X,\text{pho}}\times d_v}$，再将所有块组合起来进行拼接，得到最终的注意力 $\boldsymbol{Head}^{X,\text{pho}}\in R^{L\times d}$。多头注意力的计算

如图 3 所示，计算式如下：

$$\begin{cases}\boldsymbol{Head}^{X,\text{pho}}=Concat(\boldsymbol{Head}_1^{X,\text{pho}},\cdots,\boldsymbol{Head}_h^{X,\text{pho}})\boldsymbol{W}^X\\ \boldsymbol{Head}_i^{X,\text{pho}}=Concat(\boldsymbol{Head}_{i,1}^{X,\text{pho}},\cdots,\boldsymbol{Head}_{i,L/c^{\text{pho}}}^{X,\text{pho}})\\ \boldsymbol{Head}_{i,g}^{X,\text{pho}}=\text{softmax}(\boldsymbol{Q}_{i,g}^{X,\text{pho}}(\boldsymbol{K}_{i,g}^{X,\text{pho}})^{\text{T}}/\sqrt{d_k})\boldsymbol{V}_{i,g}^{X,\text{pho}}\end{cases}\tag{3}$$

其中 $\boldsymbol{head}_i^{X,\text{pho}}\in R^{L\times d_v}$ 是第 i 个 head 的局部注意力，描述以 $c^{X,\text{pho}}$ 的尺度划分音素序列的局部上下文依赖关系。$\boldsymbol{W}^X\in R^{hd_v\times d}$ 是注意力的维度变化参数，用于将串联的特征维度转换到音素的维度。

单词级 Conformer 模块将输入不重叠地划分为 $L/c^{X,\text{word}}$ 个长度为 $c^{X,\text{word}}$ 的块，采用多头注意力模块获取单词之间的依赖关系，最后获得单词级文本特征 $\boldsymbol{X}^{\text{word}}\in R^{L\times d}$，计算式如下：

$$\boldsymbol{X}^{\text{word}}=Conformer_{\text{word}}(\boldsymbol{X}^{\text{pho}}+\boldsymbol{X})\tag{4}$$

单词级 Conformer 学习的文本特征取决于音素级文本特征。单词级 Conformer 的计算方式和音素级 Conformer 相同。

语句级 Conformer 模块采用标准多头注意力的计算方法来获取全局上下文依赖关系，得到语句级文本特征 $\boldsymbol{X}^{\text{sen}}\in R^{L\times d}$，计算式如下：

$$\boldsymbol{X}^{\text{sen}}=Conformer_{\text{sen}}(\boldsymbol{X}^{\text{word}}+\boldsymbol{X}^{\text{pho}}+\boldsymbol{X})\tag{5}$$

语句级 Conformer 学习的文本特征取决于音素级文本特征和单词级文本特征。同样，语句级 Conformer 的计算方式也和音素级 Conformer 相同。

最后，文本编码器通过层次化结构获取文本特征 $\boldsymbol{X}^{\text{text}}$，计算式如下：

$$\begin{cases}\boldsymbol{X}^{\text{text}}=\boldsymbol{X}^{\text{sen}}+\boldsymbol{X}^{\text{word}}+\boldsymbol{X}^{\text{pho}}+\boldsymbol{X}\\ \boldsymbol{X}^{\text{sen}}=Conformer_{\text{sen}}(\boldsymbol{X}^{word}+\boldsymbol{X}^{\text{pho}}+\boldsymbol{X},c^{\boldsymbol{X},\text{sen}})\\ \boldsymbol{X}^{\text{word}}=Conformer_{\text{word}}(\boldsymbol{X}^{\text{pho}}+\boldsymbol{X},c^{\boldsymbol{X},\text{word}})\\ \boldsymbol{X}^{\text{pho}}=Conformer_{\text{pho}}(\boldsymbol{X},c^{\boldsymbol{X},\text{pho}})\end{cases}\tag{6}$$

其中文本特征 $\boldsymbol{X}^{\text{text}}\in R^{L\times d}$。每个层次的编码器使用 Conformer 的注意力机制来学习得到不同层次的文本特征，再融合不同层次的文本特征。

3.3 层次化语音编码器

在语音合成中，由于输入文本缺乏足够的声学特征来预测目标语音，因此，该模型无法完全预测语音的变化，导致合成语音不够自然。解决这一问题的方法是提供相应的声学条件作为输入，使模型学习合理的文本到语音的映射。然而，之前的研究将语音信号视为一个整体，从而忽略了固有的语音结构。本文使用层次化的语音编码器和文本语音匹配关系，来准确实现文本信号查询和语音信号合成。

语音编码器输入的帧级声学特征来自参考模块的输出。参考模块[11]将 Mel 频谱图输入到卷积层和 GRU 层，输出帧级声学特征，再将帧级声学特征进行线性变换，由 Mel 维度变换成音素级文本特征的维度。参考模块从 Mel 频谱图提取帧级声学特征 $\boldsymbol{Y}\in R^{L_{\text{mel}}\times d}$ 作为层次化语音编码器的输入，其中 L_{mel} 是 Mel 频谱图的长度。

如图 4 所示，层次化语音编码器的模型体系结构依次为音素级 Conformer 模块、单词级 Conformer 模块和语句级 Conformer 模块。

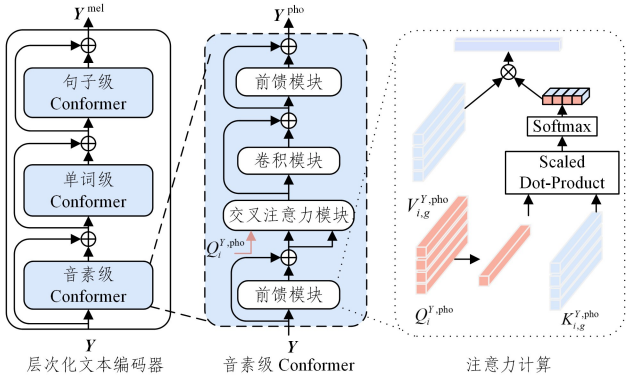


图 4 层次化语音编码器、音素级 Conformer 及其注意力计算

Fig. 4 Hierarchical speech encoder, phoneme-level Conformer and its attention calculation

音素级 Conformer 模块包括前馈模块、交叉注意力模块、卷积模块和层归一化组成。前馈模块将声学特征数据映射到高维空间，以便提取需要的信息，再映射回原来的空间。因此，前馈模块从声学特征提取主要的声学信息。

交叉注意力模块用于对齐音素级文本特征和声学特征，找出与当前音素级文本特征最相关的声学特征，得到音素级声学特征。层归一化用于得到关键字矩阵和值矩阵，查询矩阵是文本编码器中音素级 Conformer 提取的音素级文本特征。

卷积模块从声学特征学习上下文信息。随后，前馈模块和层归一化产生最终的声学特征 $\boldsymbol{Y}^{\text{pho}}$，计算式如下：

$$\boldsymbol{Y}^{\text{pho}}=Conformer_{\text{pho}}(\boldsymbol{Y},\boldsymbol{X}^{\text{pho}})\tag{7}$$

其中，$\boldsymbol{Y}^{\text{pho}}\in R^{L\times d}$，$Conformer_{\text{pho}}(\cdot)$ 表示音素级 Conformer 模块的计算，其计算式如下：

$$\begin{cases}\boldsymbol{Y}^{\text{pho}}=Layernorm\left(\boldsymbol{Y}''+\frac{1}{2}FFN(\boldsymbol{Y}'')\right)\\ \boldsymbol{Y}''=\boldsymbol{Y}'+Conv(\boldsymbol{Y}')\\ \boldsymbol{Y}'=CMHSA(X^{\text{pho}},\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{Y}})\\ \widetilde{\boldsymbol{Y}}=\boldsymbol{Y}+\frac{1}{2}FFN(\boldsymbol{Y})\end{cases}\tag{8}$$

其中，FFN 表示前馈模块，$Conv$ 表示卷积模块，$CMHSA$ 表示交叉注意力模块。交叉注意力模块将输入 $\widetilde{\boldsymbol{Y}}$ 划分为 $L_{\text{mel}}/c^{Y,\text{pho}}$ 个长度为 $c^{Y,\text{pho}}$ 的块，$c^{Y,\text{pho}}$ 表示注意力的关注范围，范围大小为文本编码器中 $c^{X,\text{pho}}$ 所对应的语音序列长度。因此，查询矩阵 $\boldsymbol{Q}_i^{Y,\text{pho}}$ 为音素级文本特征 $\boldsymbol{X}^{\text{pho}}$，关键字矩阵 $\boldsymbol{K}_{i,g}^{Y,\text{pho}}$ 和值矩阵 $\boldsymbol{V}_{i,g}^{Y,\text{pho}}$ 都为 $\widetilde{\boldsymbol{Y}}$ 划分后的第 g 块。

单词级 Conformer 模块的输入为声学特征 $\boldsymbol{Y}^{\text{pho}}$ 和单词级文本特征 $\boldsymbol{X}^{\text{word}}$。其将单词级文本特征作为查询向量，将声学特征作为关键字向量和值向量，来提取文本特征和语音特征的匹配关系。因此，语音编码器中单词级 Conformer 可以得到单词级声学特征 $\boldsymbol{Y}^{\text{word}}\in R^{L\times d}$。

语句级 Conformer 模块的输入为声学特征 $\boldsymbol{Y}^{\text{word}}$ 和语句级文本特征 $\boldsymbol{X}^{\text{sen}}$。语句级 Conformer、单词级 Conformer 和音素级 Conformer 的计算过程相同，语句级 Conformer 计算获得语句级声学特征 $\boldsymbol{Y}^{\text{sen}}\in R^{L\times d}$。

将不同层次的声学特征与层次化文本编码器得到的文本

特征 $\boldsymbol{X}^{\text{text}}$ 相加，获得特征序列 $\boldsymbol{H}\in R^{L\times d}$，计算式如下：

$$\begin{cases}\boldsymbol{H}=\boldsymbol{Y}^{\text{mel}}+\boldsymbol{X}^{\text{text}}\\ \boldsymbol{Y}^{\text{mel}}=\boldsymbol{Y}^{\text{sen}}+\boldsymbol{Y}^{\text{word}}+\boldsymbol{Y}^{\text{pho}}+\boldsymbol{Y}^{\text{frame}}\\ \boldsymbol{Y}^{\text{sen}}=conformer_{\text{sen}}(\boldsymbol{Y}^{\text{word}},\boldsymbol{X}^{\text{sen}})\\ \boldsymbol{Y}^{\text{word}}=conformer_{\text{word}}(\boldsymbol{Y}^{\text{pho}},\boldsymbol{X}^{\text{word}})\\ \boldsymbol{Y}^{\text{pho}}=conformer_{\text{pho}}(\boldsymbol{Y},\boldsymbol{X}^{\text{pho}})\end{cases}\tag{9}$$

其中，$\boldsymbol{Y}^{\text{mel}}\in R^{L\times d}$ 为层次化语音编码器提取的声学特征。每个层次的语音编码器获取不同粒度的声学特征，并且提取文本特征和语音特征的匹配关系。利用层次化的语音编码器和文本语音匹配关系，可以准确实现文本信号查询和语音信号合成。

3.4 预测器

预测器主要由韵律预测器、能量预测器、音高预测器和持续时间预测器组成。预测器通过将韵律、能量、音高和持续时间添加到文本序列中，为语音合成的一对多映射问题提供声学信息。

韵律预测器由两层时间卷积网络组成，每个卷积网络之后是层归一化，以及最后的线性层。韵律预测器的输入为上一阶段的音素序列 $\boldsymbol{H}$，韵律预测器使用两层带正则项的时间卷积来学习韵律特征，使用语音编码器的声学特征作为监督信号。然后，通过线性层将韵律特征投影到输出序列中，得到对应的预测韵律 $\boldsymbol{H}^{\text{prosody}}\in R^{L\times d}$。

能量预测器、音高预测器和持续时长预测器的模型结构与韵律预测器相同。能量预测器可以预测能量 $\boldsymbol{H}^{\text{energy}}\in R^{L\times d}$，音高预测器可以预测音高 $\boldsymbol{H}^{\text{pitch}}\in R^{L\times d}$。将韵律、能量和音高加入到文本特征后，再由持续时长预测器获得每个音素的对应时长 $\boldsymbol{H}^{\text{duration}}\in R^{L\times 1}$。然后，长度规整器根据预测的时长来扩展文本特征长度，使其编码器的文本输入长度和 Mel 频谱相同，得到扩展后的文本特征 $\boldsymbol{H}^{\text{mel}}\in R^{L_{\text{mel}}\times d}$。最后，将扩展后的文本特征送入解码器来获得 Mel 频谱图 $\boldsymbol{Z}^{\text{mel}}\in R^{L_{\text{mel}}\times d_{\text{mel}}}$，其中 d_{mel} 是 Mel 频谱图的维度。

3.5 损失函数

本文方法分两个阶段进行训练。首先，本文模型训练除韵律预测器以外的其他模块。在训练中，计算预测的 Mel 频谱图和真实的 Mel 频谱图之间的 MAE 损失。能量预测器和音高预测器由预测值和真实值使用 MAE 损失进行训练。持续时长预测器由预测值和真实值使用 MSE 损失进行训练。此时，损失函数如下：

$$\begin{cases}Loss_1=L_{\text{mel}}+L_{\text{duration}}+L_{\text{pitch}}+L_{\text{energy}}\\ L_{\text{mel}}=\dfrac{1}{m}\sum\limits_{i=1}^{m}|(\boldsymbol{Z}_i^{\text{mel}}-\boldsymbol{gt}_i^{\text{mel}})|\\ L_{\text{duration}}=\dfrac{1}{m}\sum\limits_{i=1}^{m}(\boldsymbol{H}_i^{\text{duration}}-\boldsymbol{gt}_i^{\text{duration}})^2\\ L_{\text{pitch}}=\dfrac{1}{m}\sum\limits_{i=1}^{m}|(\boldsymbol{H}_i^{\text{pitch}}-\boldsymbol{gt}_i^{\text{pitch}})|\\ L_{\text{energy}}=\dfrac{1}{m}\sum\limits_{i=1}^{m}|(\boldsymbol{H}_i^{\text{energy}}-\boldsymbol{gt}_i^{\text{energy}})|\end{cases}\tag{10}$$

其中，$\boldsymbol{gt}^{\text{energy}}\in R^{L\times d}$，$\boldsymbol{gt}^{\text{pitch}}\in R^{L\times d}$，$\boldsymbol{gt}^{\text{duration}}\in R^{L\times 1}$，$\boldsymbol{gt}^{\text{mel}}\in R^{L_{\text{mel}}\times d_{\text{mel}}}$ 为真实值，m 表示样本数。L_{mel} 是预测 Mel 频谱图和真实 Mel 频谱图之间的 MAE 损失，用于合成和真实语音更相近的 Mel 频谱图；L_{duration} 是预测持续时间和真实持续时间之间的 MSE 损失，用于保证文本和语音之间的对齐；L_{pitch} 是预测音高和真实音高之间的 MAE 损失，用于减小合成语音和真实语音的音高差距；L_{energy} 是预测能量和真实能量之间的 MAE 损失，用于减小合成语音和真实语音的能量差距。

然后联合训练本文模型和韵律预测器，其中语音编码器的输出用作监督信号，以训练韵律预测器。韵律预测器由预测韵律和语音编码器获得的韵律使用 MAE 损失进行训练。本文模型的损失函数如下：

$$\begin{cases}Loss=Loss_1+L_{\text{phone}}\\ L_{\text{phone}}=\dfrac{1}{m}\sum\limits_{i=1}^{m}|(\boldsymbol{H}_i^{\text{prosody}}-\boldsymbol{H}_i^{\text{mel}})|\end{cases}\tag{11}$$

其中，L_{phone} 是预测韵律向量和从声学编码器中提取的向量之间的 MAE 损失，以使模型在无标签的情况下合成接近真实韵律的语音。

模型的整个训练流程如算法 1 所示。

算法 1 训练过程

输入：数据集〈文本，语音〉

输出：Mel 频谱图

1. 通过预处理将文本转换为音素序列，音素序列通过音素嵌入和位置编码得到音素矩阵。
2. 利用 3.2 节的层次化文本编码器学习不同层次的文本特征。
3. 语音通过预处理转换为 Mel 频谱图，Mel 频谱图通过参考模块得到声学特征。
4. 利用 3.3 节的层次化语音编码器提供不同层次的声学特征。
5. 利用 3.4 节的能量预测器、音高预测器、持续时长预测器和解码器，依次获得能量特征、音高特征、持续时长和 Mel 频谱图的预测值。
6. 根据式(10)的损失函数，计算预测值和真实值的损失，从而训练编码器、预测器和解码器。
7. 使用训练后的语音编码器提取的声学特征作为韵律预测器的标签，根据式(11)的损失函数来训练韵律预测器和本文模型。

3.6 测试阶段

在测试阶段，无法使用语音作为参考，因此，需要从文本输入中预测韵律。模型在测试阶段的流程如算法 2 所示。

算法 2 测试过程

输入：文本

输出：Mel 频谱图

1. 模型使用层次化文本编码器学习不同层次的文本特征。
2. 模型通过预测器来预测语音中的韵律、能量、音高和持续时长。
3. 长度规整器根据持续时长对齐输入和输出的长度。
4. 将上述预测结果加入文本特征中，通过解码器来预测 Mel 频谱图。

4 实验分析

4.1 数据集

本文选取了 LJSpeech[28] 和 LibriTTS[29] 两个数据集。LJSpeech 数据集是一位女性说话者所说的 13 100 个英语音频片段，包含 25 h 的语音数据和相应的文本。LJSpeech 数据集分为 train，dev 和 test 3 组：train 用于训练(共 12 500 个样本)，dev 用于验证(共 100 个样本)，test 用于测试(共 500 个样本)。LibriTTS 数据集是一个多说话人语料库，包含 586 h 的语音数据和相应的文本。LibriTTS 数据集默认分为 train，

dev 和 test 3 组：train-clean-100 和 train-clean-360 用于训练（共 149736 个样本），dev-clean 用于验证（共 5736 个样本），test-clean 用于测试（共 4837 个样本）。

4.2 实验设置

在预处理过程中，采用的 CPU 配置为 Inter Core i9-10900X，其主频为 3.70 GHz×10 cores，缓存为 RAM 62 GB，使用 g2p 将文本序列转换为音素序列。LJSpeech 的采样率为 22050 kHz，LibriTTS 的采样率为 24000 kHz，帧长为 1024，跳跃大小为 256，窗函数采用 Hann 窗，将原始波形转换为 Mel 频谱图。

在训练过程中，本文实验采用显卡 NVIDIA GeForce RTX 3090，操作系统为 Linux 18.04。深度学习框架为 PyTorch。训练时，批处理大小设置为 64，学习率初始化为 0.0625，学习率变化沿用 FastSpeech 2 的方法，训练的第一阶段 step 为 60000，第二阶段 step 为 840000。使用 Adam 作为优化器，$\beta_1=0.9$，$\beta_2=0.98$，$\varepsilon=10^{-9}$。

4.3 评价标准

评价指标分为主观评价和客观评价两种。主观评价采用平均意见分数（Mean Opinion Score，MOS）。在 MOS 测试中，将 Mel 频谱图使用 HiFi-GAN[29] 声码器转换为语音波形。测试者听完每个合成结果后，会对合成语音的自然度进行 5 分制打分（1 为劣，2 为差，3 为中，4 为良，5 为优），打分间隔为 0.5。客观评价采用梅尔倒谱失真（Mel Cepstral Distortion，MCD）来描述语音质量。MCD 表示预测语音的 Mel 频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）特征与真实语音的 MFCC 特征的差距，其计算式如下：

$$\begin{cases} MCD(\boldsymbol{Z},\hat{\boldsymbol{Z}})=\dfrac{10\sqrt{2}}{\ln(10)}\sqrt{\sum_{t=1}^{T}\|\boldsymbol{Z}_t-\hat{\boldsymbol{Z}}_t\|} \\ \hat{\boldsymbol{Z}}=\sum_{r=1}^{R}\boldsymbol{Z}^{\mathrm{mel}}[r]\cos\left(\dfrac{\pi(r-0.5)n}{R}\right),n=1,2,\cdots,N \end{cases} \tag{12}$$

其中，$\boldsymbol{Z}$ 表示真实语音的 MFCC，$\hat{\boldsymbol{Z}}$ 表示预测语音的 MFCC，T 表示时间，R 是滤波器个数，N 是 MFCC 阶数。

4.4 对比实验

本文的对比方法可以分为以下两组：(1)带有语音编码器的模型，包括 GST[20]，Parallel Tacotron[23]，AdaSpeech[12] 和 HC-TTS[19]。(2)不带语音编码器的模型，包括 Tacotron 2[13]，Transformer TTS[2]，FastSpeech[5]，FastSpeech 2[6] 和 FastPitch[7]。表 1 列出了现有主流方法在 LJSpeech[1)] 和 LibriTTS[28] 数据集上的语音合成结果。为了避免实验的偶然性，进行了多次实验，以 95% 的置信度进行 MOS 打分。

表 1 不同模型的语音质量对比

Tabel 1 Comparison of speech quality of different models

Methods		Baseline	LJSpeech		LibriTTS	
			MOS	*MCD*	*MOS*	*MCD*
文本编码器	Tacotron 2[13]	Tacotron	3.77±0.08	7.17	3.80±0.07	7.21
	TransformerTTS[2]	Transformer	3.74±0.12	7.43	3.73±0.10	7.52
	FastSpeech[5]	Transformer TTS	4.02±0.08	6.96	4.04±0.08	6.91
	FastSpeech 2[6]	FastSpeech	4.32±0.15	6.83	4.34±0.12	6.75
	FastPitch[7]	FastSpeech	4.12±0.06	6.87	4.15±0.07	6.88
文本编码器＋语音编码器	GST[20]	Tacotron 2	3.96±0.08	6.89	3.99±0.10	6.90
	ParallelTacotron[23]	Tacotron 2	4.23±0.09	6.46	4.21±0.08	6.52
	AdaSpeech[12]	FastSpeech 2	4.40±0.11	6.41	4.39±0.12	6.44
	HC-TTS[19]	FastSpeech 2	4.41±0.08	6.39	4.35±0.10	6.40
	本文模型	FastSpeech 2	**4.45±0.08**	**6.22**	**4.47±0.08**	**6.16**

由表 1 可以看出，本文方优于其他语音合成方法。具体地：(1)带有语音编码器的模型总体上优于不带语音编码器的模型，这说明语音编码器提供了文本中缺少的声学特征，能够缓解一对多映射问题，提高合成语音的表达能力。(2)本文方法的语音效果优于 GST[20] 方法，这是因为 GST[20] 提供的粗粒度韵律标签的信息不够准确，影响了模型生成清晰的语音。本文提供不同粒度的韵律，这些韵律能够提供更具体的声学信息，说明本文方法可以有效地帮助模型生成自然的语音。(3)本文方法合成的语音质量同样高于 Parallel Tacotron[23]，AdaSpeech[12] 和 HC-TTS[19]。这是由于 Parallel Tacotron[23] 从变分自编码器获取的潜在变量产生的语音不自然且不连续。潜在变量是从高斯先验分布中独立地对每个单词或音素进行采样，可能无法正确建模时间依赖性。AdaSpeech[12] 采用一维卷积建模具有时间依赖性的声学特征。HC-TTS[19] 忽略了不同层次声学特征和文本内容具有相关性。本文方法通过语音编码器进行层次化建模，获取不同粒度的声学特征，由交叉注意力将其和相同粒度的文本特征进行对齐来确定对应的韵律。实验证明了本文方法对提升语音质量的有效性。

4.5 消融实验

4.5.1 Conformer 结构与 Transformer 结构

本文基准模型为 FastSpeech 2[6] 模型。将 FastSpeech 2 改进后的 Transformer 作为编码器，采用卷积神经网络替代前馈网络。本文采用 Conformer 结构，并使用层次化方法对编码器进行改进，以提高合成语音质量。Conformer[8] 通过一种 Macaron 的结构，将卷积模块和自注意力模块夹在两个前馈神经网络中间，从而学习更好的非线性特征。

4.5.2 层次化结构对语音合成的影响

为了验证层次化文本编码器和层次化语音编码器对本文模型性能的影响程度，本文在 LJSpeech[26] 和 LibriTTS[27] 数据集上进行了消融实验分析，并对这些消融实验进行 MOS 和 MCD 评估，结果如表 2 所列。其中，Params 表示模型参数量，GFLOPs 表示模型浮点运算数。

1) https://keithito.com/LJ-Speech-Dataset/

表 2 文本编码器和语音编码器都采用层次化结构的影响

Tabel 2 Impact of using hierarchical structure for both text and speech encoders

FastSpeech 2[6]	语句级编码器	音素级编码器	单词级编码器	*Params*/(×10⁶)	*GFLOPs*	LJSpeech		LibriTTS	
						MOS	*MCD*	*MOS*	*MCD*
√	—	—	—	32.22	18.58	4.32±0.15	6.83	4.34±0.12	6.75
√	√	—	—	32.22	18.58	4.38±0.08	6.44	4.37±0.07	6.41
√	√	√	—	33.05	20.04	4.41±0.08	6.36	4.43±0.08	6.35
√	√	√	√	33.05	20.04	**4.45±0.08**	**6.22**	**4.47±0.08**	**6.16**

从表 2 可以看出：(1)基准模型采用 Transformer 结构作为编码器和解码器，将文本和语音作为一个整体来看待，忽视了文本和语音中的结构问题，导致合成的语音质量低。(2)在基准模型上采用语句级文本编码器和语音编码器，等同于将 FastSpeech 2 中 d Transformer 结构改为 Conformer 结构。模型合成语音质量有所提升，说明语句级文本编码器和语句级语音编码器能够有效地表达全局信息特征，并且能够对齐不同粒度的声学特征和文本特征，找出与当前文本特征相关的声学特征来确定韵律。(3)加入音素级编码器后，音素级编码器能够学习细粒度的本文特征和声学特征，提供局部信息。(4)最后同时采用语句级、音素级和单词级的编码器后，学习不同时间特征之间的关系，能够找出其中需要强调的信息，准确描述不同持续时间的语音信号。消融实验表明使用层次化文本编码器和层次化语音编码器可以提高模型合成语音的质量。

4.5.3 本文方法结合不同模型的效果

将本文方法与现有模型相结合来验证其有效性。表 3 中，语音合成模型分别采用 FastSpeech 2[6] 和 FastPitch[7]。

表 3 以 Transformer 为编码器的模型采用本文方法的效果

Table 3 Effect of the model with Transformer as encoder using theproposed method

现有模型	层次化文本编码器	层次化语音编码器	LJSpeech		LibriTTS	
			MOS	*MCD*	*MOS*	*MCD*
FastPitch[7]	—	—	4.12±0.06	6.87	4.15±0.07	6.88
FastPitch[7]	√	—	4.20±0.07	6.72	4.22±0.08	6.71
FastPitch[7]	—	√	4.31±0.08	6.65	4.30±0.07	6.64
FastPitch[7]	√	√	**4.35±0.08**	**6.61**	**4.37±0.09**	**6.60**
FastSpeech 2[6]	—	—	4.32±0.15	6.83	4.34±0.12	6.75
FastSpeech 2[6]	√	—	4.35±0.09	6.71	4.35±0.07	6.70
FastSpeech 2[6]	—	√	4.40±0.08	6.44	4.39±0.09	6.52
FastSpeech 2[6]	√	√	**4.45±0.08**	**6.22**	**4.47±0.08**	**6.16**

从表 3 可以看出：(1)上述两个模型均采用 Transformer 作为编码器和解码器。(2)使用层次化文本编码器后，两个模型合成的语音质量都有所提高，这是由于它们将语句看作一个整体，难以准确地合成不同长度的语音信号。层次化文本编码器描述不同长度的文本信息，并使用 Conformer 来学习该长度信号中不同时间特征之间的关系。(3)引入层次化语音编码器后，两个模型合成的语音质量都大幅提升。由于上述方法都是从文本中预测声学条件，而这些因素并不完全包含在文本中，因此，层次化语音编码器不仅提供了语音中的声学信息，还描述了不同长度的声学信息，以合成自然的语音。(4)本文模型结合层次化文本编码器和层次化语音编码器来提取文本特征和语音特征的匹配关系，利用层次化的语音编码器和文本语音匹配关系来准确实现文本信号查询和语音信号合成，在提升 FastSpeech 2[6] 模型的同时，也能提升其他现有模型的性能，说明层次化文本编码器和层次化语音编码器能够有效提高语音合成的质量。

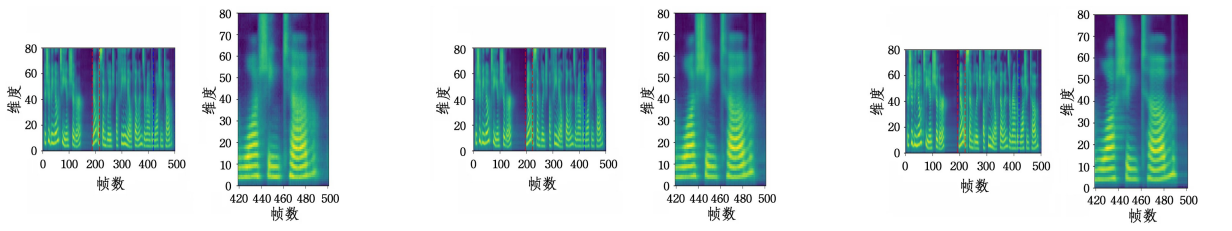
4.6 可视化分析

为了进一步证实本文模型的有效性，本文对合成语音的 Mel 频谱、音高和能量进行可视化分析。本文从测试集中选取 1 例数据“there should be no tea and sugar, no assemblage of female felons around the washing-tub”，图 5 展示了 FastSpeech 2、FastSpeech 2 + 层次化文本编码器、FastSpeech 2 + 层次化语音编码器、本文模型的 Mel 频谱图(见图5(d)左边) 和单词“washing-tub”的 Mel 频谱图(见图 5(d)右边)，以及样本真实值。图 6 和图 7 的左边分别展示了音高和能量在不同模型下和真实值之间的差距，右边放大了单词“washing-tub”的音高和能量图。

从可视化结果可以看出：

(1)现有模型关注文本的整体结构，无法准确估计细粒度单词的持续时间长度。图 5(a) 给出了单词“No”的 Mel 频谱图在第 200 帧的范围，如图中两条红色虚线所示。图 5(d) 给出了真实的 Mel 频谱图，同样用红色虚线表示“No”的范围。合成“No”的发音出现在第 200 帧之前，导致持续时间长度与真实值产生差距。本文引入层次化文本编码器来准确描述不同持续时间的语音信号，从而预测正确的发音时间，如图 5(b)所示。

(2)现有模型没有考虑不同层次的文本和语音之间的关联，存在文本没有对应正确的声学特征的问题。图 5(a)给出了该例句的 Mel 频谱图，其中“washing-tub”单词中的后缀“tub”的 Mel 频谱图在第 460-480 帧。“tub”的 Mel 频谱图在第 30－40 维对应的频率上都有响应，相互干扰而成为语音合成的噪声，使得音高和真实值之间产生了较大的差距，如图 6(a) 所示。本文引入层次化语音编码器，利用文本和语音的匹配关系，来准确实现文本特征和声学特征对齐。该方法提供了“tub”的主要频率信息。使用主要频率信息后，“tub”的 Mel 频谱图增强了在第 30－40 维中的主要频率响应，并抑制了无关频率响应，有利于产生清晰的语音信号，如图 5(c)所示。当去除无关的频率响应后，“tub”的音高与真实值的差距缩小，从而产生准确的语音信号，如图 6(c) 所示。本文方法也缩小了能量预测值和真实值之间的差距，如图 7 所示。

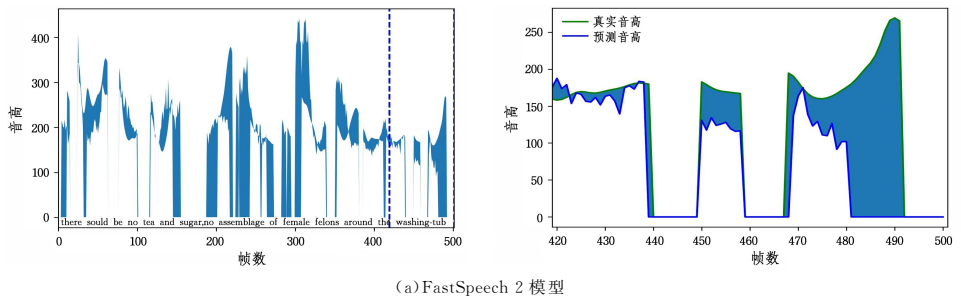


(a)FastSpeech 2 模型　(b)FastSpeech 2(with 层次化文本编码器) 模型　(c)FastSpeech 2(with 层次化语音编码器) 模型

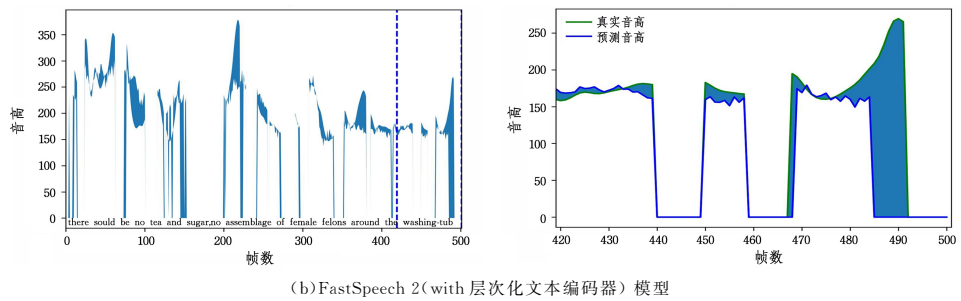
(d)本文模型　(e)Ground Truth

图 5　Mel 频谱图(电子版为彩图)

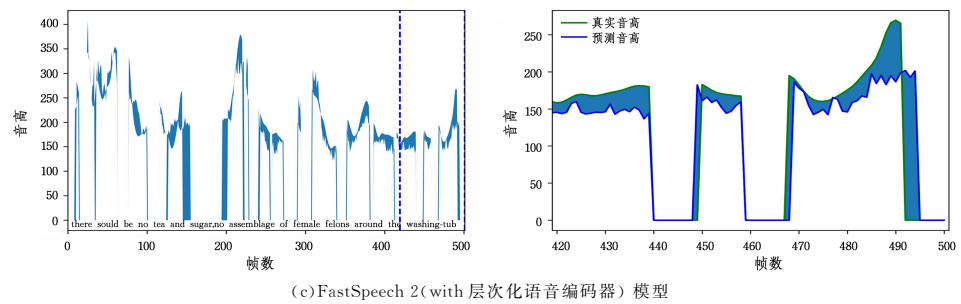
Fig. 5　Mel spectrograms



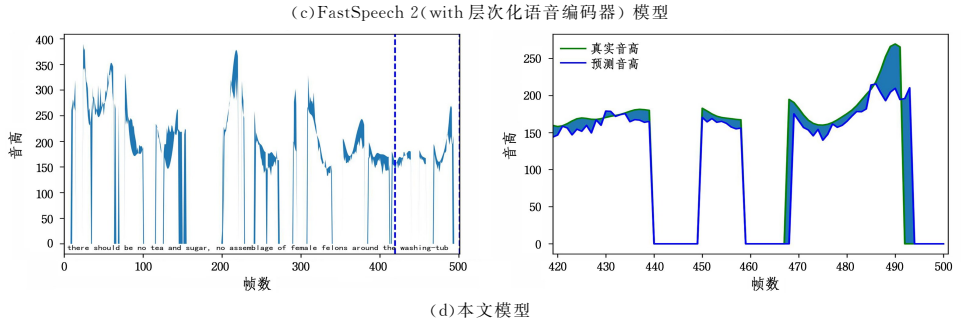
(a)FastSpeech 2 模型



(b)FastSpeech 2(with 层次化文本编码器) 模型



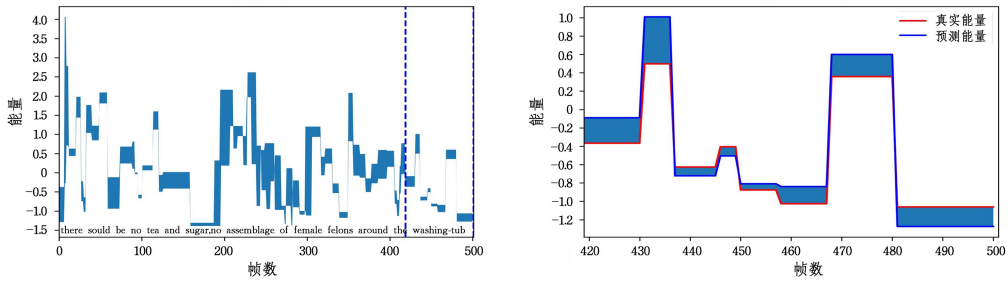
(c)FastSpeech 2(with 层次化语音编码器) 模型



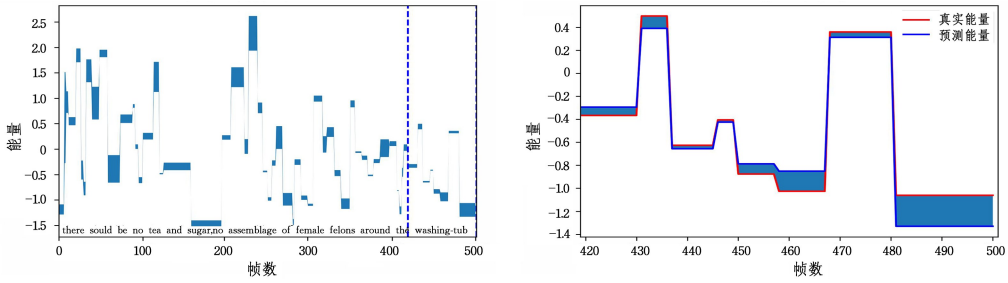
(d)本文模型

图 6　音高图

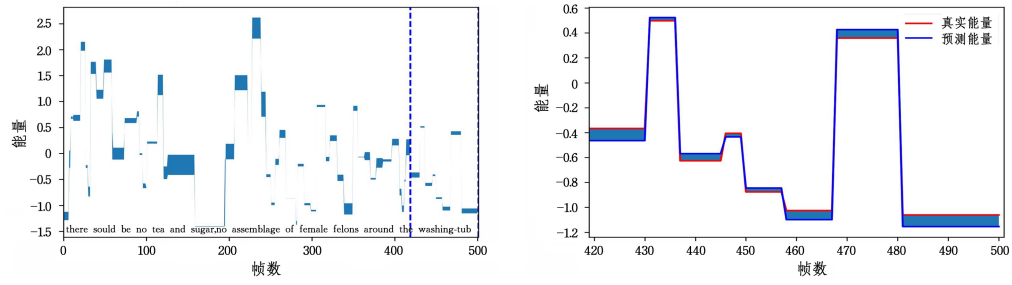
Fig. 6　Pitch charts



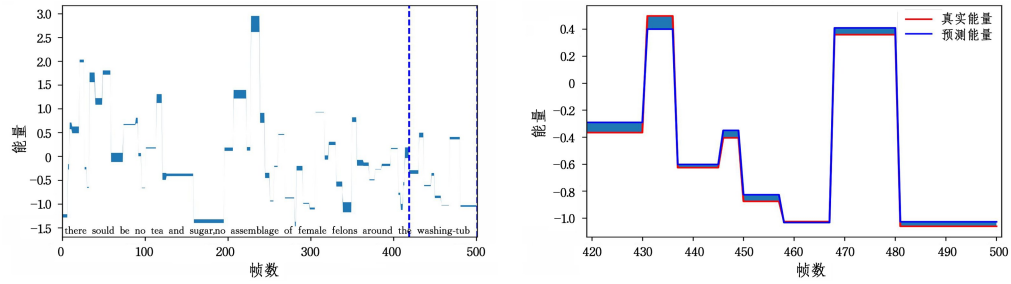
(a)FastSpeech 2 模型



(b)FastSpeech 2(with 层次化文本编码器)模型



(c)FastSpeech 2(with 层次化语音编码器)模型



(d)本文模型

图 7 能量图

Fig. 7 Energy charts

结束语 本文对现有的语音合成模型难以准确地预测出不同长度的语音信号的问题进行了研究。本文分别设计了基于 Conformer 的层次化文本编码器和基于 Conformer 的层次化语音编码器,并提出了一种基于层次化文本-语音 Conformer 的语音合成模型。模型通过层次化文本编码器来描述不同长度的文本信息,通过层次化语音编码器来提取文本特征和语音特征的匹配关系。利用层次化的语音编码器和文本语音匹配关系,来准确实现文本信号查询和语音信号合成。本文方法可以灵活地嵌入到现有的多种解码器中,通过文本和语音之间的互补,提供更为可靠的语音合成结果。在 LJSpeech 和 LibriTTS 两个数据集上的实验结果表明,本文方法能够提高合成语音的自然度。

语音合成需要大量高质量的文本和语音数据来训练模型。但是,大多数语言缺乏训练数据。因此,如何实现小样本情况下的语音合成是后续研究的方向之一。此外,语音合成模型在长文本数据集情况下仍然存在错误对齐导致的单词跳跃和重复问题,这也是后续的研究工作。

参 考 文 献

[1] AN X, DAI Z B, LI Y, et al. An end-to-end speech synthesis method based on BERT[J]. Computer Science, 2022, 49(4): 221-226.

[2] LI N, LIU S, LIU Y, et al. Neural speech synthesis with transformer network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6706-6713.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// NIPS. 2017: 5998-6008.

[4] YANG S, LU H, KANG S, et al. On the localness modeling for the self-attention based end-to-end speech synthesis[J]. Neural

networks，2020，125：121-130.
[5] REN Y，RUAN Y，TAN X，et al. Fastspeech：Fast，robust and controllable text to speech[C]// NeurIPS. 2019：3165-3174.
[6] REN Y，HU C，TAN X，et al. Fastspeech 2：Fast and high-quality end-to-end text to speech[C]// 9th International Conference on Learning Representations. Virtual Event：OpenReview. net，2021.
[7] ŁAŃCUCKI A. Fastpitch：Parallel text-to-speech with pitch prediction[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2021). IEEE，2021：6588-6592.
[8] GULATI A，QIN J，CHIU C C，et al. Conformer：Convolution-augmented transformer for speech recognition[C]// 21st Annual Conference of the International Speech Communication Association. Shanghai：ISCA，2020：5036-5040.
[9] LIU Y，XU Z，WANG G，et al. Delightfultts：The microsoft speech synthesis system for blizzard challenge 2021[J]. arXiv：2110. 12612，2021.
[10] DAI Z，YU J，WANG Y，et al. Automatic Prosody Annotation with Pre-Trained Text-Speech Model[C]// 23rd Annual Conference of the International Speech Communication Association. Incheon：ISCA，2022：5513-5517.
[11] SKERRY-RYAN R J，BATTENBERG E，XIAO Y，et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron[C]// International Conference on Machine Learning. PMLR，2018：4693-4702.
[12] CHEN M，TAN X，LI B. Adaspeech：Adaptive text to speech for custom voice[C]// 9th International Conference on Learning Representations. Virtual Event：OpenReview. net，2021.
[13] 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.
[14] HE M，DENG Y，HE L. Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS [C]// 20th Annual Conference of the International Speech Communication Association. Graz：ISCA，2019：1293-1297.
[15] ZHENG Y，LI X，XIE F，et al. Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer [C]// ICASSP 2020-2020 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2020：6734-6738.
[16] ZHAO W，HE T，XU L. Enhancing local dependencies for Transformer-based text-to-speech via hybrid lightweight convolution[J]. IEEE Access，2021，9：42762-42770.
[17] LIU Y，XUE R，HE L，et al. DelightfulTTS 2：End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders[C]// 23rd Annual Conference of the International Speech Communication Association. Incheon：ISCA，2022：1581-1585.
[18] MORIOKA N，ZEN H，CHEN N，et al. Residual Adapters for Few-Shot Text-to-Speech Speaker Adaptation[J]. arXiv：2210. 15868，2022.
[19] LEI S，ZHOU Y，CHEN L，et al. Towards Expressive Speaking Style Modelling with Hierarchical Context Information for Mandarin Speech Synthesis[C]// 2022 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2022). IEEE，2022：7922-7926.
[20] WANG Y，STANTON D，ZHANG Y，et al. Style tokens：Unsupervised style modeling，control and transfer in end-to-end speech synthesis[C]// International Conference on Machine Learning. PMLR，2018：5180-5189.
[21] STANTON D，WANG Y，SKERRY-RYAN R J. Predicting expressive speaking style from text in end-to-end speech synthesis [C]// 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE，2018：595-602.
[22] CHOI S，HAN S，KIM D，et al. Attentron：Few-shot text-to-speech utilizing attention-based variable-length embedding[C]// 21st Annual Conference of the International Speech Communication Association. Shanghai：ISCA，2020：2007-2011.
[23] ELIAS I，ZEN H，SHEN J，et al. Parallel tacotron：Non-autoregressive and controllable tts[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP 2021). IEEE，2021：5709-5713.
[24] BAE J S，YANG J，BAK T J，et al. Hierarchical and Multi-Scale Variational Autoencoder for Diverse and Natural Non-Autoregressive Text-to-Speech[C]// 23rd Annual Conference of the International Speech Communication Association. Incheon：ISCA，2022：813-817.
[25] CHIEN C M，LEE H. Hierarchical prosody modeling for non-autoregressive speech synthesis[C]// 2021 IEEE Spoken Language Technology Workshop(SLT). IEEE，2021：446-453.
[26] RAMACHANDRAN P，ZOPH B，LE Q V，et al. Searching for activation functions[C]// 6th International Conference on Learning Representations. Vancouver：OpenReview. net，2018.
[27] DAUPHIN Y N，FAN A，AULI M，et al. Language modeling with gated convolutional networks[C]// International Conference on Machine Learning. PMLR，2017：933-941.
[28] ZEN H，DANG V，CLARK R，et al. LibriTTS：A corpus derived from LibriSpeech for text-to-speech[C]// 20th Annual Conference of the International Speech Communication Association. Graz：ISCA，2019：1526-1530.
[29] 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.

WU Kewei，born in 1984，Ph. D，associate researcher，is a member of CCF (No. 42032M). His main research interests include speech synthesis，computer vision and deep learning.

XIE Zhao，born in 1980，Ph.D，associate professor. His main research interests include computer vision，image analysis and understanding，and deep learning.

（责任编辑：何杨）