黄天圆, 王超. 基于遮蔽多头注意力的 CTC-Conformer 中文语音识别模型[J]. 智能计算机与应用,2025,15(2):162-167.
DOI:10.20169/j.issn.2095-2163.250225

基于遮蔽多头注意力的 CTC-Conformer 中文语音识别模型

黄天圆, 王 超
(河北工程大学 信息与电气工程学院, 河北 邯郸 056038)

摘 要: Conformer 模型是语言处理任务中广泛应用的模型之一,其结合了 Transformer 模型和卷积神经网络的特点,既能捕捉到局部和全局的序列特征又能更好地理解输入数据的结构和上下文信息。然而,现有 Conformer 模型中的音频和文本之间对齐关系存在不确定性,同时模型采用的多头注意力还会将未来时间步输入信息泄漏到当前时间步。采用连接时序分类(Connectionist Temporal Classification, CTC)机制进行辅助训练,不仅可以提高基于 Macaron-Net 结构的 Conformer 模型鲁棒性,还可以解决音频和文本不对齐问题。在解码器部分,应用遮蔽多头自注意力机制以确保在 t 时刻模型无法查看未来时间步的输入信息,从而保证模型仅利用已生成的标记进行预测。实验结果表明,基于遮蔽多头注意力的 CTC-Conformer 模型相对于 Conformer 模型的字错率与损失率均有所下降,损失值最低达到了 3.24。

关键词: Conformer; CTC; 遮蔽多头注意力; 语言处理
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2025)02-0162-06

Combining CTC with transformer model for implementing Chinese speech recognition

HUANG Tianyuan, WANG Chao
(School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, Hebei, China)

Abstract: Conformer is one of the most widely used models for language processing tasks. It combines the features of Transformer and convolutional neural network, it can not only capture local and global sequence features, but also better understand the structure and context information of input data. On the one hand, in the current Conformer model, it is uncertain in the alignment between audio and text. On the other hand, the multi-attention will leak the input information of the future time step to the current time step. To solve the above problems, the connectionist temporal classification (CTC) is used to improve the robustness of the Conformer model based on Macaron-Net structure, and resolve the issue of audio and text misalignment. Furthermore, masking multi-head self-attention mechanism is applied, in the decoder part, to ensure that the model can not view the input information of future time step at T-moment, so that the model can only make predictions with the generated markers. The results show that both the word error rate and the loss rate of CTC-Conformer model based on masking multi-head attention are lower than that of Conformer model, the lowest loss rate is 3.24.

Key words: conformer; CTC; mask multi-headed attention; speech recognition

0 引 言

随着科技的进步与智能家电的发展,机器可以越来越好地理解人类语言,语音识别技术也因此得到了广泛研究。传统的语音识别模型包括隐马尔可夫模型、高斯模型、基于神经网络的模型和混合模型等,其建模流程相对复杂且需要较强的专业知识[1]。随着语音识别技术的发展,基于深度学习的端到端语音识别模型(end-to-end)成为热点话题,此模型与传统模型的不同在于其将整个语音识别作为一个单一的模块进行训练推断,直接从原始语音信号中学习并输出最终的文本结果,省去了传统模型中多个复杂的步骤[2-3]。

目前,主流的端对端语音识别模型包括基于注

基金项目: 河北省自然科学基金面上项目(A2020402013)。
作者简介: 黄天圆(1997—),女,硕士研究生,主要研究方向:机器学习,智慧医疗。Email:1243225144@qq.com; 王 超(1983—),男,博士,教授,硕士生导师,主要研究方向:不确定信息处理。
收稿日期: 2023-08-11

意力机制的编码器-解码器模型与 CTC 模型[4]。编码器-解码器模型一般使用 Transformer 或 Conformer 结构，通过自注意力机制捕捉输入序列的全局信息，并利用解码器生成目标序列，这种模型在处理长序列和复杂上下文时表现出色[5]；CTC 模型是一种无监督的学习方法，在训练过程中不需要对齐的标注数据。基于注意力机制的编码器解码器、CTC 机制的优点，研究人员又提出了 CTC-Attention 模型，该模型首先通过卷积神经网络或循环神经网络从语音信号中提取特征序列，这些特征序列随后被送入带有 Attention 机制的解码器中，解码器会根据当前时刻的输入和前一时刻的状态生成一个表示当前输出的向量，并计算一个对齐权重向量；最后，将解码器的输出向量与对齐权重向量进行拼接、分类，获得最终的输出结果。CTC-Attention 模型的独特之处在于其融合了 CTC 机制和 Attention 机制。CTC 机制在语音识别任务中具有良好的序列建模能力，而 Attention 机制则能够有效地对输入序列进行对齐和加权处理。CTC-Attention 模型与 CTC 机制和 Attention 机制相比，CTC-Attention 模型可以在保证准确率的同时减少对齐信息的需求，并且提高模型的泛化性能，在语音识别等任务中取得了很好的效果[6-8]。

基于 Conformer 模型的端到端语音识别是当前较为先进的一种语音识别技术，Gulati[9] 等提出 Conformer 模型，相对于传统的 Transformer 模型表现出更加优异的性能。Conformer 模型结合了 CNNs（Convolutional Neural Networks）、SA（Self-Attention）和 FFNs（Feed-Forward Neural Networks）等模块，能有效地捕捉输入语音的时序特征和上下文信息。此外，Conformer 模型还引入了一种称为相对位置编码的新型位置编码机制，用于更好地构建长序列输入的时序关系。Conformer 模型将卷积层和注意力层结合，可以更好地处理输入数据中的时间信息，从而提高识别准确率。

传统的语音识别模型需要依靠隐马尔可夫模型（Gaussian Mixture Model-Hidden Markov Model，GMM-HMM）提供对齐标注，才能实现帧与状态的对齐[10]。而 CTC 模型使用无对齐标注数据的方式训练序列建模任务，巧妙的解决了对齐问题。单独使用 Conformer 模型也会出现音频和文本不对齐问题，这是因为音频中的语音单元（音素或子词）与文本序列之间存在差异或不完整匹配，可以利用 GMM-HMM 模型解决音频和文本不对齐的问题，但是却使模型更为复杂。而使用 CTC 进行辅助训练不仅可以有效地解决音频和文本不对齐的问题，提升语音识别系统的性能，还可以简化模型，提高训练性能。传统的基于 Macaron-Net 结构的 Conformer 模型在解码器部分使用多头注意力[8]。针对多头注意力会使当前时间步的训练受到未来时间步输入信息的影响这个问题，本文将多头注意力替换为遮蔽多头注意力，从而避免信息重复、信息误差和模型无效等问题。通过使用遮蔽多头注意力，可以确保 Conformer 模型只利用已生成的标记进行预测，避免了未来信息对当前时间步训练的影响，有助于提高模型的效果和稳定性。

1　相关理论

1.1　多头注意力

Treisman 和 Gelade[11] 提出注意力机制，通过为模型中的不同部分分配不同的权重，达到提取关键数据的目的，优化了建模过程，提升建模精度。Bahdanau 等[12] 首次将注意力机制应用至自然语音处理（NLP）中。由于注意力机制良好的效果，将注意力机制与神经网络的结合已经成为各个领域共同关注的热点问题[13-14]。注意力机制分为基本注意力机制和组合注意力机制。基本注意力机制是最简单的形式，其只关注当前时刻输入和上下文信息；而组合注意力机制，又称多头注意力机制（Multi-Head Attention），是在自注意力机制的基础上进行拓展，可以提取到不同维度的特征，充当“卷积核”的作用[15]。多头注意力的结构图如图 1 所示[16]。

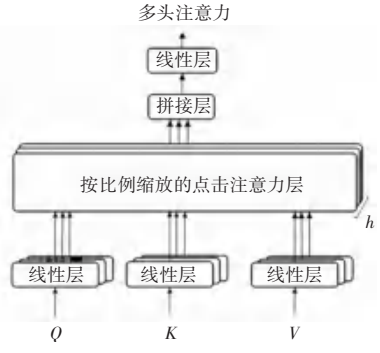


图 1　多头注意力结构

Fig. 1　Multiple head attention structure

多头注意力机制使用多个独立的注意力头并行地进行自注意力计算。每个注意力头具有自己的查询、键和值的线性投影权重矩阵，可以让模型从不同的子空间中学习不同的注意力表示，允许模型同时关注输入序列中不同位置的不同关系。传统的注意

力机制将输入序列编码作为上下文向量,捕获整体的语义信息,但忽略了不同位置之间的差异性和关系。多头注意力机制通过使用多个注意力头来解决这个问题,每个注意力头都可以学习到不同的权重分配,使得模型能够同时关注到输入序列中不同位置的不同关系。每个头都可以捕获不同的语义特征和上下文信息,并且每个头的输出被拼接在一起形成最终的表示。通过引入多头注意力机制,模型可以更好地理解输入序列的组织结构、位置关系和内部依赖关系,这样的并行计算也有助于加快训练和推理速度。

1.2 Conformer 模型

Conformer 模型是一种用于语言建模和语音识别任务的深度学习模型[17]。其由多个组件组成,包括两个前馈网络、一个子注意力模块和一个卷积模块,Conformer 模型结构如图 2 所示。在该结构中,多头注意力模块用于对输入序列进行自注意力机制的处理,以捕捉输入序列中不同位置之间的关联信息;可以将注意力集中在不同位置上,并生成具有不同权重的特征表示,可以更好地建模序列中的长距离依赖关系。

卷积模块用于在时间维度上对输入进行卷积操作,以提取局部特征,有助于捕捉输入序列的局部结构和模式。

两个前馈网络模块分别位于多头注意力模块和卷积模块的输入输出位置。前馈网络模块使用线性变换对输入进行映射和变换,以增加模型的表达能力;同时引入了 Swish 激活函数来引入非线性关系,并帮助模型更快地收敛。

为了促进信息传递和梯度流动,Conformer 模型中还使用了残差连接,允许原始输入信息直接流经模型的不同层级,从而减少了信息丢失和梯度消失的问题。

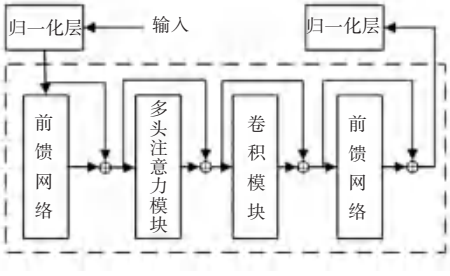


图 2 Conformer 模型结构

Fig. 2 Model of conformer structure

1.3 CTC 机制

CTC 机制的主要目标是解决时序数据建模中标签对齐的问题[18]。由于输入序列中的每个时间步都可以映射到多个输出标签或者没有映射到任何标签。而 CTC 机制引入一个特殊的“blank”标记来解决不对齐的问题,将输入序列中的每个时间步与标签序列中每个标签及“blank”标记进行对应。在训练过程中,CTC 机制使用前向-后向计算,计算每个时间步的标签路径概率,通过计算所有可能的标签路径的概率来进行训练和推断,从而解决传统方法无法处理的问题。

CTC 机制将输入序列中的每个时间步与标签序列中每个标签及“blank”标记进行对应。输入序列上引入“blank”标记以表示没有声学事件的区间;通过在输出序列中允许重复字符和使用特殊的“blank”标记符号来建立输入序列和输出序列之间的对应关系,并使用动态规划算法来解码,找到最有可能的输出序列。该算法考虑了所有可能的对齐方式,并计算每个对齐方式的概率。

2 基于遮蔽多头注意力的 CTC-Conformer 模型

基于遮蔽多头注意力的 CTC-Conformer 模型结合了 Conformer 模型和 CTC 机制优点的语音识别模型。使用 Conformer 模型并将 CTC 机制中的目标函数作为辅助任务,附加到共享编码器中,基于遮蔽多头注意力的 CTC-Conformer 模型可以直接将语音信号转化为对应的文本序列。本文设计的模型可以将测试者的语音信号转化成文本序列。

本文设计的基于遮蔽多头注意力的 CTC-Conformer 模型在桑江坤等[19]设计的基于 Macaron-Net 结构的 Conformer 模型基础上改进并融合 CTC 机制。Conformer 模型在 Transformer 模型的基础上融合深度可分离卷积,对音频序列的局部和全局依赖性进行建模,改善了 Transformer 模型提取局部特征能力较差的问题。CTC-Conformer 模型使用 Conformer 模型作为编码器、并利用 CTC 机制共同构造的语音识别模型。Mask-CTC-Conformer 模型结构如图 3 所示,包括前置处理模块(声学前置模块和文本前置模块)、编码器解码器模块(Encoder-Decoder 模块)、CTC-Conformer 损失计算模块(CTC-Conformer Loss 模块)。

2.1 前置处理模块

前置处理模块分为两部分:声学前置模块和文本前置模块如图 4 所示。声学前置模块是语音识别系统中的一个重要组成部分,负责将原始语音信号

换为适合于处理的声学特征向量；文本前置模块负责将文本标签转换为特征表示。前置处理模块将语音信号和对应的本标签转换为适合于后续处理的特征表示，以实现准确的语音识别和语义理解，通过声学前置模块和文本前置模块的协同工作，语音识别系统能够更好地理解处理语音输入。

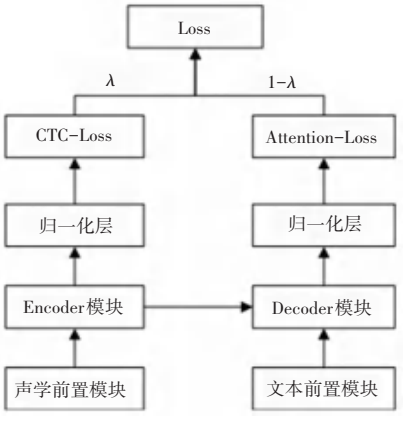
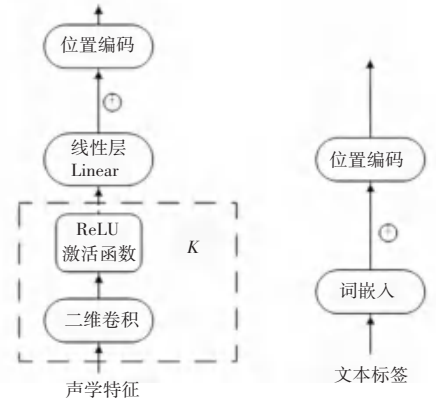


图 3　Mask-CTC-Conformer 模型结构

Fig. 3　Model of Mask-CTC-Conformer



(a) 声学前置模块　　(b) 文本前置模块

图 4　前置处理模块

Fig. 4　Pre-processing module

2.2　编码器-解码器模块

编码器-解码器模块中编码器(Encoder)部分的多头注意力模块使用相对位置嵌入，将相对位置信息编码转化为正弦函数或余弦函数的形式，并与输入序列的嵌入向量相加，从而为模型提供关于输入元素之间相对位置的表示，能够捕捉到序列中不同位置之间的相对距离关系，使模型在处理长序列时具有更好的泛化性以及更强的鲁棒性。解码器(Decoder)部分使用遮蔽多头自注意力，因为注意力计算公式 qt 无法避免的自动看到全局的信息，遮蔽多头注意力可以确保解码器能依赖于已经生成的输出，而不会受到未来时刻的信息影响。这种遮蔽机制的主要作用是避免在 t 时刻看到之后的输入，即只能观测到 $k_1, k_2, \cdots, k_{t-1}$，使用遮蔽多头注意力机制有助于保持解码器的自回归性质，并确保生成的序列是符合语法和语义规则的。为使 t 时刻及之后参与计算的值在归一化时权重为 0，故而本文中将 t 时刻及之后参与计算的值换为 $-1e^{10}$；每个子层之间使用残差连接防止梯度消失，加快模型收敛。Encoder-Decoder 模块设计如图 5 所示。

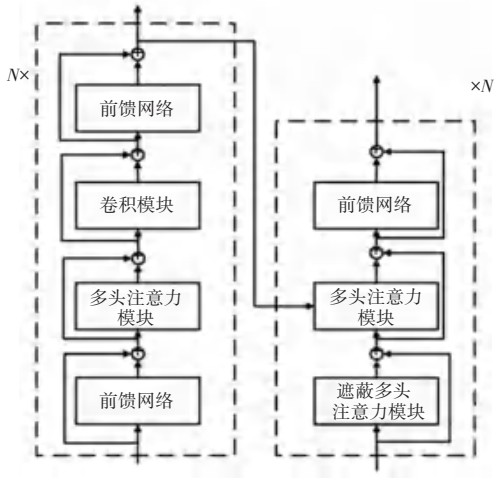


图 5　Encoder-Decoder 模块

Fig. 5　Encoder-Decoder module

2.3　损失计算模块

CTC-Conformer 模型训练过程属于多任务学习，其中涉及到两个主要的损失函数：Conformer 损失函数和 CTC 损失函数。综合考虑这两个损失函数，使用加权和的方式来计算总体损失，公式如下：

$$T_{loss} = \lambda \mathrm{CTC}_{loss} + (1-\lambda)\,\mathrm{Att}_{loss} \tag{1}$$

其中，$\lambda \in [0,1]$，为一个超参数，用于衡量 CTC 机制与 Attention 机制的权重。

通过综合考虑两个损失函数，可以提高模型在不同任务上的性能和泛化能力。

3　实验与分析

3.1　实验数据

本文实验数据来源于希尔贝壳中文普通话开源数据库 AISHELL-2。其是在 AISHELL-1 的基础上进行扩展和改进得到的，旨在提供更多样化和高质量的中文普通话语音数据。AISHELL-2 数据库不仅包含了广泛的会话场景中录制的音频、多种录音设备录制的音频、而且还收集了不同年龄段和性别的人的音频，用于语音识别、语音合成等领域的研究和开发。

3.2　实验环境

在本文实验使用 Intel(R) Core(TM) i7-

10700K CPU @ 3.80 GHz 3.79 GHz 的服务器,内存 16 G,GPU 显卡型号为 NVIDIA GeForce RTX 3080,深度学习框架为 Pytorch1.8-gpu。

3.3 实验参数

本文实验中卷积核大小为 3×3,步长为 2。Encoder 与 Decoder 部分中多头注意力输出维度为 256,注意力头数为 4,前馈层输出维度为 1 024,使用 glue 激活函数。损失计算模块中权重 λ 值取 0.3。模型训练过程中 $echo = 600$,使用 $\beta_1 = 0.9$,$\beta_2 = 0.98$,$\varepsilon = 10^{-9}$ 的 Adam 优化器。在训练过程中学习率是一个重要的超参数,控制着模型权重更新的步长,为了达到优化模型的性能和收敛速度的目的,在实验过程中动态调整学习率。Dong L 等[20]采用动态调整策略调整学习率,本文在实验过程中使用与其一致的策略,动态调整学习率。

3.4 结果分析

训练过程:前置模块将输入的语音特征及其对应的文本标签转化为特征向量,Encoder-Decoder 模块中 Encoder 将声学特征向量映射到隐藏层向量,此隐藏向量作为 CTC-Decoder 的输入,分别计算 CTC 与 Conformer 的损失值,并通过加权的方式计算总损失值。模型中字错率计算公式如下:

$$CER = (S + D + I)/N \tag{2}$$

其中,S 代表参考例句转化为预测样本时替换单词的数量;D 代表预测样本转化为参考例句时删除的单词数量;I 代表预测样本转化为参考例句时额外插入的单词数量;N 代表整句的总字数。

实验总共训练 600 轮,损失值变化趋势如图 6 所示。根据图 6 可以看出在训练 30 轮后损失值趋于平缓,最低达到了 3.24;在 300 轮之后有上升趋势,这是由于训练轮数过多导致了过拟合问题,进而导致损失值上升。

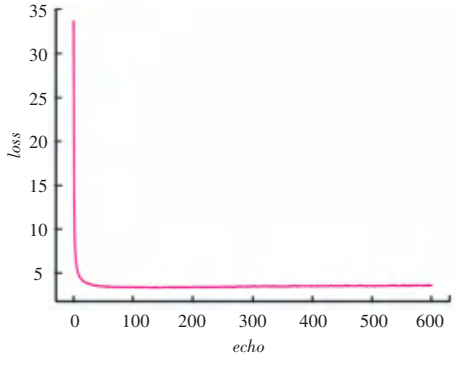


图 6 损失值

Fig. 6 Loss values

本文使用字错率来判断模型识别的准确程度,随着训练轮数的增加,Transformer 模型、Conformer 模型与 CTC-Conformer 模型的字错率对比如图 7 所示,可以看出 Conformer 模型的准确性优于 Transformer 模型,而基于遮蔽多头注意力的 CTC-Conformer 模型的准确率优于 Conformer 模型,基于遮蔽多头注意力的 CTC-Conformer 模型字错率比 Conformer 模型的字错率低,达到了 5.0%的错误率。

基于遮蔽多头注意力的 CTC-Conformer 模型不仅在字错率方面有了明显的下降,该模型还具有更强的鲁棒性和稳定性。因此,该模型可以广泛应用于实际场景的语音识别任务中,具有较高的实用价值。

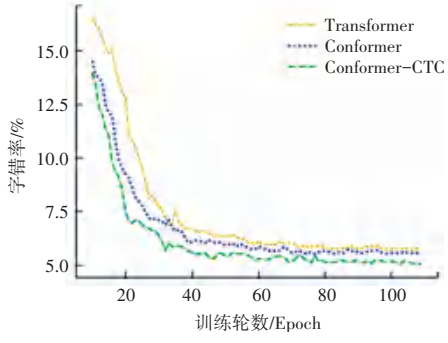


图 7 验证集字错率对比

Fig. 7 Verification set word error rate

4 结束语

本文提出的基于遮蔽多头注意力的 CTC-Conformer 中文语音识别模型,将解码器部分的多头注意力替换为遮蔽多头自注意力,并使用 CTC 机制进行辅助训练。使用数据库 AISHELL-2 验证模型在中文语音识别中的效果,该模型表现出较好的性能,最终达到了字错率 5.0%的效能。未来的研究中还会探索不同的参数对模型的影响,还会将此模型应用到医学领域实现自动化诊断疾病。

参考文献

[1] 谢旭康. 基于端到端的语音识别模型研究及系统构建[D]. 苏州:江南大学,2022.

[2] 方明弘, 万里, 戴凡杰. 基于双层记忆网络的多领域端到端任务型对话系统[J]. 计算机应用研究, 2023,1(7): 7-20.

[3] 邵娜,李晓坤,刘磊等. 基于深度学习的语音识别方法研究[J]. 智能计算机与应用,2019,9(2):135-142.

[4] HADWAN M, ALSAYADI H A, AL-HAGREE S. An end-to-end transformer-based automatic speech recognition for Qur´an reciters[J]. Computers, Materials & Continua, 2023, 74(2): 3471-3487.

[5] 沈逸文, 孙俊. 结合 Transformer 的轻量化中文语音识别[J]. 计算机应用研究,2023,40(2):424-429.

[6] MIAO H, CHENG G, ZHANG P, et al. Online hybrid CTC/

attention end-to-end automatic speech recognition architecture [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020(28): 1452-1465.
[7] PARK H, KIM C, SON H, et al. Hybrid CTC-attention network-based end-to-end speech recognition system for Korean language [J]. Journal of Web Engineering, 2022, 21(2): 265-284.
[8] 谢旭康，陈戈，孙俊，等. TCN-Transformer-CTC 的端到端语音识别[J]. 计算机应用研究，2022，39(3)：699-703.
[9] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv，2005.08100, 2020.
[10] 李云红，梁思程，贾凯莉，等. 一种改进的 DNN-HMM 的语音识别方法[J]. 应用声学，2019，38(3)：371-377.
[11] YU C, YU J, QIAN Z, et al. Endangered Tujia language speech recognition research based on Audio-Visual Fusion [C]// Proceedings of the 5th Artificial Intelligence and Cloud Computing Conference. 2022: 190-195.
[12] BAHDANAU D. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv，1409.0473, 2014.
[13] GUO M H，XU T X，LIU J J. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.
[14] TANVEER M, GANAIE M A, BEHESHTI I, et al. Deep learning for brain age estimation: A systematic review [J]. Information Fusion, 2023(96): 130-143.
[15] REN Z, YOLWAS N, SLAMU W, et al. Improving hybrid CTC/attention architecture for agglutinative language speech recognition[J]. Sensors, 2022, 22(19): 7319.
[16] HAO M, XU B, LIANG J Y, et al. Chinese short text classification with mutual-attention convolutional neural networks [J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2020, 19(5): 1-13.
[17] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv，2005.08100, 2020.
[18] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]// Proceedings of International Conference on Machine Learning. PMLR, 2021: 5530-5540.
[19] 桑江坤，努尔麦麦提·尤鲁瓦斯. 基于 Conformer 的端到端语音识别模型的压缩优化策略[J]. 信号处理，2022，38(12)：2639-2649.
[20] DONG L, XU S, XU B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition [C]// Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signalprocessing. Piscataway, NJ: IEEE, 2018: 5884-5888.