基于改进Conformer的新闻领域端到端语音识别

张济民[1,2]，早克热·卡德尔[1,2,*]，申云飞[2,3]，艾山·吾买尔[1,2,*]，汪烈军[1,2]
1.新疆大学信息科学与工程学院，新疆　乌鲁木齐
2.新疆大学新疆多语种信息技术实验室，新疆　乌鲁木齐
3.新疆大学软件学院，新疆　乌鲁木齐
{e_zhangjimin,shenyunfei}@stu.xju.edu.cn,{zuhra,hasan1479,wljxju}@xju.edu.cn

摘要

目前，开源的中文语音识别数据集多为面向通用领域，缺少面向新闻领域的开源语音识别语料库，因此本文构建了面向新闻领域的中文语音识别数据集CH_NEWS_ASR并使用ESPNET-0.9.6框架的RNN、Transformer和Conformer等模型对数据集的有效性进行了验证，实验表明本文所构建的语料在最好的模型上CER为4.8%，SER为39.4%。由于新闻联播主持人说话语速相对较快，本文构建的数据集文本平均长度为28个字符是Aishell_1数据集文本平均长度的2倍，且以往的研究中训练目标函数通常为基于字或词水平，缺乏明确的句子水平关系，因此本文提出了一个句子层级的一致性模块与Conformer模型结合直接减少源语音和目标文本的表示差异，在开源的Aishell_1数据集上其CER降低0.4%，SER降低2%；在CH_NEWS_ASR数据集上其CER降低0.9%，SER降低3%，实验结果表明该方法不提升模型参数量的前提下能有效提升语音识别的质量。

关键词： 端到端语音识别；Conformer；句子层级一致性

End-to-End Speech Recognition in News Field based on Conformer

Jimin Zhang[1,2],Zaokere Kadeer[1,2,*],Yunfei Shen$_{2,3}$,Aishan Wumaier[1,2,*],Liejun Wang[1,2]
1.School of Information Science and Engineering, Xinjiang University
Urumqi Xinjiang
2.Key Laboratory of Multilingual Information Technology in Xinjiang Uyghur Autonomous Region,Urumqi Xinjiang
3.School of Software, Xinjiang University Urumqi Xinjiang
{e_zhangjimin,shenyunfei}@stu.xju.edu.cn,{zuhra,hasan1479,wljxju}@xju.edu.cn

Abstract

At present, the open source Chinese speech recognition data sets are mostly for the general domain, and there is a lack of open source speech recognition corpus for the news domain. Therefore, this paper constructs a news-oriented Chinese speech recognition data set CH_NEWS_ASR and uses the RNN, Transformer and Conformer models of ESPNET-0.9.6 framework to verify the validity of the data set. Experiments show that the CER and SER of the corpus constructed in this paper are 4.8% and 39.4% on the best model. As news broadcast hosts speak relatively fast, the average text length of the dataset constructed in this paper is 28 characters, which is 2 times of the average text length of Aishell_1 dataset. In addition, in previous studies, the training objective function is usually based on word or word level, and there is no clear sentence level relationship. In this paper, we propose a sentence-level consistency module combined with the Conformer model to directly reduce the representation differences between source speech and target text. On the open source Aishell_1 dataset, the CER decreases

第二十届中国计算语言学大会论文集，第851页-第861页，呼和浩特，中国，2021年8月13日至15日。

by 0.4% and the SER decreases by 2%. On the CH_NEWS_ASR dataset, the CER decreases by 0.9% and the Ser decreases by 3%. The experimental results show that the proposed method can effectively improve the quality of speech recognition without increasing the number of model parameters.

Keywords: End-to-End Speech Recognition , Conformer , Sentence-Level Agreement

1 引言

语音识别是利用智能算法将人类的语音转换为文本或控制信号的过程，在许多生物识别系统和语音控制自动化系统中起着至关重要的作用。与将语音识别任务分解为多个子任务（词汇模型、声学模型和语言模型）的传统方法不同(王勇和et al., 2018)，端到端的语音识别模型能够根据输入的音频特征生成与之对应的文本信息，在一定程度上简化了语音识别任务中对模型的训练过程(张宇et al., 2018)，使得输入的序列长度远大于输出序列长度的问题得到了解决。近年来，人们研究了各种模型，包括(Graves and Jaitly, 2014)使用基于链接时序主义(Connectionist temporal classification,CTC)、(Miao et al., 2015; Chan et al., 2016)使用基于注意力机制的编码器-解码器模型以及(Chan et al., 2016; Zeyer et al., 2018)混合模型和(Graves et al., 2013)RNN-T模型。本文主要使用基于链接时序主义和基于注意力机制的编码器-解码器模型。

当前中文语音识别开源的数据集多为面向通用领域的，针对新闻领域的语音识别研究较少，资源也相对溃泛。但是在我国新闻联播内容丰富不仅可以从中发现商机、了解国家政策红利、预知风险，还可以做舆情、政情分析把握时代发展的大方向。能够看懂、理解新闻联播具有重大的意义，因此本文构建了面向新闻领域的中文语音识别数据集CH_NEWS_ASR。

数据集	总时长/小时	主要领域
Speechocean	10	日常对话
THCHS30	30	诗句、书籍等
Prime words	100	语音聊天和智能语音控制
ST-CMDS	122	历史、书籍等
Aishell_1	178	科技、体育、娱乐等
Aidatatang_200zh	200	普通电话语料库
MAGICDATA	755	互动问答、音乐搜索等
Aishell_2	1000	智能家居、无人驾驶等
Aidatatang_1505zh	1505	普通电话语料库

Table 1: 主要开源的中文语音识别数据集

语音识别的训练目标是尽量减少识别音频中的字或词与对应文本中字或词之间的损失，且在端到端语音识别中，音频和文本之间存在着一种对应关系，然而这种关系仅利用整个神经网络表示出来，并且训练目标是按字或词的水平计算。因此本文提出了一个句子层级一致性方法来直接减少音频和对应文本的表示差异，在开源的(Bu et al., 2017)Aishell_1数据集上其CER降低0.4%，SER降低2%，在CH_NEWS_ASR数据集上其CER降低0.9%，SER降低3%，表明该方法能有效提升语音识别质量。

本文结构组织如下：第二章对研究方法进行简要介绍；第三章介绍本文构建的面向新闻领域的语音识别数据集，包括标注方法、数据统计和基线系统等；第四章介绍本文提出方法、实验数据、实验内容、实验结果及分析，验证句子一致性方法的有效性；第五章对本文进行总结。

基金项目：国家重点研发计划（2017YFB1002103）、自治区“天山创新团队计划”申报书(编号：202101642)

2 研究方法

2.1 Conformer模型

(Gulati et al., 2020)提出的Conformer模型是(Vaswani et al., 2017)Transformer模型的一种变种，其在编码端利用Transformer模型擅长捕捉基于内容的全局交互，卷积神经网络能够有效地利用局部特征的特性，通过参数有效的方式将二者融合更好地提取音频序列的局部和全局依赖性。其整体结果如图Figure 1所示，主要由位置嵌入、多头自注意力层、掩码多头自注意层、卷积模块、残差连接、层归一化等组成。

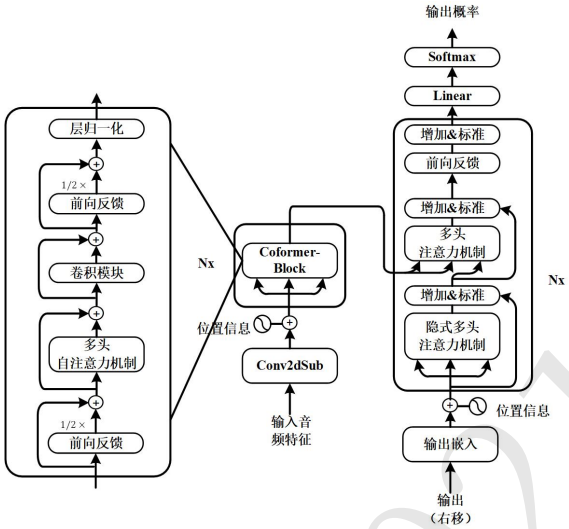

Figure 1: Conformer模型的整体结构

Conformer模型主要分为两大部分，即编码器和解码器。编码器是通过堆叠其编码器子层而形成的，在编码器子层中，首先将输入的数据传入多头注意力层，在这里，自注意力（即Q=K=V）来对输入数据所具有的上下文关系进行计算。当从多头注意力层输出后将会进行层归一化等操作，然后输入数据将通过残差连接与多头的输出进行加和，再输入卷积模块，将其输出进行残差连接后输入到前向反馈层。另外，多头注意力和前向反馈层的输出也将被层归一化。

解码器由一堆解码器子层组成，包含能够计算输出文本的上下文信息的自注意力层、相互注意力层、前向反馈层和对应的残差连接。另外，因为解码器对自注意力进行计算的时候只能看到当前时刻及其之前时刻的信息，因此使用Mask机制来确保训练和解码的一致性，计算公式如下：

$$\mathrm{mas}k(e_{ij}) = \begin{cases} e_{i,j}, j \leq i \\ -\infty, otherwise \end{cases} \tag{1}$$

编码器主要负责将输入（语言序列）映射到隐藏层，然后解码器将隐藏层映射成自然语音的序列。Conformer使用位置嵌入来加入语言的顺序信息，使用自注意力机制、卷积模块和全连接层进行计算，这将在下面进行详细描述。

位置编码,由于Consformer模型没有具有循环神经网络的迭代操作，因此必须将每个字或词的位置信息提供给Consformer，以便它可以识别语言中的序列关系。通过使用不同频率的正弦和余弦函数生成位置编码，然后将其添加到相应位置的字或词向量中，位置向量的维数必须与单词数量的维数相同。计算公式如下：

$$PE(pos, 2i) = sin(pos/10000^{(2i/d\_model)}) \tag{2}$$

$$PE(pos, 2i{+}1) = \cos(pos/10000^{(2i/d\_model)}) \tag{3}$$

式中，pos指的是序列中某个字的位置，取值范围是[0,max_sequence_length]，i指的是字向量的维度序号，取值范围是[0,embedding_dimension/2]，d_model指的是embedding_dimension的值。

在Conformer中使用多头的点乘注意力（Multi-head Scaled Dot Product Attention）,其结构如下图所示:

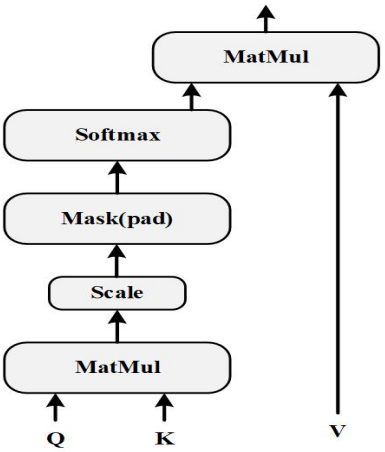

Figure 2: 缩放点乘注意力数据流

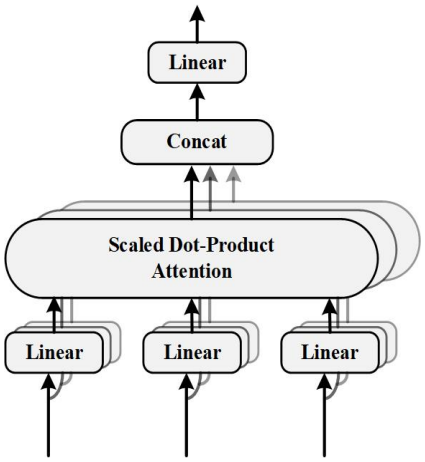

Figure 3: 添加多头机制后的数据流

在上图中，Q和K的运算有一个可选的Mask过程。在编码器中，我们不需要使用它限制注意力模块所关注的序列信息；而在解码器中，我们需要它限制注意力模块只能关注到当前时间步及以前时间步的信息。这个过程可以简洁地表示为函数Attention(Q,K,V)：

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

根据式(4)可以看出，首先要计算Q与K^T之间的点乘，为了防止其结果过大，接着除以一个尺度标准$\sqrt{d_k}$，其中d_k为一个Q或K向量的维度。最后使用Softmax这一操作来把结果转换为在[0,1]区间内的概率分布，然后再与矩阵V相乘获得权重求和的表示。

多头注意力则是通过不同的线性变换对Q，K，V进行投影，然后将经过缩放点乘注意力的结果拼接起来，如式(5)、式(6)所示：

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^o \tag{5}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$

卷积模块如Figure 4所示其包含逐点卷积、线性门单元、一维的深度卷积和批次归一化以及Swish激活层，通过卷积模块来更好地学习音频的局部特征。

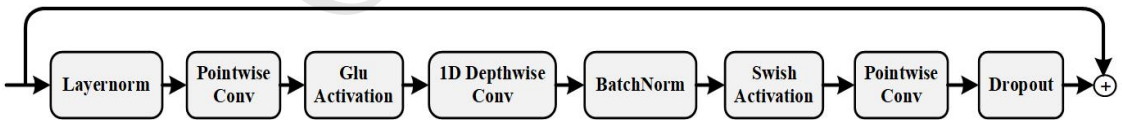

Figure 4: 卷积模块结构图

残差连接是输入和输入的非线性变化的叠加，通过将上一层的信息无差异地传递到下一层，仅仅关注差异部分可以减轻深度网络层数的增加带来的一系列问题，例如：梯度的消失、梯度的爆炸、模型的过度拟合以及计算资源的消耗等。层归一化的功能是将神经网络中的隐藏层进行归一化转换为标准正态分布，从而加快训练的速度，并起到加速收敛的作用。

2.2 评测标准

模型训练完成后，需要合适的评测来评估和分析其可靠性。为了能够充分地反应模型的性能，本文使用在中文语音识别中最常用的字错误率(CER)、句子错误率(SER)作为评测标准。如式(7)、式(8)所示：

$$CER = \frac{S + D + I}{N} \times 100\% \tag{7}$$

其中S表示被替换的错误数，D表示被删除掉的错误数，I表示位置插入的错误数，N表示在句子中所有的字数目。需要注意的是，由于编辑距离的特殊性，错误率存在超过100%的可能性。

$$SER = \frac{S.Error}{S.Total} \times 100\% \tag{8}$$

其中S.Error为句子中至少有一个词或字错误的句子个数，S.Total为总的句子个数。

$$W.Corr = \frac{N - D - S}{N} \times 100\% \tag{9}$$

其中N表示在句子中所有的字数目，D表示被删除的错误数，S表示被替换的错误数。

3 面向新闻领域语音识别数据集的构建

3.1 数据集的构建

本章构建的数据主要来源于央视新闻联播、各省级新闻联播等官方新闻联播网站，主要目标是构建一个面向新闻领域的中文语音识别数据集。语料的构建如Figure 5所示主要分为以下步骤:

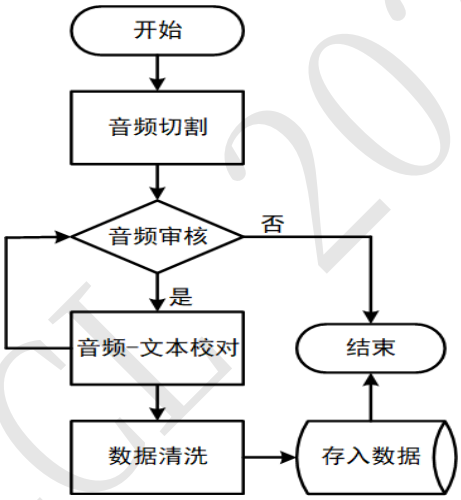

Figure 5: 语料构建流程图

1）音频切分及文本识别，使用(Giannakopoulos, 2015)PyAudioAnalysis工具对音频文件进行切分，将切分得到的音频分别使用百度和腾讯的接口对音频进行识别，然后使用Mteval工具计算二者的BLEU值对文本进行筛选;

2）音频审核，判断筛选后的音频是否完整;

3）音频-文本校对，将音频内容与文本进行校对，保证文本与音频对应。

最后数据的存储格式根据希尔贝壳提供的Aishell_1数据集格式相同，下面对音频切分、文本识别、音频审核和音频-文本校对进行详细介绍

3.1.1 音频切分及文本识别

由于下载的音频为时长30分钟左右、混合说话人的连续性音频，而准备构建的新闻联播语料库每条数据需为单一说话人、时长较短的音频文件，因此本文选用PyAudioAnalysis 工具包对下载的音频进行切分。

PyAudioAnalysis是一个非常好用且强大的python音频分析开源工具，能实现音频的特征提取、分类和回归模型的训练和执行，以及其他一些实用的功能，本文主要使用了其分割的功

能。音频分割是音频分析过程中非常重要的处理环节，其目的是把连续的音频信号分割成均匀的或者“同类”的音频片段。其操作过程如下：首先根据静音检测将音频拆分成相应时长的片段，并抽取这些片段的特征；然后使用预先训练好的模型对每个片段依据说话人进行分类；最后将相同说话人的片段放到同一个文件夹内。

根据数据整理过程的经验，当参数调好后主要删除的部分为采访、开头和结尾的内容，其他地方基本可以准确切分。将切分得到的音频分别使用百度和腾讯的接口对音频进行识别，然后使用(Bharati et al., 2004)Mteval工具计算二者的BLEU值，选取BLEU值大于0.5的进行人工音频审核。

3.1.2 音频审核

使用工具进行音频切分时，无法保证其能完全正确，因此需要人工对音频是否完整、是否只有一个说话人、说话环境是否干净等进行审核，从而保证音频的质量。本文使用的审核方法如下，其中✓表示可用，×表示不可用：

1）判断是否只有一个说话人

音频内容	是否可用
（李梓萌）以下来看详细报道	✓
（李梓萌）以下来看详细报道-（郭志坚）国务院发布最新消息	×

Table 2: 音频审核判断标准1

评判标准为当单个音频中有两个或多个说话人时此音频不能用，由Table 2可以看出，第一句话只包含李梓萌一个说话人因此可用，第二句话包含李梓萌和郭志坚两个说话人不符合要求故不能使用。

2）判断音频是否完整

音频内容	是否可用
今天出版的人民日报发表评论员文章	✓
今天出版的人民日报发表评论员文zha	×

Table 3: 音频审核判断标准2

评判标准为当单个音频内容表达不完整时不能使用，由Table 3可以看出，第一句话内容完整表达可用，第二句话最后“章”字未完全发音因此不能使用。

3）说话环境是否干净

音频内容	是否可用
本台最新消息（背景音干净）	✓
现在位于叙利亚的首都（背景音有枪炮声）	×

Table 4: 音频审核判断标准3

判断标准为当单个音频中有大风、枪炮等背景音时不能用，由Table 4可以看出，第一句话背景音是干净的故可用，第二句话的背景音有枪炮声因此不能用。

单个音频只有同时满足上面三个判断标准时才属于合格音频，计入语料库中作为能使用的数据，未满足标准则舍弃。

3.1.3 音频-文本校对

由于无法保证音频文件切分后的单个音频的文件内容，并且新闻联播单人说话字数较多且无法确保音频的截取位置完全正确，所以在语料规模不大的情况下尽可能保证语料的质量，需

人工根据音频实际内容对筛选得到的识别文本进行调整，删除涉及敏感政治问题、用户隐私、色情、暴力等不适当内容(牛米佳et al., 2020)，并将数字、日期、百分比等转换为中文读取方式，将诸如《，》，[,],\,/,=等符号删除，最后还需根据3.1.2的音频审核标准对其再进一步审核，确保语料的质量。

3.2 数据集统计

本章最终构建了72034条新闻领域的中文语音识别数据CH_NEWS_ASR，共计时长127小时。训练集共100小时包含56891条句子，开发集共15小时包含8312条句子，测试集共12小时包含6831条句子，训练集和测试集均为随机抽取。数据集的分布情况如Figure 6所示：

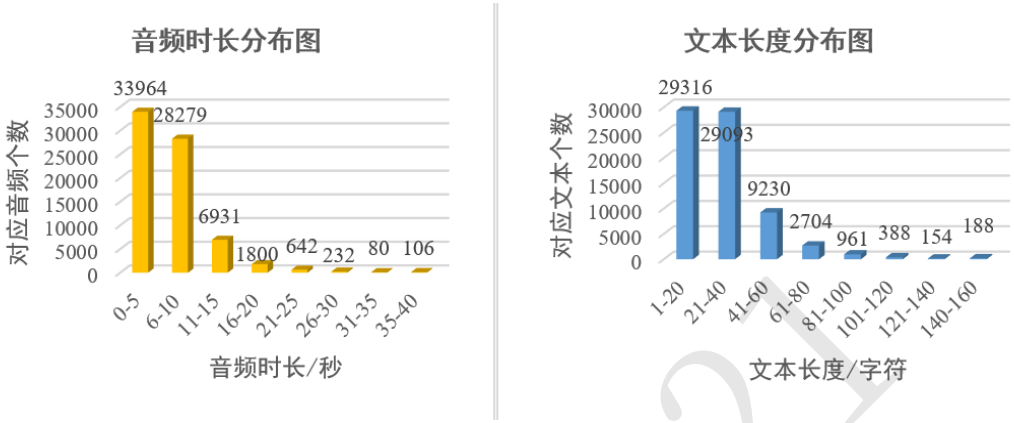

Figure 6: CH_NEWS_ASR数据集分布

数据集的音频平均时长约为6秒，平均文本长度约为28个字符，大多数文本长度在10-80个字符，只有少量文本的长度在100以上，句子长度数量和时长数量分布集中。在构建数据时我们发现，新闻联播主持人说话语速相对人们日常口语交流语速较快，单个说话人存在一句话音频较长字符较多现象。因此为了减少模型的训练时长和参数量，在保证语料完整性的基础上对文本字符长度大于160的、音频时长超过40秒的内容进行舍弃。

3.3 实验结果及分析

使用基于PyTorch的(Watanabe et al., 2018)ESPNET-0.9.6框架，使用端到端语音识别模型进行相关实验，通过实验证明该语料库是进行面向新闻领域语音识别的可靠数据库，并提供关于此数据库的语音识别基线。

3.3.1 实验环境

所有的实验均在Centos7.8.2003操作系统上进行，单块NVIDIA Tesla V100(16GB)，Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz处理器。提取80维的Fbank音频特征，使用Spec Augment方法对语音识别数据进行增强。将字错误率(CER)、句错误率(SER)作为模型识别准确与否的评价标准，计算方式与1.2节相同。本节将使用Conformer、Transformer和RNN等模型进行对比实验，验证数据集的有效性。其模型配置参数如下：

1）Conformer模型：编码器为12层其使用4头注意力，输入注意力机制的维度为256，dropout 0.1，使用swish激活函数；解码器为6层的4头注意力，dropout 0.1，其他参数为默认值，训练50个epoch。

2）Transformer模型：编码器为12层并使用4头注意力，输入维度为256，dropout 0.1；解码器为6层的4头注意力，dropout 0.1，其他参数为默认值，训练50个epoch。

3）RNN模型：编码器为VGG+BiLSTM，其中BiLSTM为3层，隐藏层节点为1024，dropout 0.0；解码器为2层的LSTM，隐藏层节点数为1024，Batch Size为30，其他参数使用默认值，训练10个epoch。

3.3.2 实验结果及分析

基于链接时序主义模型(CTC)借鉴了马尔科夫假设有效的解决了序列的动态对齐问题，基于注意力机制的模型通过一个注意力机制解决了声学帧和标签之间的对齐问题，证明了两种模型联合训练和联合解码优于上述单独一种方式训练的效果。因此，本文使用联合模型对构建的数据集进行训练，CTC权重的选择如Table 5所示，在Conformer模型中权重为0.5时性能最好，在Transformer模型中权重为0.3时性能最好，在RNN模型中权重为0.6时性能最好。因此，在后续的训练构成中CTC权重选择对应模型其性能最好的。

	RNN			**Transformer**			**Conformer**		
CTC权重	**W.corr**	**SER**	**CER**	**W.corr**	**SER**	**CER**	**W.corr**	**SER**	**CER**
0.1	91.9	60.4	8.4	93.4	46.5	7.4	94.7	42.5	5.9
0.2	92.3	58.6	8.2	93.6	46.1	7.4	92.5	45.8	8.1
0.3	92.5	55.3	8.0	**94.2**	**44.5**	**6.5**	95.2	41.3	5.5
0.4	92.6	56.3	7.9	92.4	46.6	9.3	95.2	41.5	5.5
0.5	92.8	54.2	7.6	92.2	48.1	9.4	**95.5**	**41.3**	**5.1**
0.6	**93.1**	**53.4**	**7.3**	92.3	47.8	9.4	94.8	41.9	5.6
0.7	92.7	53.8	7.8	92.1	48.5	9.5	94.6	42.4	5.9
0.8	92.6	56.4	7.9	92.0	48.8	9.5	94.2	43.5	6.2
0.9	93.8	58.4	8.1	92.1	48.6	9.5	92.2	44.8	6.5

Table 5: 基于不同模型的CTC权重选择实验结果

由Table 6可以看出，当使用融合CTC和音速扰动时该数据集的识别效果有较好的提升，且Conformer+SP+CTC模型的性能最好，与理论相符合。因为融合CTC能够更好地利用CTC损失函数来辅助模型更好地学习文本到语音的对齐，使用音速扰动(SP)即使用0.9和1.1的因子对训练集的语速进行扰动，使训练集的容量增大了3倍，同时对训练数据的容量进行了随机扰动，该技术有助于使神经网络模型对测试数据的速度和音量不变性更加鲁棒，因此在使用CTC和音速扰动(SP)时性能较好。

模型	**W.corr(%)**	**SER(%)**	**CER(%)**
RNN+CTC	93.1	53.4	7.3
RNN+SP+CTC	94.0	52.2	6.6
Transformer+CTC	94.2	44.5	6.5
Transformer+SP+CTC	95.1	43.8	5.6
Conformer	90.5	48.5	9.9
Conformer+SP	94	39.9	6.4
Conformer+CTC	95.5	41.3	5.1
Conformer+SP+CTC	**95.8**	**39.4**	**4.8**

Table 6: CH_NEWS_ASR数据集在不同模型下识别结果

通过实验本文构建的数据集在Conformer+SP+CTC模型中其测试集的字错误率和字正确率分别为4.8%和95.8%，实验结果证明了本文构建的数据集的有效性。但对于端到端语音识别所需数据量而言，本文构建的数据集规模较小，包含的省份和日期还不够全面，下一步将构建面向全国各个省份的近十年的新闻联播数据，实现新闻专门领域的语音识别。

4 融合句子层级一致性的Conformer模型

4.1 模型的改进

与传统的级联语音识别模型不同，端到端语音识别采用了一种注意机制来帮助输出文本和输入音频对齐，它基于对每个目标文本的所有输入音频的概率分布的估计。然而，音频和文本

处于不同的表示空间中，它们仍然需要经历一个漫长的信息处理过程，可能导致源音频被错误地识别成目标文本。

在以往的研究中，训练目标函数通常基于字或词水平，缺乏明确的句子水平关系。尽管Conformer模型已经在端到端语音识别中已经是最好的模型，但是更多的是通过自注意力网络关注字或词层级关系。句子层级一致性方法已经应用于许多自然语言处理任务中，(Aliguliyev, 2009)利用句子相似度测量技术进行摘要的自动生成；(Liang et al., 2010)基于VSM的句子相似度算法有助于解决FAQ问题；(Su et al., 2016)为了提高句子相似度，提出了一种面向口语对话系统的句子相似度方法；(Rei and Cummins, 2016)提出了提高话题相关度的句子相似度度量方法；(Wang et al., 2018)使用句子相似度选择具有相似领域的句子。在这些研究的启发下，本文直接在深度神经网络中建立了一个句子层级的一致通道，用于缩短源语音和目标文本句子级嵌入之间的距离，不仅考虑了字或词层面的识别，还考虑了句子层面的识别。其结构如Figure 7所示:

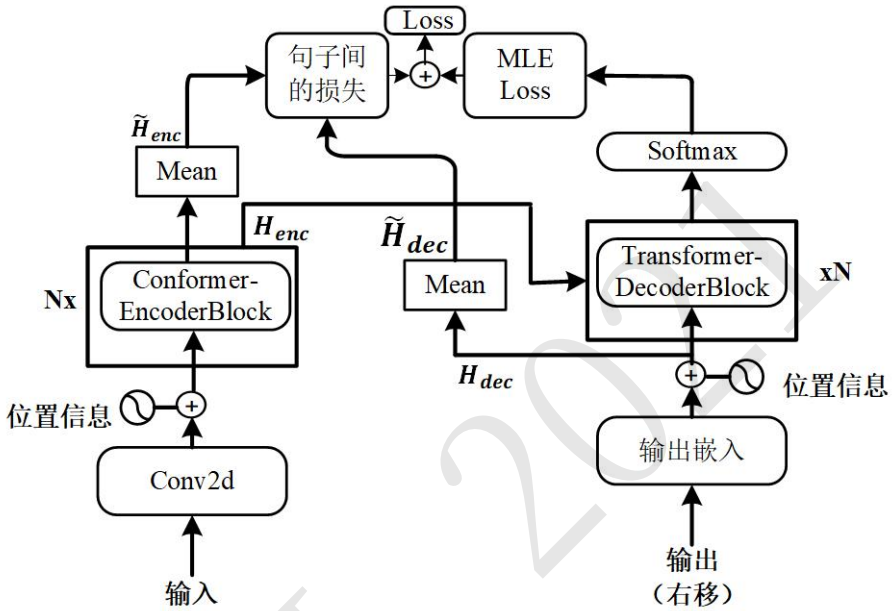

Figure 7: 融合句子层级一致性的Conformer模型

首先，我们需要获取源语音和目标文本的句子层级表示。(Le and Mikolov, 2014)研究表明，平均运算是序列词句子表示的一种有效方法。$\widetilde{H}_{enc}$表示对编码器的输出H_{enc}进行平均运算所得，$\widetilde{H}_{dec}$表示对解码器的输入H_{dec}进行平均运算所得。我们设计句子层级的损失函数L_{mse}用于衡量源语音和目标文本句子水平向量之间的距离，如式(10)所示:

$$L_{mse} = \left\| \widetilde{H}_{enc} - \widetilde{H}_{dec} \right\|^2 \tag{10}$$

最后，我们的目标是提高语音识别的水平，缩短句子水平上的距离。因此，我们模型的最终目标由两部分组成，如式(11)所示:

$$L = L_{mle} + L_{mse} \tag{11}$$

其中L_{mle}表示Conformer模型原本的损失计算公式。

4.2 实验结果及分析

4.2.1 实验数据

本章分别在公开的中文普通话数据集Aishell_1和本文构建的CH_NEWS_ASR数据集上进行实验。Aishell_1数据集包含约178小时的开源版数据，其录音环境是安静的客厅或专业录音室，录音内容为不同领域的话题。在实验的过程中将约150小时的语音作为训练集，约18小时的语音

作为开发集，约10小时的语音作为测试集。训练集、开发集和测试集的讲话者没有重叠，所有的录音都是16kHz的WAV格式。CH_NEWS_ASR数据集为第三章所构建的面向新闻领域的语音识别语料，训练集、开发集和测试集分别为100小时、15小时和12小时。

数据集	数据分布	句子数	时长
Aishell_1	训练集	120098	150
	开发集	14326	18
	测试集	7176	10
CH_NEWS_ASR	训练集	56891	100
	开发集	8312	15
	测试集	6831	12

Table 7: 训练使用数据集详情

4.2.2 实验设置

基于ESPnet2系统对Conformer模型进行改进实现本章方法。实验中提取80维的Fbank特征，将词嵌入向量和隐藏层向量的维度设置为256；编码器设置为12层的Conformer结构其注意力头数为4、卷积模块的cnn_module_kernel为15，dropout 0.1使用swish激活函数；解码器设置为6层的Transformer结构其注意力头数为4，dropout 0.1 其他参数为默认值；总共训练50个epoch。解码时采用了Beam Search策略对其进行预测，beam size的大小为20。

4.2.3 实验结果与分析

表Table 8给出了Aishell_1数据集和CH_NEWS_ASR数据集语音识别任务的结果。表中的第二列使用的不同模型，“Conformer(Base)”为基准模型仅使用ESPnet中的Conformer模型，“Conformer+loss”为4.1的句子层级的端到端语音识别模型。

数据集	模型	W.corr	SER	CER	模型大小/MB
Aishell_1	Conformer (Base)	93.3	43.6	6.8	180.44
	Conformer+loss	**93.7**	**41.6**	**6.4**	180.44
CH_NEWS_ASR	Conformer (Base)	90.5	48.5	9.9	179.45
	Conformer+loss	**91.8**	**45.5**	**9.0**	179.45

Table 8: 融合句子层级一致性Conformer模型语音识别结果

观察Table 8，首先对Aishell_1数据集识别结果进行对比，使用句子层级一致性后在模型大小不变的前提下，字错误率降低了0.4%且句子错误率降低了2%，这表明使用句子层级一致性后能够显著提升句子错误率对字错误率也有较好的提升。然后对CH_NEWS_ASR数据集识别结果进行对比，在使用句子层级一致性后模型的字错误率降低了0.9%和句子错误率降低了3%，相比于Aishell_1数据集有较高的提升，这可能是因为本文构建的数据集句子平均长度较长，使用句子层级一致性促进了句子间的对齐从而更好地提升了模型的识别性能。

5 总结与展望

针对面向新闻领域语音识别资源不足的现状，本文在第三章人工标注构建了数据集CH_NEWS_ASR，该数据集共包含72034条音频-文本对应文件，共计127小时。使用ESPNET框架的Conformer、Transformer、RNN、CTC等模型对数据集进行实验验证了数据集的有效性。此外，本文还提出了融合句子层级层级一致性的Conformer模型，分别在本文构建的数据集和开源的Aishell_1数据集上进行实验，实验结果证明句子层级一致性方法对语音识别非常有用且模型的参数并未增加。未来工作中，如何对模型的内部结构进行改动，更好地融合编码器与解码器的特征信息是需要探索的方向。

参考文献

Ramiz M Aliguliyev. 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, 36(4):7764–7772.

Akshar Bharati, Rajni Moona, Smriti Singh, Rajeev Sangal, and Dipti Mishra Sharma. 2004. Mteval: an evaluation methodology for machine translation systems. In *Proc. SIMPLE Symp on Indian Morphology, Phonology and Lang Engineering*. Citeseer.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Xu Liang, Dongjiao Wang, and Ming Huang. 2010. Improved sentence similarity algorithm based on vsm and its application in question answering system. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, volume 1, pages 368–371. IEEE.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE.

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. *arXiv preprint arXiv:1606.03144*.

Bo-Hao Su, Ta-Wen Kuan, Shih-Pang Tseng, Jhing-Fa Wang, and Po-Huai Su. 2016. Improved tf-idf weight method based on sentence similarity for spoken dialogue system. In *2016 International Conference on Orange Technologies (ICOT)*, pages 36–39. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2018. Sentence selection and weighting for neural machine translation domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1727–1741.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018. Improved training of end-to-end attention models for speech recognition. *arXiv preprint arXiv:1805.03294*.

张宇, 张鹏远, and 颜永红. 2018. 基于注意力lstm 和多任务学习的远场语音识别.

牛米佳, 飞龙, and 高光来. 2020. 蒙古语长音频语音文本自动对齐的研究. 中文信息学报, 34(1):51–57.

王勇和, 飞龙, and 高光来. 2018. 基于tdnn-fsmn 的蒙古语语音识别技术研究. 中文信息学报, 32(9):28–34.