*INTERSPEECH 2019年*

# 2019年9月15日至19日，奥地利格拉茨

Online Hybrid CTC/Attention Architecture for End-to-end Speech Recognition

用于端到端语音识别的在线混合CTC/注意结构

*苗浩然*1*,*2*、程高峰*1*,*2*，张鹏远*1*,*2*，李塔*1*,*2*，颜永红*1*,*2*,*3 1

中国声学研究所语音声学与内容理解重点实验室

中国科学院大学3

中国科学院新疆理化技术研究所新疆少数民族语音与语言信息处理实验室

{苗浩然、程高峰、张鹏远、利塔、颜永红}@hccl.ioa.ac公司中国

# 摘要

The hybrid CTC/attention end-to-end automatic speech recognition (ASR) combines CTC ASR system and attention ASR system into a single neural network. Although the hybrid CTC/attention ASR system takes the advantages of both CTC and attention architectures in training and decoding, it remains challenging to be used for streaming speech recognition for its attention mechanism, CTC prefix probability and bidirectional encoder. In this paper, we propose a stable monotonic chunkwise attention (sMoChA) to stream its attention branch and a truncated CTC prefix probability (T-CTC) to stream its CTC branch. On the acoustic model side, we utilize the latencycontrolled bidirectional long short-term memory (LC-BLSTM) to stream its encoder. On the joint CTC/attention decoding side, we propose the dynamic waiting joint decoding (DWDJ) algorithm to collect the decoding hypotheses from the CTC and attention branches. Through the combination of the above methods, we stream the hybrid CTC/attention ASR system without much word error rate degradation.

Index Terms: speech recognition, CTC, attention, online decoding

混合CTC/注意力端到端自动语音识别（ASR）将CTC-ASR系统和注意力ASR系统结合成一个单一的神经网络。尽管混合式CTC/attention-ASR系统在训练和解码方面综合了CTC和attention两种体系结构的优点，但由于其注意机制、CTC前缀概率和双向编码等优点，在流式语音识别中的应用仍然具有挑战性。本文提出了一种稳定的单调分块注意流（sMoChA）和一种截断的CTC前缀概率流（T-CTC）。在声学模型方面，我们利用延迟控制的双向长-短期存储器（LC-BLSTM）来传输编码器。在联合CTC/注意解码方面，我们提出了动态等待联合解码（DWDJ）算法来收集来自CTC和注意分支的解码假设。通过上述方法的结合，在不降低误码率的前提下，实现了CTC/attention-ASR混合系统的流化。

索引词：语音识别，CTC，注意力，在线解码

2019 interspeech

# 1.     介绍

端到端语音识别系统在大规模自动语音识别（ASR）任务中表现出了良好的性能，与传统的混合系统相比具有一定的竞争力。端到端系统集成了声学模型、词汇和语言模型，将声学特征直接转化为目标标签。端到端语音识别领域采用了两种主流框架。一种是帧同步预测，即每个输入帧有一个目标标签。连接主义时间分类（CTC）[1]丢失函数一直被用来训练帧同步模型。另一种是标签同步预测，即ASR模型决定何时输出目标标签。基于注意的编码器-解码器体系结构[2，3]被广泛应用于这样的框架中，其中注意决定应该关注哪些编码器特性。

到目前为止，已经提出了一些改进来提高CTC框架和基于注意的编解码框架的性能。为了消除CTC的条件独立性假设，一些工作已经将注意机制纳入了CTC框架[4，5，6]。Listen，attent和Spell（LAS）[7]应用了金字塔BLSTM，使注意力更容易从子采样特征中建模更广泛的输入上下文。此外，多头注意[8,9]，自我注意网络[10,11]和

此外，还介绍了其他一些复杂的注意事项。最近，有人提出了混合CTC/注意结构[12]，将CTC和注意框架结合成一个单一的神经网络。在这个架构中，CTC分支将引导注意力进行单调的排列，因此CTC和注意力的联合可以产生高质量的假设。

由于双向编码网络和全局注意机制的存在，目前大多数竞争激烈的端到端ASR系统不适合在线ASR任务。在混合CTC/注意体系结构中，CTC和注意分支在联合CTC/注意解码中以离线方式执行[13]。幸运的是，已经有一些工作专注于低延迟双向声学建模[14,15]和在线注意，如硬单调注意[16]和单调分块注意（MoChA）[17]。此外，最近还提出了一些在线端到端ASR模型[17、18、19]。

这项工作是第一次尝试流混合CTC/注意架构。首先，我们发现标准MoChA在我们的系统中是不稳定的，因此我们提出了一种稳定MoChA（sMoChA），它改变了注意力权重的计算方式，从而取代了全局注意力。第二，利用基于CTC的网络对音频进行分割，并在分割后的音频上而不是在完整的音频上计算截断CTC（T-CTC）前缀的概率。这项研究表明，T-CTC前缀概率与原始CTC前缀概率非常接近[20]。在对CTC和attention分支进行流化处理后，设计了动态等待联合译码（DWJD）算法来解决波束搜索中这两个分支异步预测标签的问题。最后，本文实现了一个在线混合CTC/attention体系结构，并在LibriSpeech上进行了实验。与离线混合CTC/attention结构相比，在线混合CTC/attention结构在test clean/test other上的绝对字错误率下降了1.8%/3.3%。

# 2.     相关工程

## 2.1.    混合CTC/注意力架构

混合CTC/attention体系结构由编码器、基于注意的解码器和基于CTC的网络组成。给定-长度声学特征={x1，···，xT}，编码器产生-长度编码器特征={h1，···，hU}（≤T）。基于注意的解码器在注意分支中逐步接收和预测目标标签。以下*T型十U型小时U型小时*

**http://dx.doi.org/10.21437/Interspeech.2019-2018**

公式描述了具体过程：

*小时*=编码器（X），（1）=注意（si−1，H），（2）∼解码器（si−1，yi−1，ci），（3）*ci公司彝语*

在基于注意的解码器中，隐藏状态、“上下文”向量和输出标签在哪里。考虑到长度标签序列={y1，···，yL}，CTC和注意分支将分别计算序列的后验概率（Y | X）和（Y | X）。最后，混合CTC/注意力架构的训练目标定义如下：*四、词、义我是的第*反恐委员会*第*收件人

*我*=λlogPctc（Y | X）+（1−λ）logPatt（Y | X），（4）其中是满足0≤λ≤1的超参数。*λ*

## 2.2.    单调的分块注意（摩卡）

MoChA的目标是学习编码器特性和标签序列之间的单调对齐。除此之外，摩卡在一个长度的区块内执行局部注意。算法1[17]总结了MoChA如何在解码阶段计算“上下文”向量。是预测标签时块停止的位置。第3-13行通过保持向前移动来执行单调行为。具体来说，在第4行中定义为选择的概率。在6-9号线，如果≥0.5，MoChA将在至之间进行局部关注。*小时是的w型ci公司钛彝语钛皮，jhj公司彝语hj公司*−宽+1*hj公司皮，j*

算法1摩卡解码算法

输入：编码器特性={h1，···，hU}，输出索引，解码器隐藏状态，输出标签，端点，sigmoid函数（·），*小时我硅彝语钛σ*

注意块宽度*w型*

1： 初始化=0，y0=hsosi，t0=1，i=1*s码*0

2： 而−6=heosi do*彝语*1

3： for=ti−待办事项*日本*1 *U型*

4： =σ（能量（si−））*皮，j*1*，黄杰*

5： 如果≥0.5，则*皮，j*

6： for=j−w+1待办事项*k公司日本*

7： =能量（si−）*用户界面，k*1*，香港*

8： 结束

9:

10:

11： 中断

12： 结束if

13： 结束

14： 如果0.5，∀j∈{ti−···，U}，那么*皮，j<*1*,*

15： =0，=ti负极*ci公司钛*1

16： 结束if

17： ∼解码器（si−-），i=i+1*彝语*1*，彝语*1*，ci公司*

18： 结束时



然而，算法1不能适应反向传播框架。我们必须根据算法2[17]计算训练阶段的期望值。为了模拟向前移动的行为，在算法2的第5行中提供了选择的预期概率。*ci公司αi，jhj公司彝语*

能量和能量函数的选择是

>

能量(

(5)

能量（si−1，hj）=v>tanh（Wssi−1+Whhj+b），

(6)

其中是标量、向量和矩阵。它们都是可学习的参数。*g、 右v、 b类西，西*

算法2摩卡训练算法

输入：编码器特征={h1，···，hU}，输出索引，解码器隐藏状态，输出标签，sigmoid函数（·），注意块宽度，高斯噪声*小时我硅彝语σw型*

1： =0，y0=hsosi，α0,0=1，α0，k=0（k6=0），i=1*s码*0

2： 而−6=heosi do*彝语*1

|  |  |
| --- | --- |
| 6: | 结束 |
| 7: | for=1待办事项*日本U型* |
| 8: | *乌伊，j*=能量（si−）1*，黄杰* |

3： 对于=1到4：=σ（能量(*日本U型皮，j*

5:

9:

*日本*

12： ∼解码器（si−-），i=i+1*彝语*1*，彝语*1*，ci公司*

13： 结束时



## 2.3.    延迟控制BLSTM（LC-BLSTM）

LC-BLSTM旨在减少上下文敏感块BPTT中的冗余计算[14]。具体来说，首先将输入序列分割成固定长度的块。然后，在每个块之后连接未来的帧作为正确的上下文。对于每个区块，前向LSTM的隐藏状态从前一个区块复制而来，反向LSTM的隐藏状态由正确的上下文而不是完整的未来上下文提供。因此，LC-BLSTM的延迟被限制为+Nr帧[21]。*数控编号数控*

# 3.        在线混合CTC/Attention架构

## 3.1.    稳定单调分块注意（sMoChA）

在第2.2节中可以发现，除非在算法2中是离散的，否则解码阶段不等价于训练阶段，即∈{0,1}。这被称为训练和解码场景之间的不匹配。为了缓解这个问题，MoChA通过初始化到等式5中的负值或正值[16]，强制≈0或≈1。然而，我们发现这种初始化策略会导致注意力权重迅速衰减到零。我们可以按以下形式详细重写：*ci公司皮，j皮，j第第右αi，j*

*.* (7)

显然，序列{αi，·}指数衰减（1−p），因此当≈1时，注意力权重将沿指数衰减为零。类似地，序列{α·，j}指数衰减，因此当≈0时，注意力权重将沿指数衰减为零。因此，在标准摩卡咖啡中存在着离散性和稳定性之间的两难选择。*日本第第我第*

为了解决这个问题，我们提出了sMoChA，去掉了期望选择概率的项并计算如下：*αi*−1，j

*日本*−1

*αi、 日本*=p*i、 日本*Y（1−p）*i、 k公司*).(8)

*k公司*=1

我们初始化为负值，以保证序列{αi，·}的离散性，并防止序列{αi，·}急剧衰减。此外，我们还在第4.2节中评估了标准摩卡咖啡和我们的摩卡咖啡的稳定性和离散性。*右皮，j*

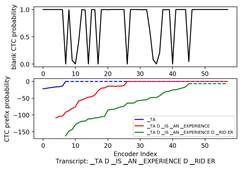


图1：完整和截断CTC前缀概率的比较：上图是CTC空白符号概率分布的示例。下图显示了CTC前缀概率在对数域中的动态求和。实线表示计算部分，虚线表示省略部分。实线和虚线的交点是截断点。

## 3.2.    截断CTC（T-CTC）前缀概率

在混合CTC/注意体系结构中，采用联合CTC/注意解码[13]来提高性能。CTC概率被认为可以有效地排除由注意分支产生的错误假设。在波束搜索中，我们将最终假设的前缀称为部分假设，并将其表示为。联合CTC/注意力解码得分定义如下：*我我*

            分数（l）=λlogPctc（l | H）+（1−λ）logPatt（l | H）。(9)

在波束搜索中，所有作为前缀的标签序列的累积CTC概率用于评分，称为CTC前缀概率[20]，可通过以下公式计算：*我我*

*U型*

*第*ctc（l，…| H）=XPctc（l·ν| H）=XPctc（l | H[1:t]），（10）

*νt*=1

其中是任意后缀字符串。（l，…| H）是等式10中（l | H）的替代物，但它依赖于整个帧，妨碍在线解码。为了消除全局依赖性，我们可以利用峰值CTC概率分布来分割编码器特征。我们建议T-CTC前缀概率截断等式10中的总和，如下所示：*ν第*反恐委员会*第*反恐委员会*我*

*k公司*

*第*反恐委员会（长|高）≈XP反恐委员会（长|高）[1:t]),(11)

*t型*=1

其中求和的上限满足CTC分支在完成生成。因此，（HBI×HK（1））0和（HBI×HK）1在CTC概率分布中，其中HBI代表空白符号。我们需要跟踪每个假设的扩展。此外，图1证明了T-CTC前缀概率是CTC前缀概率的合理近似值。*k公司我香港第第k公司*

## 3.3.    动态等待联合译码

由于在线注意力和CTC分支的同步预测在理论上没有保证，我们提出了动态等待联合译码算法，其中两种算法3动态等待联合译码（DWJD）

输入：声学特征={x1，x2，···}，输出索引，编码器索引，编码器特征，解码器隐藏状态，输出标签，起始点和，布尔变量doAtt和doCTC，sigmoid函数（·），阈值*十我日本hj公司硅彝语t型*收件人*t型*反恐委员会*σθ*

|  |  |
| --- | --- |
| 1： =0，y0=l=hsosi，tatt=tctc=1，j=i=1，θ=0.5*s码*0  2： 而−6=heosi do*彝语*1 | |
| 三： | doAtt=true，doCTC=true，j=min{tatt，tctc} |
| 4: | 而不是最后的编码器功能做*hj公司*−1 |
| 5: | *hj公司*=编码器（X） |
| 6: | 如果≥tatt和doAtt那么*日本* |
| 7: | *皮，j*=σ（能量（si−））1*，黄杰* |
| 8: | 如果≥0.5，则*皮，j* |
| 9: | *t型*收件人=j，doAtt=false |
| 10: | 结束if |
| 11: | 结束if |
| 12: | 如果和doCTC那么*j>t*反恐委员会 |
| 13: | 计算（hbi | hj）*第* |
| 14: | 如果（hbi | hj−1）<θ且（hbi | hj）≥θ，则*第第* |
| 15: | *t型*反恐委员会=j，doCTC=false |
| 16: | 结束if |
| 17: | 结束if |
| 18: | 如果doAtt=false，doCTC=false，则 |
| 19: | 打破 |
| 20: | 结束if |
| 21: | *日本*=j+1 |
| 22: | 结束时 |
| 23: | 如果doCTC=true，则 |
| 24: | *t型*反恐委员会=j−1 |
| 25: | 结束if |
| 26: | *c级我*=sMoChA（s−1]）*我，小时*[1:*t型*收件人 |
| 27: | *彝语*∼解码器（si−-）1*，彝语*1*，ci公司* |
| 28: | *我*=l·yi，i=i+1 |
| 29: | 计算ctc（l | H[1:tctc]），Patt（l | H[1:tatt]）*第* |
| 30:结束时 | |



在线分支在生成部分假设时相互等待。该算法在算法3中给出。第4-22行中的内环在预测下一个标签时计算必要的编码器特征。在第6-11行中，记录注意块开始的位置，doAtt决定是否执行注意。在第12-17行中，记录第3.2节中讨论的截断点，doCTC确定是否计算CTC概率。为了等待较慢的分支，编码器索引总是设置为最小值和。*彝语t型*收件人*t型*反恐委员会*日本t型*收件人*t型*反恐委员会

在我们的实验中，我们还利用一个单独训练的LSTM语言模型[22]，以及算法3中计算的（l | H[1:t]）和（l | H[1:t]），对波束搜索中的部分假设进行评分。*第*反恐委员会反恐委员会*第*收件人收件人*我*

# 4.     实验

我们的实验是在一个1000小时阅读的英语语音语料库LibriSpeech上进行的。所有的模特都在960小时的火车上训练。有效集由dev clean和dev other组成。单词错误率（WER）分别在test clean和test other上报告。我们的基线和在线模型基于ESPNet[23]。所有实验均采用83维特征，包括80个滤波器组、基音、增量基音和归一化互相关函数（NCCF），用25ms窗口计算，每10ms移位一次，并选择子词[24]作为输出标签来解决词汇表外的问题。一个5000大小的子词集是通过基于单图语言模型的子词分割算法实现的[25]。我们在混合CTC/注意结构和外部语言模型中都使用了这个子词集。

## 4.1.    基线

在我们的基线，编码器包含2块VGG层

[26]然后是5层BLSTM。两个VGG区块的下采样率为四分之一。译码器是一个2层LSTM，它在训练阶段接收地面真值标签作为前一个预测。位置感知注意机制[3]被用来提高性能。LSTM和注意的隐态维数为1024。外部语言模型包含一个单层LSTM，在LibriSpeech的规范化LM训练文本上进行训练。所有实验在解码阶段使用相同的外部语言模型。清洁试验和其他试验的基线WER分别为4.2%和13.4%，见表1第一行。

## 4.2.    通过sMoChA和T-CTC流式传输CTC/注意力

我们的第一项工作是评估摩卡和斯摩查咖啡的性能。我们分别用MoChA和sMoChA替换基线中的位置感知注意，发现在不同的配置下，根据标准MoChA计算的注意权重容易衰减为零，而sMoChA计算的注意权重更稳定。在图2（b）中，当初始化为负值时，标准摩卡中的注意力权重会沿着解码器索引迅速衰减，如第3.1节所述。sMoChA的注意块宽度为3，等式5中的初始偏差为−4。如表1第3行所示，在test clean和testother上，sMoChA模型的WER分别为4.7%和13.6%。*右右*

我们在图2中描述了通过不同注意机制学习的注意权重，以进行比较。通过比较图2（c）和图2（d），我们发现解码和训练在注意权重上几乎没有差异，证明了我们的sMoChA的离散性。对比图2（a）和图2（c），图2（c）中的注意权重被限制在一个小的块中，不管子词发音多长，这是导致准确度损失的主要原因。

其次，我们研究了T-CTC前缀概率在联合CTC/注意解码中的性能。比较表1中的第1行和第2行，T-CTC前缀概率导致0.1%/0.2%的绝对退化。通过sMoChA和T-CTC前缀概率的结合，我们的系统在没有太多退化的情况下实现了4.8%/13.9%的WERs，如表1的第4行所示。

## 4.3.    通过VGG-LC-BLSTM传输CTC/注意力

我们将基线中的VGG-BLSTM替换为VGG-LC-

BLSTM，保持其余不变。具体来说，我们训练由基线初始化的VGG-LC-BLSTM模型，而不是从头开始。在LC-BLSTM中，块长度为32，右上下文长度为16。VGG-LCBLSTM模型在试验清洁/其他条件下的WER为5.4%/16.4%，见表1第5行。*数控编号*

最后，我们的在线模型由VGG-LC-BLSTM编码器、基于sMoChA的解码器和基于CTC的网络组成。我们通过基线初始化VGG-LC-BLSTM网络。采用DWJD算法实现在线解码过程。我们的在线模型在testclean/other上的WER为6.0%/16.7%，如表1的最后一行所示。第1行、第5行和第7行表明VGG-LC-BLSTM的绝对降解率为1.2%/3.0%，而sMoChA和T-CTC前缀概率的绝对降解率为0.6%/0.3%。表1还提供了与LibriSpeech上其他已发布的在线模型[27]的比较。

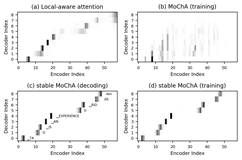
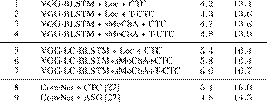


图2：不同注意下的注意权重可视化。需要注意的是，MoChA和sMoChA在解码阶段对一个块进行注意，而在训练阶段对整个帧进行注意。

表1：*不同模型的误字率（WER）：我们在不同的编码器（Enc）上进行了实验，包括VGG-BLSTM、VGG-LSTM和VGG-LC-BLSTM，不同的注意机制（Att）包括位置感知（Loc）和sMoChA。在联合CTC/注意解码中，我们使用CTC或T-CTC前缀概率。第二栏详细介绍了各种组合。*







# 5.     结论

在本文中，我们提出了sMoChA和T-CTC前缀概率来计算注意权重和在线解码话语。在此基础上，提出了基于DWJD的CTC/attention联合解码算法，设计了一种基于VGG-LC-BLSTM编码器、基于sMoChA解码器和基于CTC网络的在线混合CTC/attention语音识别体系结构。我们的混合CTC/attention架构可以在线，在test clean/other of LibriSpeech上，WERs的绝对退化率为1.8%/3.3%。实验还表明，VGGLC-BLSTM编码器的性能损失最大，而sMoChA和T-CTC前缀概率造成的性能下降仅为0.6%/0.3/%绝对WERs。因此，我们的sMoChA和T-CTC前缀概率是可靠的在线方法。研究更好的低延迟编码器网络将包括在我们未来的工作。

# 6.     致谢

本工作得到了国家重点研究开发计划（2016YFB0801203、2016YFB0801200）、国家自然科学基金（11590774、11590770）、新疆维吾尔自治区科技攻关项目（2016A03007-1）、通用电气装备预研项目的部分资助信息系统(编号：JZX2017-0994/Y306），与索尼（中国）有限公司合作项目。

# 7.     工具书类

[1] A.Graves、S.Fernndez、F.J.Gomez和J.Schmidhuber，

“连接主义时间分类：用递归神经网络标记未分段序列数据”，国际机器学习会议，2006年。

[2] J.Chorowski、D.Bahdanau、K.Cho和Y.Bengio，“使用基于注意的递归神经网络进行端到端连续语音识别：第一个结果”，Eprint Arxiv，2014年。

[3] J.Chorowski、D.Bahdanau、D.Serdyuk、K.Cho和Y.Bengio，“基于注意力的语音识别模型”，计算机科学，第10卷，第4期，第429-4392015页。

[4] H.Sak、M.Shannon、K.Rao和F.Beaufays，“递归神经对齐器：序列到序列映射的编码器-解码器神经网络模型”，Interspeech，2017，第1298-1302页。

[5] L.Dong，S.Zhou，C.Wei，和X.Bo，“扩展用于普通话流式端到端语音识别的递归神经对齐器”，2018年。

[6] A.Das、J.Li、Z.Rui和Y.Gong，“用注意力建模推进连接主义时间分类”，2018年。

[7] W.Chan、N.Jaitly、Q.Le和O.Vinyals，“听、听和拼写：用于大词汇量会话语音识别的神经网络”，IEEE国际声学会议，2016年。

[8] C.C.Chiu、T.N.Sainath、Y.Wu、R.Prabhavalkar、P.Nguyen、Z.Chen、A.Kannan、R.J.Weiss、K.Rao和K.Gonina，“具有序列间模型的最先进语音识别”，2018年。

[9] T.Hayashi、S.Watanabe、T.Toda和K.Takeda，“端到端语音识别的多头解码器”，arXiv预印本附件十四：1804.08050, 2018.

[10] L.Dong，S.Xu和B.Xu，“语音转换器：用于语音识别的无重复序列到序列模型”，2018 IEEE声学、语音和信号处理国际会议（ICASSP）。IEEE，2018年，第5884-5888页。

[11] A.Vaswani、N.Shazeer、N.Parmar、J.Uszkoreit、L.Jones、A.N.Gomez、L.Kaiser和I.Polosukhin，“关注是你所需要的一切”，2017年。

[12] S.Watanabe、T.Hori、S.Kim、J.R.Hershey和T.Hayashi，“用于端到端语音识别的混合ctc/注意力架构”，IEEE信号处理精选主题期刊，第11卷，第8期，第1240–1253页，2017年。

[13] T.Hori、S.Watanabe和J.Hershey，“端到端语音识别的联合ctc/注意力解码”，计算语言学协会第55届年会论文集（第1卷：长论文），第1卷，2017年，第518-529页。

[14] C.Kai和H.Qiang，“通过上下文敏感块bptt方法训练lvcsr的深层双向lstm声学模型”，IEEE/ACM音频语音和语言处理学报，第24卷，第7期，第1185-119320016页。

[15] Y.Zhang，G.Chen，D.Yu，K.Yaco，S.Khudanpur和J.Glass，“用于远程语音识别的公路长-短期记忆RNN”，2016 IEEE声学、语音和信号处理国际会议（ICASSP）。IEEE，2016年，第5755-5759页。

[16] C.Raffel，M.-T.Luong，P.J.Liu，R.J.Weiss和D.Eck，“通过强制单调排列实现在线和线性时间注意”，《第34届机器学习国际会议论文集》第70卷。JMLR公司。org，2017年，第2837-2846页。

[17] C.-C.Chiu和C.Raffel，“单调的分块注意”，arXiv预印本附件十四：1712.05382, 2017.

[18] R.Fan，P.Zhou，W.Chen，J.Jia和G.Liu，“基于在线注意力的语音识别模型”，arXiv预印本附件十四：1811.05247, 2018.

[19] L.Dong，F.Wang，和B.Xu，“自我注意对齐器：使用自我注意网络和区块跳跃的asr的一种新型控制端到端模型”，arXiv预印本附件十四：1902.06450, 2019.

查看出版物统计信息

[20] K.Kawakami，“递归神经网络监督序列标记”，慕尼黑工业大学博士论文，2008年。

[21]V.Peddinti、Y.Wang、D.Povey和S.Khudanpur，“使用时间卷积和lstms的低延迟声学建模”，《IEEE信号处理快报》，第25卷，第3期，第373–377页，2018年。

[22]T.Hori、S.Watanabe、Y.Zhang和W.Chan，“使用深度cnn编码器和rnn lm的联合ctc基于注意力的端到端语音识别的进展”，arXiv预印本附件十四：1706.02737, 2017.

[23]S.Watanabe，T.Hori，S.Karita，T.Hayashi，J.Nishitoba，Y.Unno，N.E.Y.Soplin，J.Heymann，M.Wiesner，N.Chen等人，“Espnet:端到端语音处理工具包”，arXiv预印本附件十四：1804.00015, 2018.

[24]R.Sennrich、B.Haddow和A.Birch，“具有子词单位的稀有词的神经机器翻译”，arXiv预印本附件十四：1508.07909, 2015.

[25]T.Kudo，“子词正则化：改进具有多个子词候选的神经网络翻译模型”，arXiv预印本附件十四：1804.10959, 2018.

[26]K.Simonyan和A.Zisserman，“用于大规模图像识别的超深卷积网络”，arXiv预印本附件十四：1409.1556, 2014.

[27]V.Liptchinsky、G.Synnave和R.Collobert，“基于门控网络的语音识别”，arXiv预印本附件十四：1712.09444, 2017.