Transformer-Based Online Ctc/Attention End-To-End Speech Recognition Architecture

基于变压器的在线CTC/注意端到端语音

识别体系结构

*苗浩然*1*,*2*、程高峰*1*，高长峰*1*,*2*，张鹏远*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

新疆理工大学新疆少数民族语言文字信息处理实验室

中国科学院理化研究所

# 摘要

Recently, Transformer has gained success in automatic speech recognition (ASR) field. However, it is challenging to deploy a Transformer-based end-to-end (E2E) model for online speech recognition. In this paper, we propose the Transformer-based online CTC/attention E2E ASR architecture, which contains the chunk self-attention encoder (chunk-SAE) and the monotonic truncated attention (MTA) based self-attention decoder (SAD). Firstly, the chunk-SAE splits the speech into isolated chunks. To reduce the computational cost and improve the performance, we propose the state reuse chunk-SAE. Sencondly, the MTA based SAD truncates the speech features monotonically and performs attention on the truncated features. To support the online recognition, we integrate the state reuse chunk-SAE and the MTA based SAD into online CTC/attention architecture. We evaluate the proposed online models on the HKUST Mandarin ASR benchmark and achieve a 23:66% character error rate (CER) with a 320 ms latency. Our online model yields as little as 0:19% absolute CER degradation compared with the offline baseline, and achieves significant improvement over our prior work on Long Short-Term Memory (LSTM) based online E2E models.

Index Terms— Transformer, end-to-end speech recognition, online speech recognition, CTC/attention speech recognition

近年来，Transformer在自动语音识别领域取得了成功。然而，部署一个基于变压器的端到端（E2E）模型用于在线语音识别是一个挑战。本文提出了一种基于Transformer的在线CTC/attention E2E-ASR体系结构，包括块自关注编码器（chunk-SAE）和基于单调截断注意（MTA）的自关注解码器（SAD）。首先，组块SAE将语音分割成孤立的组块。为了降低计算量和提高性能，我们提出了状态重用的方法。其次，基于MTA的SAD算法对语音特征进行单调截断，并对截断后的特征进行关注。为了支持在线识别，我们将状态重用块SAE和基于MTA的SAD集成到在线CTC/attention体系结构中。我们在香港科大普通话ASR基准上评估了所提出的在线模型，并在320毫秒的延迟下达到了23.66%的字符错误率（CER）。与离线基线相比，我们的在线模型仅产生0.19%的绝对CER降解，并且与我们先前基于长短时记忆（LSTM）的在线E2E模型相比取得了显著的改进。

*索引项*-变压器、端到端语音识别、在线语音识别、CTC/注意力语音识别

*2020*.01.15 arxiv, icassp 2020

# 1导言

近年来，端到端（E2E）自动语音识别（ASR）在ASR界得到了广泛的应用[1，2，3，4，5，6]。E2E-ASR模型简化了DNN/HMM混合ASR模型，用一个深度神经网络代替声学、语音和语言模型，直接将语音转录成文本。迄今为止，E2E ASR模型在ASR领域取得了显著的改进[4，5，6]。混合连接主义时间分类（CTC）/注意E2E ASR体系结构[6]由于结合了CTC模型和注意模型的优点而备受关注。在训练过程中，CTC目标作为辅助任务附加到基于注意的编解码模型中。在解码过程中，波束搜索采用联合CTC/注意解码的方法[7]。然而，事实并非如此很难部署在线CTC/attention E2E ASR架构，因为全局注意机制[8]和CTC前缀分数[6,9]，它们依赖于整个输入语音。我们之前的工作[10,11]从模型结构和解码算法两个方面对这种架构进行了研究。在模型结构方面，我们提出了稳定的单调分块注意机制（sMoChA）[10]和单调截断注意机制（MTA）[11]，并采用时延控制的双向长短时记忆（LC-BLSTM）作为低时延编码器。在译码方面，我们提出了在线联合译码方法，包括截断CTC（T-CTC）前缀分数和动态等待联合译码（DWJD）算法[10]。

最近，变压器[12]在ASR领域取得了成功[13，14，15]。基于变压器的模型是可并行的，并与递归神经网络竞争[16]。然而，vanilla Transformer不适用于在线任务，原因有二：第一，自注意编码器（SAE）计算整个输入帧上的注意权重；第二，自注意解码器（SAD）计算SAE的整个输出上的注意权重。

在本文中，我们将变压器流化并将其集成到CTC/e ASR体系结构中。在SAE方面，我们提出了块SAE，它将输入语音分割成固定长度的孤立块。受Transformer-XL[17]的启发，我们进一步提出了状态重用块SAE，它重用先前块的存储状态，以降低计算成本。在SAD方面，我们提出了基于MTA的SAD，它关注SAE被截断的历史输出。最后，我们通过在线联合译码方法提出了基于变压器的在线CTC/attention E2E ASR体系结构[10]。实验结果表明，该模型在HKUST上的字符错误率为23.66%，与离线模型相比，绝对错误率仅下降0.19%。

本文的其余部分组织如下。在第2节中，我们描述了我们先前工作中提出的在线CTC/attention E2E体系结构[10，11]。在第3节中，我们将介绍Transformer体系结构。在第4节中，我们描述了基于在线变压器的CTC/注意体系结构。第5节和第6节分别介绍了实验和结论。

# 2在线CTC/E体系结构

在我们之前的工作[10]中，我们提出了一种在线混合CTC/attention E2E ASR架构，它由LC-BLSTM编码器、sMoChA和LSTM解码器组成。在培训过程中，我们引入CTC目标作为辅助任务，损失函数定义如下：

                                    L=αLdec+（1−α）Lctc，（1）

其中是超参数，Ldec和Lctc是解码器和CTC的损失函数。在解码过程中，我们采用在线联合解码方法，其定义如下：*α*

## *是的*ˆ=arg max{λlogPdec（Y | X）+（1−λ）logPt（Y | X）*-*反恐委员会

*是的*∈Y∗

                                                            +γlogPlm（Y）}，（2）

其中（Y | X）和（Y | X）是以来自解码器和T-CTC的输入帧为条件的假设的概率[10]，并且（Y）是语言模型概率。超参数和可调。对于在线解码，我们提出了DWJD算法[10]：1）协调编码器中的前向传播和解码器中的波束搜索；2）解决基于sMoChA的解码器和CTC输出的不同步预测。*第*12月*第*t型*-*反恐委员会*是的十第*流光溢彩*λγ*

MTA[11]在截断的历史编码器输出上执行注意，通过利用更长的历史，它的性能优于sMoChA。形式上，我们将和分别表示为第-个解码器状态和第-个编码器输出。类似于单调的分块注意[18]，MTA将截断编码器输出的概率定义为：**qhh公司***我日本我日本皮，j日本*

## *皮，j*=乙状结肠(

其中矩阵、向量和标量是可训练的参数。然后，注意力权重由以下公式计算：**WWbv公司**12*克右艾，j*

*,* (4)

其中表示在处截断编码器输出并在之前跳过编码器输出的概率。在解码期间，MTA通过以下步骤确定第-个解码器的截断端点：*艾，j***啊***日本日本钛我*

*齐，j*=I（pi，j>0.5∧j≥ti−1），（5）其中表示在处截断或不截断编码器输出的指示符，I表示指示符函数。根据等式5中的条件≥ti−1，MTA强制端点以从左到右的模式移动。对于某些情况，Once=1，MTA设置为。最后，MTA注意截断的编码器输出：*齐，j***小时***日本日本齐，j日本钛日本*

*,* (6)

其中是第-个解码器步骤的逐字母隐藏向量。**右***我我*

在培训期间，MTA对整个编码器输出进行关注：

*,* (7)

其中表示编码器输出的数量。*T型*

# 三。变压器结构

Transformer[12]遵循编码器-解码器架构，使用编码器和解码器的堆叠自关注和位置前馈层。在本节中，我们将简要介绍Transformer体系结构。

## 3.1. 多头注意

Transformer采用缩放的点积注意将查询和一组键值对映射到输出，如下所示：

                              注意（Q，K，V）=Softmax（，（8）

*米*

其中矩阵∈R×d、∈R×d和∈R×d表示查询、键和值，表示查询和键（或值）的个数，表示表示表示维数。**Q K V公司***不米米米米米不米dm公司*

Transformer没有执行单一的注意功能，而是使用多个头部注意，从不同的表示子空间联合学习查询和键之间的不同关系，如下所示：

## 多头（Q，K，V）=Concat（头1，···，头

headh=注意（QWhQ，KWhK，VWhV），（10）其中表示头数和=dm/H。矩阵∈R×dand∈R×dare可训练参数。*小时丹麦***栈单***Od级米米总部，K，Vd级米k公司*

由于Transformer缺乏对序列顺序的建模，在[12]中的工作建议使用不同频率的正弦和余弦函数来执行位置编码。

### 3.2. 自我注意编码器（SAE）

SAE由一堆相同的层组成，每个层有两个子层，即一个自关注层和一个位置前馈层。SAE的输入是ASR任务中的声学帧。自我注意层采用多头部注意，其中查询、键和值是前一层的输入。此外，SAE在每个子层之后使用剩余连接[19]和层规范化[20]。

### 3.3. 自我关注解码器（SAD）

SAD还包括一堆相同的层，每个层有三个子层，即一个自注意层、一个编码器注意层和一个位置前馈层。SAD的输入是右移输出标签的嵌入。为了防止将来在self-attention中访问输出标签，随后的位置被屏蔽。在编解码器中，查询是当前层的输入，而键和值是SAE输出。此外，SAD算法在每个子层之后还使用了剩余连接和层归一化。

# 4基于变压器的在线CTC/e体系结构

在本节中，我们提出了基于Transformer的在线E2E模型，该模型由具有或不具有重用存储状态的块SAE和基于MTA的SAD组成。基于变压器的在线CTC/E2E架构如图1所示。

## 4.1. 块SAE

为了流化SAE，我们首先提出了块SAE，它将语音分割成中心长度不重叠的孤立块。为了获得上下文信息，我们将每个块前面的左框架拼接为历史上下文，后面的右框架拼接为历史上下文*数控荷兰编号*

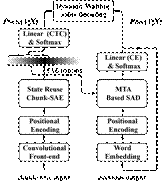


图1。基于变压器的在线CTC/E体系结构。

未来背景。拼接帧仅作为上下文，不提供输出。使用预定义的参数和，每个区块SAE输出的感受野限制为+Nc+Nr，区块SAE的延迟限制为。*数控荷兰编号荷兰编号*

## 4.2. 状态重用块SAE

在区块SAE中，为每个区块重新计算历史上下文。为了减少计算量，我们将计算出的隐藏状态存储在中心上下文中。然后，在计算新的块时，我们在与历史上下文相同的位置重用先前块中存储的隐藏状态，这一点受到TransformerXL[17]的启发。图2示出了具有或不具有重用隐藏状态的块SAE之间的差异。正式地，和**小时**分别表示第-层中第-个块的已存储和新计算的隐藏状态。然后，第-th self attention层中的第-thchunk的查询、键和值定义如下：*τ我τ我*

**千伏百出***我τ,我τ,τ我*=小时*我τ*−1*,*电子*τ我*−1*,*电子*我τ*−1*,* (11)

                                *我我*−1升−1

               式中=Concat（SG（sτ），hτ）。(12)

在等式12中，函数SG（·）表示停止梯度。因此，状态重用块SAE的复杂度降低了（Nl+Nc+Nr）的因子。*荷兰/*

此外，状态重用块SAE捕获了块之外的长期依赖性。假设状态重用区块SAE是分层的，左侧的感受野延伸到远至·Nl帧，比区块SAE宽得多。*我我*

## 4.3. 基于MTA的SAD

为了流式传输SAD，我们提出了基于MTA的SAD，以从左到右的单调方式截断感受野，并对SAE的截断输出进行注意。具体来说，我们用MTA代替每个SAD层中的编码器-解码器注意，如图2所示。假设表示维度为，MTA在训练期间并行执行，如下所示：*dm公司*

             MTA（Q，cumprod（1−P））VVV，（13）

**QWWK公司***问k公司*>>

**第**=乙状结肠(√+r+ε），(14)

*d级米*

其中矩阵∈Rand标量偏差是可训练参数，表示噪声。我们将={pi，j}定义为**可湿性粉剂**·*d级米*×天*米右*

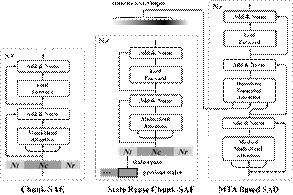


图2。区块SAE、状态重用区块SAE和基于MTA的SAD的示例。

截断概率矩阵，其中指示为了预测第-个输出标签而截断第-个SAE输出的概率。在等式13中，累积积函数cumprod（x）=*皮，j日本我*

而cumprod（·）应用于。符号表示元素的乘积。**第**

MTA通过可训练标量学习等式14中乙状结肠前激活的适当偏移量。为了防止cumprod（1−P）消失为零，我们将其初始化为负值，例如在我们的实验中=−4。为了鼓励截断概率的离散性，我们只需在训练期间将零均值、单位方差高斯噪声添加到sigmoid激活前。*右右右ε*

在解码过程中，我们必须逐行计算={pli，j}中的元素，其中是第-层的截断概率矩阵。在预测第-个输出标签时，我们定义为属于第-层的截断端点。然后，终点由以下公式确定：**第页***我我我tli公司我我*

*,* (15)

其中表示截断或不截断第层SAE输出的指示符，I表示指示符函数。一旦=1，我们设置为，这意味着第-层的感受野仅限于SAE输出。假设基于MTA的SAD由层组成，那么每个解码步骤都有端点。每层中截断的SAE输出的数量不会影响其他层。因此，我们将端点的最大值定义为基于MTA的SAD的感受野。*zi，jl日本我zi，jl日本tli公司日本我tli公司我我我*

# 5实验

## 5.1. 语料库

我们使用科大华语会话电话[21]评估了我们的模型。香港科大由约200小时的训练列车及约5小时的测试列车组成。我们从train集合中提取了4000个语句作为我们的开发集合。为了提高识别的准确率，我们将速度扰动应用于系数为0.9和1.1的静止序列集。

## 5.2. 模型描述

我们使用ESPnet工具箱构建了所有的在线模型[22]。对于输入，我们使用了83维特征，包括80维可过滤1。香港科技大学不同变形器ASR模型之字元错误率。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编码器 | 解码器 | 州  重新使用 | 编码器速度比 | 开发 | 试验 |
| 美国汽车工程师协会 | 悲伤 | – | 2.8 | 24.12 | 23.47 |
| 块- | MTA公司- | × |  |  |  |
| 美国汽车工程师协会 | 悲伤 |  | 1.5 | 24.45 | 23.65 |

音高、音高、增量音高和归一化互相关函数。特征以25ms窗口计算，每10ms移位一次。输出采用3655个词汇量集，包括3623个汉语普通话字符、26个英语字符以及6个非语言符号，分别表示笑声、噪声、发声噪声、空白、未知字符和sos/eos。

我们使用两层卷积神经网络（CNN）作为前端。每个CNN层有256个滤波器，每个滤波器的核尺寸为3×3，步长为2×2，因此前端的时间减少了1/4。SAE和SAD分别有12层和6层。所有子层以及嵌入层都产生了256维的输出。在多头注意网络中，头数为4。在位置前馈网络中，内维数为2048。此外，我们以科大成绩单为外部语言模型，训练了一个2层1024维的LSTM网络，并采用了上述3655大小的词汇集。

在训练期间，我们使用了CTC/注意力联合训练（=0.7）和Adam优化器以及Noam学习率计划（25000个温暖步骤）[12]，并训练了30个时期。为了防止过度拟合，我们在每个子层中使用了dropout[23]（dropout rate=0.1），在输出层中使用了uniform label smoothing[24]（惩罚=0.1），并使用了模型平均方法来平均最近10个时期的模型参数。在解码过程中，我们采用在线联合解码方法，结合T-CTC前缀分数（=0.5）和语言模型分数（=0.3）对假设进行删减，波束大小为10。*αλγ*

## 5.3. 是否重用状态

在表1中，我们比较了具有或不具有重用状态的chunkSAE的速度和性能。在比较期间，在线模型的上下文配置保持不变，即=Nc=Nr=64。首先，我们使用Intel（R）Xeon（R）Silver 4114 CPU，2.20GHz的服务器测量了各种编码器在解码过程中的速度。为了比较清楚，我们将chunkSAE的速度设置为1.0，并给出了其他编码器的速度比。在表1的第1行和第2行中，由于历史和未来上下文的冗余计算，区块SAE比SAE慢。在表1的第2行和第3行中，我们观察到状态重用块SAE比块SAE快1.5倍，这与第4.2节中的理论分析一致。除了更快的速度外，状态重用区块SAE在HKUST开发和测试集上的相对CER减少率分别比区块SAE高出1.53%和0.38%。由于具有更快的速度和更好的性能，我们在后续的实验中采用了状态重用块SAE。*荷兰*

## 5.4. 语境调查

在表2中，我们研究了我们的在线模型在不同历史、中心和未来上下文长度下的性能。首先，比较表2中的第2-4行，我们可以看到未来的上下文带来了更多的表2。香港科技大学不同情境配置之线上变形金刚ASR模式之CER。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 不。 | 模型 | *荷兰* | *数控* | *编号* | 开发 | 试验 |
| 1 | SAE+萨德 | – | – | – | 24.12 | 23.47 |
| 3 | 状态重用 | 64 | 64 | 0 | 29.97 | 28.41 |
| 4 | 块SAE | 0 | 64 | 64 | 24.94 | 24.10 |
| 5 | + | 64 | 64 | 64 | 24.45 | 23.65 |
| 6 MTA-SAD 64 32 24.67 24.05 | | | | | | |

表3。与香港科技大学公布的ASR模型比较。

|  |  |  |
| --- | --- | --- |
| 模型 | 大小 | 试验 |
| TDNN混合，无晶格MMI[25] | 1900万 | 23.69 |
| 离线自我注意校准器[26] | 3800万\* | 24.12 |
| 在线自我注意校准器[26] | 24米\* | 26.52 |
| 离线BLSTM CTC/att模型[6] | 112米 | 27.43 |
| 在线LC-BLSTM CTC/att模型[11] | 112米 | 27.84 |
| 基于在线变压器的CTC/att模型 | 3100万 | 23.66 |

\*

根据模型配置估计模型参数大小。

改进比历史背景，这表明未来背景对我们的在线模型的性能更为关键。其次，对比表2中的第5-7行，我们发现在减少状态重用延迟的同时，增加历史上下文的长度可以有效地保持识别的准确性。第三，比较表2中第7行和第8行，我们发现随着中心上下文长度的增加，CER降低。

最后，与表2第1行中的离线基线相比，我们最好的在线模型实现了23.65%的CER，640毫秒的延迟和0.18%的绝对CER降低。在表3中，我们还比较了基于变压器的在线CTC/注意模型和其他已发布的ASR模型。为了公平比较，表3中列出的在线E2E模型的潜伏期为320ms。这些模型在香港科技大学进行了速度扰动训练，但在线自我注意校准模型除外。

# 6结论

本文提出了基于Transformer的在线E2E-ASR模型，该模型由状态重用块SAE和基于MTA的SAD组成，并将提出的基于Transformer的在线E2E-ASR模型集成到CTC/attention-ASR体系结构中。与简单块SAE相比，状态重用块SAE的性能更好，所需的计算成本更低，因为它通过在以前的块中存储状态而具有更广泛的历史背景。与SAD算法相比，基于MTA的SAD算法对SAE输出进行了从左到右的单调截断，并对截断后的SAE输出进行关注，使其适用于在线识别。我们在香港科技大学评估了所提出的基于变压器的在线CTC/attention E2E模型，在320毫秒的延迟下达到了23.66%的CER，这比我们以前基于LSTM的在线E2E模型要好。在未来，我们计划采用师生学习的方式来进一步减少模型的延迟。

# 7参考文献

[1] A.Graves、S.Fernandez、F.Gomez和J.Schmidhu-´ber，“连接主义时间分类：用递归神经网络标记未分段序列数据”，《第23届机器学习国际会议论文集》，美国纽约，2006年，ICML&apos;06，第369-376页，ACM。

[2] J.Chorowski、D.Bahdanau、D.Serdyuk、K.Cho和Y.Bengio，“基于注意力的语音识别模型”，《第28届神经信息处理系统国际会议论文集》第1卷，美国马萨诸塞州剑桥，2015年，NIPS&apos;15，第577-585页，麻省理工学院出版社。

[3] 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”，2016年IEEE声学、语音和信号处理国际会议（ICASSP），2016年3月，第4960–4964页。

[4] D.Amodei，S.Ananthanarayanan，R.Anubhai，J.Bai，E.Battenberg，C.Case，J.Casper，B.Catanzaro，Q.Cheng，G.Chen，et al.，“深度语音2：英语和汉语的端到端语音识别”，国际机器学习会议，2016年，第173-182页。

[5] C.Chiu，T.N.Sainath，Y.Wu，R.Prabhavalkar，P.Nguyen，

Z、 Chen，A.Kannan，R.J.Weiss，K.Rao，E.Gonina，N.Jaitly，B.Li，J.Chorowski和M.Bacchiani，“序列到序列模型的最新语音识别”，2018年IEEE声学、语音和信号处理国际会议（ICASSP），2018年4月，第4774-4778页。

[6] 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12月。

[7] 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），加拿大温哥华，2017年7月，第518-529页，计算语言学协会。

[8] D.，K.Cho和Y.Bengio，“通过联合学习对齐和翻译实现神经机器翻译”，第三届学习表征国际会议，ICLR 2015，圣地亚哥，加利福尼亚州，美国，2015年5月7-9日，会议跟踪记录，2015年。

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

[10] 苗浩、郑国光、张培章、田良良、颜义明，“端到端语音识别之线上混合CTC/注意架构”，载于《语音识别研究进展》。Interspeech 2019，2019，第2623-2627页。

[11] H.Miao和G.Cheng，“流式关注”https://github.com/HaoranMiao/streaming-attention，2020年，GitHub存储库。

[12] A.Vaswani、N.Shazeer、N.Parmar、J.Uszkoreit、L.Jones、A.N.Gomez、L.Kaiser和I.Polosukhin，“注意力就是你所需要的一切”，摘自《第31届神经信息处理系统国际会议论文集》，美国，2017年，NIPS&apos;17，第6000–6010页，Curran Associates Inc。

[13] L.Dong，S.Xu和B.Xu，“语音转换器：用于语音识别的非连续序列对序列模型”，2018 IEEE声学、语音和信号处理国际会议（ICASSP）。IEEE，2018年，第5884-5888页。

[14] S.Karita、N.E.Y.Soplin、S.Watanabe、M.Delcroix、A.Ogawa和T.Nakatani，“改进基于变压器的

连接主义时间分类和语言模型集成的端到端语音识别。Interspeech 2019，2019，第1408-1412页。

[15] N.Pham，T.Nguyen，J.，M.Mller和A.Waibel，“用于端到端语音识别的非常深入的自我关注网络”，过程。Interspeech 2019，2019，第66-70页。

[16] S.Karita，N.Chen，T.Hayashi，T.Hori，H.Inaguma，Z.Jiang，

M、 Someki，N.Yalta，R.Yamamoto，X.Wang，S.Watanabe，T.Yoshimura和W.Zhang，“语音应用中变压器与rnn的比较研究”，2019年9月。

[17] 戴志伟、杨志伟、杨勇、卡博内尔、乐国伟和

R、 Salakhutdinov，“Transformer xl:超出固定长度上下文的注意力语言模型”，更正，第abs/1901.028602019卷。

[18] C.Chiu和C.Raffel，“单调的分块注意力”，第六届学习表征国际会议，ICLR 2018，加拿大不列颠哥伦比亚省温哥华，2018年4月30日至5月3日，会议记录，2018。

[19] K.He，X.Zhang，S.Ren和J.Sun，“图像识别的深度剩余学习”，2016年IEEE计算机视觉和模式识别会议（CVPR），2016年6月，第770-778页。

[20] L.J.Ba、J.R.Kiros和G.E.Hinton，“层标准化”，更正，第abs/1607.06450卷，2016年。

[21]刘彦，冯平，杨彦，蔡瑞利，黄国胜，格拉芙，“香港科技大学/mts：一个非常大规模的普通话电话语音语料库”，载于《国际汉语口语处理会议》，2006年。

[22]S.Watanabe，T.Hori，S.Karita，T.Hayashi，J.Nishitoba，

Y、 Unno，N.E.Y.Soplin，J.Heymann，M.Wiesner和N.Chen，“Espnet:端到端语音处理工具包”，2018年。

[23]N.Srivastava、G.Hinton、A.Krizhevsky、I.Sutskever和R.Salakhutdinov，“辍学：防止神经网络过度拟合的简单方法”，J.Mach。学习。《研究》，第15卷，第1期，1929-1958页，2014年1月。

[24]C.Szegedy，V.Vanhoucke，S.Ioffe，J.Shlens和Z.Wojna，“重新思考计算机视觉的初始架构”，IEEE计算机视觉和模式识别会议记录，2016年。

[25]D.Povey、V.Peddinti、D.Galvez、P.Ghahremani、V.Manohar、X.Na、Y.Wang和S.Khudanpur，“基于无晶格mmi的asr纯序列训练神经网络”，INTERSPEECH，2016年。

[26]L.Dong，F.Wang，B.Xu，“自我注意校准器：使用自我注意网络和区块跳跃的asr延迟控制端到端模型”，ICASSP 2019-2019 IEEE声学、语音和信号处理国际会议（ICASSP），2019年5月，第5656-5660页。