三菱电气研究实验室http://www.merl.com

**用于端到端语音识别的混合CTC/注意结构**

渡边，S.；Hori，T.；Kim，S.；好时，J.R.；Hayashi，T。

                                                                    TR2017-190 2017年10月

**摘要**

传统的基于隐马尔可夫模型（HMM）/深度神经网络（DNN）的自动语音识别（ASR）是一个非常复杂的系统，由声学、词汇和语言模型等模块组成。它还需要语言资源，如发音词典、标记化和语音上下文依赖树。另一方面，端到端ASR已经成为一种流行的替代方法，通过用单一的深度网络结构表示复杂的模块，并用数据驱动的学习方法代替语言资源的使用，大大简化了传统ASR系统的建模过程。ASR的端到端结构主要有两种：基于注意的方法利用注意机制实现声帧与识别符号之间的对齐；连接主义时间分类（CTC）利用马尔可夫假设通过动态规划有效地解决序列问题。本文提出了混合CTC/注意端到端ASR，有效地利用了这两种结构在训练和译码方面的优势。在训练过程中，我们采用多目标学习框架来提高鲁棒性和快速收敛性。在解码过程中，我们将基于注意和CTC的分数结合在一个单程波束搜索算法中进行联合解码，以进一步消除不规则对齐。以英语（WSJ和CHiME-4）任务为例的实验证明了所提出的多目标学习方法在CTC和基于注意的编解码基线上的有效性。此外，将该方法应用于两个大规模ASR基准测试（自然日语和汉语普通话），基于多目标学习和无需语言资源的联合解码的优点，其性能与传统DNN/HMM ASR系统相当。

*IEEE信号处理选刊*

不得出于任何商业目的复制或复制本作品的全部或部分内容。出于非营利的教育和研究目的，允许复制全部或部分内容而不支付费用，前提是所有此类全部或部分内容包括以下内容：经三菱电气研究实验室有限公司许可的此类复制通知；对作者和个人贡献的确认以及版权声明的所有适用部分。为任何其他目的复制、复制或再版需要许可证，并向三菱电气研究实验室有限公司支付费用。保留所有权利。

版权所有c Mitsubishi Electric Research Laboratories，Inc.，2017 201 Broadway，Cambridge，Massachusetts 02139

# 用于端到端语音识别的混合CTC/注意结构

渡边真治，IEEE高级成员，Takaaki Hori，IEEE高级成员，Suyoun Kim，学生

*IEEE成员，*John R.Hershey，IEEE高级成员，和Tomoki Hayashi，IEEE学生成员

*摘要*-传统的基于隐马尔可夫模型（HMM）/深度神经网络（DNN）的自动语音识别（ASR）是一个非常复杂的系统，由声学、词汇和语言模型等模块组成。它还需要语言资源，如发音词典、标记化和语音上下文依赖树。另一方面，端到端ASR已经成为一种流行的替代方法，通过用单一的深度网络结构表示复杂的模块，并用数据驱动的学习方法代替语言资源的使用，大大简化了传统ASR系统的建模过程。ASR的端到端结构主要有两种：基于注意的方法利用注意机制实现声帧与识别符号之间的对齐；连接主义时间分类（CTC）利用马尔可夫假设通过动态规划有效地解决序列问题。本文提出了混合CTC/注意端到端ASR，有效地利用了这两种结构在训练和译码方面的优势。在训练过程中，我们采用多目标学习框架来提高鲁棒性和快速收敛性。在解码过程中，我们将基于注意和CTC的分数结合在一个单程波束搜索算法中进行联合解码，以进一步消除不规则对齐。通过对英语（WSJ和CHiME-4）任务的实验，证明了所提出的多目标学习方法在CTC和基于注意的编解码基线上的有效性。此外，将该方法应用于两个大规模ASR基准测试（自然日语和汉语普通话），基于多目标学习和无需语言资源的联合解码的优点，其性能与传统DNN/HMM ASR系统相当。

*索引项*-自动语音识别，端到端，连接主义时间分类，注意机制，混合CTC/注意。

## 一、 导言

|  |
| --- |
| A |

自动语音识别（ASR）是实现自然人机界面的关键技术。它已经成为一套成熟的技术，得到了广泛的应用，在语音搜索等界面应用中取得了巨大的成功。基于概率噪声信道模型[1]，一个典型的ASR系统被分解为几个模块，包括声学、词汇和语言模型。在过去的十年里

S、 Watanabe、T.Hori和J.R.Hershey与美国三菱电气研究实验室（MERL）合作，电子邮件：

{渡边、托里、好时}@merl.com网站.

S、 Kim供职于美国卡内基梅隆大学（CMU），电子邮件：suy-oun@cmu.edu。

T、 Hayashi在日本名古屋大学工作，电子邮件：hayashi.tomoki先生@g、 sp.m.公司。是名古屋吗-美国。ac.jp公司.

2017年4月1日收到稿件；修订？？？, 20??.

声学和语言模型是由机器学习技术驱动的，称为深度学习[2]。然而，当前的系统严重依赖于围绕传统技术开发的复杂遗留体系结构的支架。他们提出了以下问题，我们可以设法消除。

•逐步细化：构建准确的模块需要许多特定于模块的过程。例如，当我们从头开始建立一个声学模型时，首先要建立一个隐马尔可夫模型（HMM）和高斯混合模型（GMMs）来获得绑定状态的HMM结构和语音对齐，然后才能训练深层神经网络（DNNs）。

•语言信息：为了很好地分解声学和语言模型，我们需要有一个词典模型，它通常基于一个手工制作的发音词典来将单词映射到音素序列。由于音素是使用语言知识设计的，因此它们容易受到人为错误的影响，而完全由数据驱动的系统可能会避免这些错误。最后，有些语言没有显式的单词边界，需要标记化模块

[3], [4].

•条件独立假设：当前的ASR系统在上述分解过程中经常使用条件独立假设（特别是马尔可夫假设），并使用GMM、DNN和-gram模型。真实世界的数据不一定遵循这样的假设，从而导致模型的错误说明。•复杂解码：推理/解码必须通过集成所有模块来执行。虽然这种集成通常由有限状态传感器有效地处理，但优化传感器的构造和实现非常复杂[5]，[6]。*不*

•优化中的不一致性：上述模块分别以不同的目标进行优化，这可能会导致优化中的不一致性，其中每个模块没有经过训练以匹配其他模块。

因此，对于非专家来说，为新的应用程序，特别是新的语言，使用/开发ASR系统是相当困难的。

端到端ASR的目标是将上述基于模块的体系结构简化为深度学习框架内的单一网络体系结构，以解决上述问题。ASR的端到端架构有两种主要类型：基于注意的方法使用注意机制在声学帧和识别符号之间执行对齐，而连接主义时间分类（CTC）使用马尔可夫假设通过动态规划有效地解决序列问题[7]，[8]。

所有的ASR模型都旨在阐明给定语音特征序列的单词序列的后验分布（W | X）。端到端方法直接实现这一点，而传统模型将（W | X）分解为语言模型（W）和声学模型似然（X | W）等模块，前者可以在纯语言数据上进行训练，后者可以在具有相应语言标签的声学数据上进行训练。端到端ASR方法通常只依赖于成对的声学和语言数据。如果没有额外的语言数据，他们可能会遇到数据稀疏或词汇量不足的问题。为了提高泛化能力，并处理词汇表外的问题，在语言输出序列中通常使用字母表示法而不是单词表示法，我们在下面的描述中采用了这种方法。*第W型十第第第*

基于注意的端到端方法通过使用编码器-解码器架构将语音特征序列映射到文本，从而解决了ASR问题。解码器网络使用注意机制来寻找输出序列的每个元素与声学编码器网络针对声学输入的每个帧生成的隐藏状态之间的对准[7]、[9]、[10]、[11]。在每个输出位置，解码器网络在每个输入时间计算其隐藏状态和编码器网络的状态之间的匹配分数，以形成时间对齐分布，然后使用该分布来提取相应编码器状态的平均值。

这种基本的时间注意机制太灵活了，因为它允许非常不连续的排列。这对于输入和输出词序不同的机器翻译等应用来说可能是很好的[12]，[13]。然而，在语音识别中，特征输入和相应的字母输出通常以相同的顺序进行，只有很小的词内偏差（例如，单词“iron”，它将“r”和“o”的声音转置）。另一个问题是，ASR中的输入和输出序列可能有非常不同的长度，并且它们在不同的情况下有很大的差异，这取决于说话速率和书写系统，使得跟踪对齐更加困难。

然而，注意机制的一个优点是不需要任何有条件的独立性假设，并且可以解决上面提到的所有问题。尽管[7]，[14]中使用了各种机制部分地解决了基于注意的机制的对齐问题，但是这里我们提出了更严格的约束，使用基于CTC的对齐来指导训练。

CTC允许使用动态规划[15]，[8]有效地计算严格单调对齐，尽管它需要单独的语言模型和基于图形的解码[16]，除了在大量训练数据的情况下[17]，[18]。我们建议在基于CTC/注意的混合系统中利用约束CTC对齐。在训练过程中，我们提出了一种多目标学习方法，将一个CTC目标作为正则化附加到一个基于注意的编码网络上[19]。这大大减少了不规则排列的话语数量，而无需任何启发式搜索技术。在解码过程中，我们提出了一种联合解码方法，该方法将基于注意和CTC的分数结合在一个重排序/单程波束搜索算法中，以消除不规则对齐[20]。

该方法首次应用于英语语音阅读任务中，主要验证了混合CTC/注意结构的多目标学习的有效性。然后，该方法进一步应用于日语和汉语的ASR任务，这些任务除了需要语音词典外，还需要额外的语言资源，包括词法分析器[3]或分词[21]，以在传统的DNN/HMM ASR中提供准确的词汇和语言模型。令人惊讶的是，当同时使用多目标学习和联合译码时，在不使用上述语言资源的情况下，该方法取得了与几种最先进的HMM/DNN-ASR系统相当的性能，并且在某些情况下优于它们。

本文总结了我们以前对混合CTC/注意体系结构的研究[19]，[20]，分别侧重于它的训练和解码功能。本文扩展了[19]和[20]，提供了从传统HMM/DNN系统到当前端到端ASR系统的更详细的公式（第二节），用于训练和解码的混合CTC/注意体系结构的一致公式（第三节），以及更多的实验讨论（第四节）。

二。从HMM/DNN到端到端ASR

     本节提供了常规的

HMM/DNN ASR和CTC或基于注意的端到端ASR。该公式旨在阐明概率因子分解和条件独立性假设（Markov假设），这是描述这三种方法的重要性质。

### A、 嗯/DNN

ASR处理从一个长语音特征序列={xt∈R | t=1，···，t}到一个长单词序列={wn∈V | n=1，···，n}的序列映射。这里，是帧处的一维语音特征向量（例如，log Mel filterbank），是词汇表中位置处的单词V。*T型十D级不W型***十***t型D级t型wn公司不*

ASR是用贝叶斯决策理论在数学上表示的，其中最可能的词序列是在所有可能的词序列V\*中估计的，如下所示：*W型*ˆ

*W型*ˆ=arg max p（W | X）。(1)

*W型*∈V\*

因此，ASR的主要问题是如何获得后验分布（W | X）。*第*

当前ASR的主流是基于混合HMM/DNN[22]，它使用贝叶斯定理并引入HMM状态序列={st∈{1，···，J}| t=1，···，t}，将（W | X）分解为以下三个分布：*S码第*

arg最大p（宽X）

*W型*∈V\*

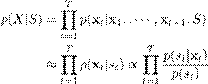
|  |  |
| --- | --- |
| =arg最大值（X | S，W）p（S | W）p（W）十*第*  *W型*∈V\*  *S码* | (2) |
| ≈arg max（X | S）p（S | W）p（W）。十*第* | (3) |

*W型*∈V\*

*S码*

这三个因素（X | S）、（S | W）和（W）分别是声学模型、词汇模型和语言模型。式（3）通过条件独立性假设（即（X | S，W）≈p（X | S））获得，这是简化声学模型相关性的合理假设。*第第第第*

*1)声学模型第*（X | S）（X | S）通过使用概率链规则和条件独立性假设进一步分解，如下所示：*:第*

(4)

*,* (5)

其中，用强大的DNN分类器通过使用所谓的伪似然技巧计算的帧态后验分布（st | xt）/p（st）替换帧态似然函数（xt | st）[22]。公式（5）中的条件独立性假设通常被认为太强，因为它不考虑任何输入和隐藏状态上下文。因此，具有长上下文特征的dnn或递归神经网络常被用来缓解这一问题。为了训练逐帧后验，我们还需要提供一个逐帧状态对齐，作为一个目标，这通常是由HMM/GMM系统提供的。*第第圣*

*2)词汇模型第*（S | W）（S | W）也通过使用概率链规则和条件独立性假设（一阶马尔可夫假设）进行因子分解，如下所示：*:第*

*T型*

*第*（西南）=是的*第*（st | s1，…，st−1，W）(6)

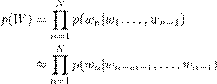
*t型*=1吨

                                            ≈（st | st−1，W）。(7)是的*第*

*t型*=1

这个概率由给定的HMM状态转换表示。从HMM状态到HMM状态的转换是通过音素表示使用发音字典来确定的。*W型W型*

*3)语言模型第*（W） 类似地，（W）通过使用概率链规则和条件独立性假设（（m−1）阶马尔可夫假设）作为gram模型进行因子分解，即。，*:第米*

(8)

*.* (9)

尽管递归神经网络语言模型（RNNLM）可以避免这种条件独立性假设问题[23]，但它使译码变得复杂，并且RNNLM通常与基于重排序技术的-gram语言模型相结合。*米*

因此，传统的HMM/DNN系统通过使用因子分解和条件独立性假设，使得式（1）中的ASR问题可行，代价是第1节中讨论的五个问题。

### B、 连接主义时间分类（CTC）

CTC公式也遵循贝叶斯决策理论（公式（1））。注意，CTC公式使用长度字母序列={cl∈U | l=1，·····，l}，以及一组不同的字母U。此外，CTC还使用“空白符号”，明确表示字母边界来处理字母符号的重复。使用空白符号，将增强的字母序列定义为*我C级C级*0

|  |  |
| --- | --- |
| *C级*0 ={<b><b><b>···，cL，<b>}*，c级*1*,，c级*2*,,* | (10) |
| ={c0l∈U∪{<b>}| l=1，····，2L+1}。 | (11) |

在中，当是奇数时，增广字母始终为空“<b>”，而当是偶数时，增广字母始终为字母*C级*0*c级*0*我我我*

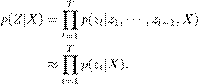
与第II-A节类似，通过引入带有附加空白符号的逐帧字母序列，={zt∈U∪{<b>}| t=1，···，t}，后验分布（C | X）被分解如下：*Z轴*[[1]](" \l "_ftn1" \o ")*第*

|  |  |
| --- | --- |
| *第*（C | X）=（C | Z，X）p（Z | X）十*第*  *Z轴* | (12) |
| ≈（C | Z）p（Z | X）。十*第* | (13) |

*Z轴*

与式（3）类似，CTC使用条件独立性假设来获得式（13）（即，（C | Z，X）≈p（C | Z）），这是简化CTC声学模型（Z | X）和CTC字母模型（C | Z）依赖性的合理假设。*第第第*

*1） CTC声学模型：*与第II-A1节类似，（Z | X）通过使用概率链规则和条件独立性假设进一步分解，如下所示：*第*

(14)

(15)

逐帧后验分布（zt | X）以所有输入为条件，使用双向长短时记忆（BLSTM）建模非常简单[25]，*第十*

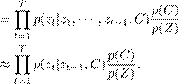
[26]:

|  |  |
| --- | --- |
| *第*（zt | X）=软最大值（LinB（ht））， | (16) |
| **小时***t型*=BLSTM（X）。*t型* | (17) |

Softmax（·）是一个sofmax激活函数，LinB（·）是一个线性层，用于将隐藏向量转换为（| U |+1）维向量（+1表示CTC中引入的空白符号），具有可学习的矩阵和偏置向量参数。BLSTM（·）接受完整的输入序列和输出隐藏向量。**小时***t型t型t型*

*2） CTC信函型号：第*（Z | X）通过使用贝叶斯规则、概率链规则和条件独立性假设重写，如下所示：

(18)

(19)

(20)

其中（zt | zt−1，C）、（C）和（Z）分别是状态转移概率、基于字母的语言模型和状态先验概率。CTC有一个基于字母的语言模型（C）*第第第第*，通过使用字母到单词的有限状态转换器，我们还可以在解码过程中在CTC中加入基于单词的语言模型[16]。在最初的CTC公式中没有引入[15]。然而，这一因素的理论依据和实验有效性如[27]所示。

状态转移概率（zt | zt−1，C）*第*，用式（11）中的增广字母表示，如下所示：*第*（zt | zt-）1*，C级*

         1 zt=c0l和*zztt公司*−−11 =cc00l尽一切可能*我，我*

=l负极1

*.*

对于所有可能的偶数*兹特*−1*我*

0   否则

(21)

在式（21）中，第一种情况表示自跃迁，而第二种情况表示状态跃迁。第三种情况是通过跳过“blank”从letter进行特殊的状态转换，其中*我*是一个偶数，并且总是表示一个字母，如等式（10）所示。请注意，在实现中，这些转换值不会随时间而标准化*兹特*（即，不是概率值）[16]，[28]，类似于HMM状态转换实现[29]。

利用式（21）中的状态转移形式，很明显，CTC具有单调对准特性，即：。，

              什么时候？*,* 则=式中≥m.（22）*兹特c级*0*我我*

这个特性是ASR的一个重要约束，因为ASR序列到序列的映射必须遵循与机器翻译不同的单调对齐。HMM/DNN也满足这种单调对齐特性。

*3） 目标：*用等式。（15） （20）后面的（C | X）最后表示为*第*

*.* (23)

,*第*反恐委员会（立方厘米）

尽管等式（23）必须处理所有可能的求和，但由于马尔可夫性，它可以通过使用动态规划（维特比/前向-后向算法）有效地计算。我们还定义了CTC目标函数（C | X），在后面的公式中使用，它通常不包括（C）/p（Z）。*Z轴第*反恐委员会*第*

CTC公式与HMM/DNN的公式相似，只是它将Bayes规则应用于（C | Z）而不是（W | X）。因此，CTC具有三个类似于HMM/DNN情况的分布分量，即逐帧后验分布（zt | X）、转移概率（zt | zt−1，C）和（基于字母的）语言模型（C）。CTC还使用了一些条件独立性假设（马尔可夫假设），并没有充分利用端到端ASR的优点，如第一节所述。但是，与HMM/DNN系统相比，具有字符输出表示的CTC仍然具有端到端的优点，即它不需要语音词典，并且省略了HMM/GMM构造步骤。*第第第第第*

### C、 注意机制

与HMM/DNN和CTC方法相比，基于注意的方法不做任何条件独立性假设，而是根据概率链规则直接估计后验概率（C | X），如下所示：*第*

#### lp（C | X）=p（cl | C，···，cl，X），（24）是的1−1

*我*=1



                                           |{z}

,*第*收件人（立方厘米）

其中（C | X）是基于注意的目标函数。（cl | c1，···，cl−1，X）通过*第*收件人*第*

**小时***t型*=编码器（十） 哦，(25)

(

内容注意（ql−1，ht）

                                                 *中高音*=,(26)

位置注意

*T型*

**右侧***我*=Xa*书信电报t型,* (27)

*t型*=1

*第*（cl | c1，···，cl−1，X）=解码器（rql，l−1，cl−1）。(28)

方程式。（25）和（28）分别是编码器和解码器网络。在等式（26）中是注意权重，并且表示基于隐藏向量的加权和的每个输出的隐藏向量的软对齐，以形成等式（27）中的字母式隐藏向量。式（26）中的ContentAttention（·）和LocationAttention（·）分别基于有卷积特征和无卷积特征的基于内容的注意机制[9]。我们将在下面更详细地解释每个模块。*中高音***人力资源***t型氯我*

*1） 编码器网络：*式（25）将输入特征向量转换为逐帧隐藏向量，BLSTM通常用作编码器网络，即。，*十***小时***t型*

                                  编码器（X），BLSTM（X）。(29)*t型*

注意，输出通常是二次采样的，以降低编码器网络的计算复杂度[9]，[10]。

|  |  |  |
| --- | --- | --- |
| *2） 基于内容的注意机制：*注意（·）表示如下： | 等式。 | (26), |
| *英语教学*=g>tanh（Lin（ql−1）+LinB（ht））， |  | (30) |

*中高音*=Softmax软件*.* (31)

**克**是可学习的向量参数。是一个*T型*-维向量，即双曲正切激活函数，Lin（·）是一个线性层，具有可学习的矩阵参数，但没有偏置矢量参数。

*3)位置感知注意机制：*基于内容的注意机制被扩展到处理卷积（位置感知注意）。当我们使用=**一***我*−1

，式（26）中的位置注意（·）表示如下：

{f} =1=K\*a−1（32）=g>tanh（Lin（ql−1）+Lin（ht）+LinB（ft）），（33）=*t型Tt公司我,   英语教学中高音*Softmax软件*.* (34)

\*表示沿输入特征轴的一维卷积，带有卷积参数，以产生*t型***K公司***T型*特征。

*4)解码器网络：*等式（28）中的解码器网络是除逐字母隐藏向量外，类似于RNNLM的、以先前输出和隐藏向量为条件的另一循环网络。我们使用以下单向LSTM：*氯*−1**qr码***我*−1*我*

                      解码器（·），Softmax（LinB（LSTM（·）））。(35)*我*

LSTM（·）是一个单向LSTM单元，输出隐藏向量如下：*我***问***我*

**问***我*=LSTM公司*我*（rql，l−1，cl−1）。(36)

此LSTM接受字母隐藏向量的串联向量和前一输出的一个热表示作为输入。**右***我氯*−1

*5)目标：*注意模型的训练目标由式（24）中的后验序列（C | X）近似计算如下：*第*收件人

#### lp（C | X）p（cl | C，···，cl，X），p（C | X），（37）L=1收件人约Y∗1∗−1∗收件人

哪里*c级*∗*我*是前面人物的基本事实。这是基于注意力的方法的有力假设，即等式（37）对应于基于简单多类分类的字母目标组合，每个输出中有条件的基本事实历史，*我*，并没有充分考虑序列级目标，如[10]所指出的。

总之，基于注意的ASR并没有明确地将每个模块分开，并且可能处理第一节中提出的所有五个问题。它隐含地将声学、词汇和语言模型结合为编码器、注意和解码器网络，这些网络可以作为一个单独的网络进行联合训练。但是，与具有合理单调对齐特性的HMM/DNN和CTC相比，如中所述

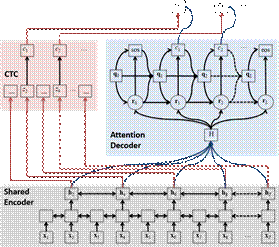


图1：混合CTC/基于注意力的端到端架构。共享编码器由CTC和注意模型目标同时训练。共享编码器将输入序列{xt···xt}转换为高级特征={ht···ht}，注意力解码器生成字母序列{c1···cL}。*小时*

第II-B2节，注意机制并不维持这种约束。对齐由所有帧的加权和表示，如等式（27）所示，通常提供不规则对齐。本文的重点是通过提出混合的CTC/attention体系结构来解决这个问题。

## 三、 混合CTC/注意

本节解释了我们的CTC/注意力体系结构，它在ASR的训练和解码步骤中利用了CTC和注意力的优点。

### A、 多目标学习

所提出的训练方法使用CTC目标函数作为辅助任务，在多目标学习（MOL）框架内训练注意模型编码器[19]。图1说明了框架的总体架构，其中同一BLSTM与CTC和注意力编码器网络（即Eqs）共享。（17） 和（29）。与单一注意模型不同，CTC的前向-后向算法可以在训练过程中实现语音序列和标签序列的单调对齐。也就是说，CTC中的前向-后向算法有助于加速估计期望对齐的过程，而不是仅仅依赖于数据驱动的注意方法来估计长序列中的期望对齐。最大化目标是CTC和注意力目标的对数线性组合，即（C | X）*第*反恐委员会在式（23）和式（37）中：

         我摩尔*,* (38)

其中可调参数满足0≤λ≤1。公式（38）的另一个优点是，注意目标是一个近似的字母目标，如第II-C5节所述，而CTC目标是一个序列级目标。因此，这种多目标学习除了有助于估计所需比对的过程之外，还可以减轻序列级CTC目标的这种近似。文[30]中还介绍了端到端ASR中的多目标学习策略，它结合了分段条件随机场（CRF）和CTC。*λ*

### B、 联合译码

我们的混合CTC/基于注意的端到端语音识别的推理步骤是通过标签同步解码和类似于传统基于注意的ASR的波束搜索来完成的。然而，我们考虑了CTC概率，找到了一个更适合输入语音的假设，如图1所示。在下文中，我们描述了一般的基于注意的解码和传统技术，以减轻对齐问题。然后，我们提出了一种混合CTC/注意结构的联合译码方法。

*1） 一般来说，基于注意的解码：*端到端语音识别推理通常被定义为在给定语音输入的情况下，寻找最可能的字母序列的问题，即。*C级*ˆ*十*

*C级*ˆ=arg max logp（C | X）。(39)

*C级*∈U\*

在基于注意的ASR中，（C | X）由公式（24）计算，并通过波束搜索技术找到。*第C级*ˆ

设Ωl是长度的一组部分假设。在波束搜索开始时，Ω0仅包含一个带有起始符号<sos>的假设。对于=1到，通过添加可能的单个字母扩展Ωl−1中的每个部分假设，新假设存储在Ωl中，其中是要搜索的假设的最大长度。每个新假设的得分在对数域中计算为*我我我*最大值*我*最大值

*α*（h，X）=α（g，X）+logp（c | gl−1，X），(40)

其中是Ωl−1中的部分假设，是附加的字母，是新的假设，以便=g·c。如果是表示序列结束的特殊符号，<eos>，则添加到Ω，而不是Ωl，其中Ω表示一组完整的假设。最后，通过*克c级克小时小时c级小时*ˆˆ*C级*ˆ

*C级*ˆ=argmaxα（h，X）。(41)

*小时*∈Ωˆ

在波束搜索过程中，为了提高搜索效率，允许Ωl只保留有限数量的分数较高的假设。

然而，基于注意的ASR可能容易包括删除和插入错误（参见图3和相关讨论），因为它具有灵活的对齐特性，可以关注编码器状态序列的任何部分以预测下一个标签，如第II-C节所述。由于注意是由解码器网络产生的，它可能会过早地预测序列结束标签，即使它没有注意到所有的编码器帧，使假设太短。另一方面，通过关注与之前关注的部分相同的部分，它可以高概率地预测下一个标签。在这种情况下，假设变得非常长，并且包含相同标签序列的重复。

*2） 传统解码技术：*为了缓解对齐问题，通常使用长度惩罚项来控制要选择的假设长度[9]，[31]。利用长度惩罚，等式（39）中的解码目标改变为

*C级*ˆ=arg max{logp（C | X）+γ| C |}，(42)

*C级*其中| C |是序列的长度，是一个可调参数。然而，即使经过仔细调整，也很难完全排除太长或太短的假设。用最小值和最大值在一定程度上控制假设长度也是有效的，其中最小值和最大值与输入语音长度的比值是固定的。然而，由于与输入语音相比，存在异常长或异常短的转录本，因此很难平衡保存这些异常转录本和防止不相关长度的假设。*C级γγ*

另一种方法是最近在[14]中提出的覆盖项，该覆盖项包含在本文的解码目标中

式（42）asˆ=arg max{logp（C | X）+γ| C |+η·coverage（C | X）}，*C级*

*C级*∈U\*

(43)

其中保险期限的计算方法为

                      新闻报道 *.* (44)

*η*是可调参数。覆盖率项表示接收到的累计注意力大于的帧数。因此，当第一次密切关注某些帧时，它增加，但当再次关注相同帧时，它不增加。这个属性对于避免同一个标签序列在一个假设中循环是有效的。然而，覆盖项没有明确的机制来避免序列末端标签的过早预测，这使得假设太短，导致了大量的删除错误。此外，仍然很难获得、和可选的最小/最大长度的公共参数设置，以便它们适用于来自不同任务的任何语音数据。*ττγητ*

*3） 联合解码：*我们的混合CTC/注意方法在推理步骤和训练步骤中结合了CTC和基于注意的序列概率。假设式（23）中的（C | X）和式（24）中的（C | X）分别是CTC和注意模型给出的序列概率。解码目标的定义类似于等式（38），如下所示：*第*反恐委员会*第*收件人

*C级*ˆ=arg max{λlogp（C | X）+（1−λ）logp（C | X）}。反恐委员会收件人

*C级*∈U\*

(45)

CTC概率强制执行单调对齐，不允许第II-B2节中讨论的相同帧的大跳跃或循环。此外，它还可以避免覆盖项无法处理的序列结束标签的过早预测。因此，可以选择具有更好对齐的假设并且排除不相关的假设，而不依赖覆盖项、长度惩罚或最小/最大长度。

在波束搜索过程中，解码器需要使用等式（40）计算每个部分假设的分数。然而，在波束搜索中，将CTC和基于注意的分数结合起来是非常重要的，因为注意解码器同步地执行它的输出标签，而CTC同步地执行它的帧。为了在假设得分中加入CTC概率，我们提出了两种方法。

*重新排序：*第一种方法是两遍方法，其中第一遍使用波束搜索获得一组完整的假设，其中仅考虑基于注意的序列概率。第二步使用CTC和注意概率重新审视完整的假设，其中CTC概率通过CTC的前向算法获得[15]。重新扫描通过根据

*C级*ˆ=argmax{λα反恐委员会（h，X）+（1−λ）α收件人（h，X）}，(46)

*小时*∈Ωˆ

哪里

(

*α*反恐委员会（h，X），logp（h | X）反恐委员会

*.* (47)

*α*收件人（h，X），logp（h | X）收件人

*单程解码：*第二种方法是单程解码，利用CTC和注意模型计算每个部分假设的概率。这里，我们利用CTC前缀概率[32]，定义为所有标签序列的累积概率，这些标签序列的前缀为：*小时*

*第*反恐委员会（高，…| X）=十*第*反恐委员会（h·ν| X），(48)

*ν*∈（U∪{<eos>}）+定义CTC得分为（h，X），logp（h，…| X），（49）*α*反恐委员会反恐委员会

其中表示除空字符串以外的所有可能的标签序列。CTC分数不能像式（40）中那样递归地获得，但是可以通过保持每个部分假设的输入帧上的前向概率来有效地计算。然后它与（h，X）结合使用。*να*收件人*λ*

一次译码的波束搜索算法如算法1所示。Ωl和Ω在算法的第2行和第3行中初始化，这两行实现为分别接受部分长度假设和完全假设的队列。在第4–25行中，Ωl−1中的每个部分假设由标签集U中的每个标签扩展。每个扩展的假设在第11行中打分，其中CTC和基于注意力的分数由（）和（）获得。之后，如果=<eos>，假设假设假设是完整的，并存储在第13行的Ω中。如果6=<eos>，则存储在第15行的Ωl中，其中第16行检查了Ωl中的假设数。如果数值超过波束宽度，则最差分数为Ωl的假设，即。，ˆ*我克c级小时α*反恐委员会*α*收件人*c级小时*ˆ*c级小时*

*小时*最差的=arg minα（h，X），∈Ωl*小时*

通过第17行中的REMOVEWORST（）从Ωl中删除。

|  |
| --- |
|  |
|  |  |

我们可以选择性地应用端点检测技术，通过在算法1联合CTC/注意单通解码之前停止波束搜索来减少计算量*我*

|  |  |
| --- | --- |
| 1： 过程ONEPASSBEAMSEARCH（，）*十我*最大值  2： Ω0←{<sos>} | |
| 三： | Ωˆ ← ∅ |
| 4: | 对于=1…Ldo*我*最大值 |
| 5: | Ωl←∅ |
| 6: | 当Ωl−16=∅do时 |
| 7: | *克*←磁头（Ωl-）1 |
| 8: | 出列（Ωl−1） |
| 9: | 对于每个∈U∪{<eos>}do*c级* |
| 10: | *小时*←克·c |
| 11: | *α*（h） ←λα（h，X）+（1−λ）α（h，X）反恐委员会收件人 |
| 12: | 如果=<eos>那么*c级* |
| 13: | 排队（Ω）ˆ*，小时* |
| 14: | 其他的 |
| 15: | 排队（Ωl，h） |
| 16: | 如果|Ωl |>则*波束宽度* |
| 17: | 移除最坏值（Ωl） |
| 18: | 结束if |
| 19: | 结束if |
| 20: | 结束 |
| 21: | 结束时 |
| 22: | 如果ENDDETECT（Ωˆ，l）=真，则 |
| 23: | 断开回路出口*.* |
| 24: | 结束if |
| 25: | 结束 |
| 26:返回argmaxC∈Ωˆ（C）*α*  27:结束程序 | |

达到。如果随着将来分数的增加，几乎不可能找到分数更高的完整假设，则第22行中的函数ENDDETECT（Ω）返回true。在我们的实现中，如果*我*最大值ˆ*，左我*

*米*−1 "( ) #

#### X−最大α（h0，X）<D=M，最大α（h，X）结束

*小时*∈Ω：ˆ| h |=l−m h0∈Ωˆ

*米*=0

(50)

其中和是预先确定的阈值。*D级*结束*米*

如果最近完成的假设的分数与解码过程中到目前为止所有完成的假设的最佳分数相比都足够小，则该等式变为真。在公式（50）的总和中，第一个最大值对应于最近生成的完整假设中的最佳分数，其长度| h |为−m，其中=0，…，m−1（例如，=3）。第二个最大值对应于所有完整假设中的最佳分数。如果这些最大分数之间的差值小于阈值（例如，=−10），则Iverson括号[·]返回1，否则返回0。因此，如果所有的差异都小于阈值，那么求和的结果就是。*我米米*Ωˆ*D级*结束*D级*结束*米*

在第11行中，计算每个部分假设的CTC和注意模型分数。注意力得分很容易获得，方法与式（40）相同，而CTC得分需要一个改进的前向算法，同步计算它。算法执行函数（h，X）*α*反恐委员会，如算法2所示。Let和算法2 CTC标签序列得分

|  |  |  |  |
| --- | --- | --- | --- |
| 1： 函数（）*α*反恐委员会*h、 十*  2： ←h。分为最后一个标签和其他标签*g、 c级小时c级克* | | | |
| 三： | 如果=<eos>那么*c级* |  | |
| 4: | 返回 |  | |
| 5: | 其他的 |  | |
| 6: | 1 0 | 如果=<sos>否则*克* | |
| 7:  8:  9: | for=2…不做*t型* | |  |
| 10: | 0 | | 如果last（g）=c，则 |
| 11:  12:  13: | Ψ←Ψ+Φ·p（zt=c | X） | |  |
| 14: | 结束 | |  |
| 15: | 返回对数（Ψ） | |  |
| 16： 结束if  17： 结束函数 | | | |
|  |  |  |  |

假设的前向概率，在时间范围1…t内，上标（n）和（b）*小时*表示所有CTC路径分别以非空或空符号结束的不同情况。在开始波束搜索之前，对和进行初始化*t型*=1，…，T等于

<sos>）=0，（51）

*t型*

<sos>）=（<sos>）·p（zτ=<b>| X），（52）=1是的*γτ*（b）−)1*τ*

我们假设<sos>)=1，<b>是空白符号。注意，由于编码器的子采样技术，时间索引和输入长度可能与输入语句的时间索引和输入长度不同[24]，[10]。*t型T型十*

在算法2中，假设首先被分割成最后一个标签，其余的在第2行。如果是<eos>，则返回前向概率的对数，假设第4行是一个完整的假设。前向概率由下式给出*小时c级克c级小时小时*

(53)

根据和的定义。如果*c级*不是

|  |
| --- |
|  |
|  |  |

<eos>，它计算前向概率和前缀概率，Ψ=p（h，…| X），假设反恐委员会*小时*不是一个完整的假设。第6-14行描述了这些概率的初始化和递归步骤。在这个函数中，我们假设每当我们计算概率时Ψ，前向概率和已经通过波束搜索过程获得，因为*克*是| g |<| h |的前缀。因此，可以有效地计算每个假设的前缀和前向概率，并且可以通过波束搜索期间的CTC得分排除具有不相关对齐的部分假设。因此，一次搜索表I:ASR任务。*小时*

|  |  |  |
| --- | --- | --- |
| 钟声4[35] | #话语 | 长度（h） |
| 培训 | 8,738 | 18 |
| 发展 | 3,280 | 5.6 |
| 评价 | 2,640 | 4.4 |
| 《华尔街日报》[33]，[34] | #话语 | 长度（h） |
| 培训（WSJ0 si84） | 7,138 | 15 |
| 培训（WSJ1 si284） | 37,416 | 80 |
| 发展 | 503 | 1.1 |
| 评价 | 333 | 0.7 |
| 中国证监会[36] | #话语 | 长度（h） |
| 培训（10万） | 100,000 | 147 |
| 培训（学术） | 157,022 | 236 |
| 培训（全套） | 445,068 | 581 |
| 评估（任务1） | 1,288 | 1.9 |
| 评估（任务2） | 1,305 | 2.0 |
| 评估（任务3） | 1,389 | 1.3 |
| 科大[37] | #话语 | 长度（h） |
| 培训 | 193,387 | 167 |
| 训练（速度扰动） | 580,161 | 501 |
| 发展 | 4,000 | 4.8 |
| 评价 | 5,413 | 4.9 |

与重定标方法相比，该方法能以较少的计算量减少搜索错误的数目。

## 四、 实验

我们使用四种不同的ASR任务来演示我们的实验，如表一所示。实验的第一部分使用了著名的英语干净语音语料库WSJ1和WSJ0[33]，[34]，以及一个有噪声的语音语料库CHiME-4[35]。编钟-4是在日常环境中使用平板设备录制的：咖啡馆、街道交叉口、公共交通和行人区。利用这些语料库进行的实验旨在关注我们的混合CTC/注意结构的多目标学习部分（第III-A节）的有效性，由于这些语料库的规模相对较小，因此具有不同的学习配置。

实验的第二部分利用自发日语语料库（CSJ）[36]和香港科技大学普通话-汉语会话电话语音识别语料库（HKUST）[37]扩大了语料库的规模。这些实验主要显示了我们联合解码的有效性，如第三节B所述。选择这两种语言的主要原因是，这些表意文字语言的字母序列长度（即）相对短于字母语言，这大大降低了计算复杂性，并且使得在解码器网络中处理上下文信息变得容易。事实上，我们的初步调查表明，日语和汉语端到端的ASR可以很容易地扩展，并且在不使用为大规模英语任务开发的各种技巧的情况下表现出合理的性能。*我*

表二列出了所有实验中常见的实验超参数。每个实验部分都描述了特定于任务的超参数。本文还严格遵循了端到端的ASR概念，没有使用任何发音词典、语言模型、GMM/HMM或DNN/HMM，我们的混合CTC/注意体系结构是用Chainer实现的[28]。

表二：常用实验参数。

|  |  |
| --- | --- |
| 参数初始化 | 均匀分布[-0.1，0.1] |
| #BLSTM单元的编码器 | 320 |
| #编码器投影单元 | 320 |
| 编码器子采样 | 第二层和第三层底层  （每2个特征跳过一次，产生4/T） |
| #解码器LSTM单元 | 320 |
| 优化 | 阿达德尔塔 |
| 阿达德尔塔*ρ* | 0.95 |
| 阿达德尔塔  阿达德尔塔  *γ* | 2 |
| 端点检测长度阈值*D级*结束 | 对数1E−10 |
| 终点检测得分阈值*米* | 3 |

### A、 《华尔街日报》和CHiME-4

如表一所示，对WSJ0和WSJ1的1）“eval92”和2）进行了评估

CHiME-4的“et05实际隔离1ch轨道”，而对WSJ0和WSJ1的1）“dev93”和CHiME-4的2）“dt05多隔离1ch轨道”执行超参数选择。

作为输入特征，我们使用40个mel尺度滤波器组系数及其一阶和二阶时间导数来获得每帧总共120个特征值。对于注意模型，我们只使用了32个不同的标签：26个字符、撇号、句点、破折号、空格、噪声和sos/eos标记。CTC模型使用SUBE代替SOS/EOS，我们的MOL模型既使用SOS/EOS又使用空白。编码器是一个四层BLSTM，每层和每个方向有320个单元，每个BLSTM层后面是320个单元的线性投影层。编码器的第二和第三底部LSTM层在下面的网络中每隔一秒读取一次状态特征，从而将话语长度减少四倍，即4。译码器为单层单向LSTM，320个单元。其他实验装置见表二。对于我们的MOL，我们测试了三种不同的任务权重：0.2、0.5和0.8。*T型/λ*

对于注意力和MOL模型的解码，我们使用了一种类似于[38]的传统波束搜索算法

波束尺寸为20，减少了计算量。对于CHiME4，我们手动将输出序列的最小和最大长度分别设置为输入序列长度的0.1倍和0.18倍，公式（42）中的长度惩罚设置为0.3。对于WSJ，最小和最大长度分别设置为输入序列长度的0.075倍和0.2倍，没有长度惩罚（即=0）。对于CTC模型的解码，我们采用了Viterbi序列作为结果。*γγ*

表III中的结果表明，我们提出的模型MOL在噪声蜂鸣-4和清洁WSJ任务的CER方面显著优于CTC和注意模型。我们的模型显示，验证集和评估集的相对改善率分别为6.0-8.4%和5.4-14.6%。我们观察到我们的混合CTC/注意MOL表III：干净语料库WSJ0和WSJ1的字符错误率（CER）以及嘈杂语料库CHiME-4。

                                            型号CER（有效）CER（评估）

|  |  |  |
| --- | --- | --- |
| WSJ1 SI284（80小时） | 第93版 | 评估92 |
| 反恐委员会 | 11.48 | 8.97 |
| 注意（基于内容） | 13.68 | 11.08 |
| 注意（+位置感知） | 11.98 | 8.17 |
| 摩尔（=0.2）*λ* | 11.27 | 7.36 |
| 摩尔（=0.5）*λ* | 12.00 | 8.31 |
| 摩尔（=0.8）*λ* | 11.71 | 8.45 |
| WSJ0 SI84（15小时） | 第93版 | 评估92 |
| 反恐委员会 | 27.41 | 20.34 |
| 注意（基于内容） | 28.02 | 20.06 |
| 注意（+位置感知） | 24.98 | 17.01 |
| 摩尔（=0.2）*λ* | 23.03 | 14.53 |
| 摩尔（=0.5）*λ* | 26.28 | 16.24 |
| 摩尔（=0.8）*λ* | 32.21 | 21.30 |
| 蜂鸣-4（18小时） | dt05实数 | et05真实 |
| 反恐委员会  注意（基于内容）  注意（+位置感知）  摩尔（=0.2）*λ*  摩尔（=0.5）*λ*  摩尔（=0.8）*λ* |  |  |

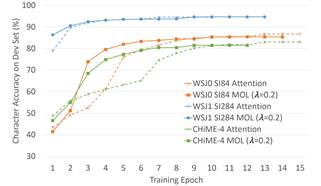


图2：学习曲线：位置感知注意模型和MOL=0.2。注意，注意力和我们的MOL的近似精度是根据地面真实历史获得的，如第II-C5节所述。*λ*

在嘈杂的CHiME-4和干净的WSJ任务中使用=0.2时，获得了最佳性能。作为参考，我们还计算了模型MOL（=0.2）的单词错误率（WER），得分为18.2%，略好于[39]中模型的WER。*λλ*

除了CER的改进之外，MOL还可以非常有助于加速所需对齐的学习。图2显示了CHiME-4、WSJ0 SI84和WSJ1 SI284验证集在训练期间的字符精度学习曲线。注意，注意力和我们的MOL的近似精确度*λ*=0.2，如第II-C5节所述，我们无法直接比较MOL和注意之间验证特征准确度的绝对值。然而，从学习曲线行为来看，MOL训练比注意训练收敛更快。

|  |
| --- |
| （a） 注意1 epoch（b）注意3 epoch（c）注意5 epoch（d）注意7 epoch（e）注意9 epoch                 （f） 摩尔1历元（g）摩尔3历元（h）摩尔5历元（i）摩尔7历元（j）摩尔9历元  图3：基于位置的注意力模型（第1行）和我们的模型MOL（第2行）在训练时期（1、3、5、7和9）的字符（y轴）和声音帧（x轴）之间学习对齐速度的比较。所有对齐都针对一个手动选择的语句（F05 442C020U）CAF（真实）在嘈杂的CHiME-4评估集中。 |

表四：日本自发语音识别（CSJ）任务语料库中常规注意的字符错误率（CERs）和提出的CTC/注意端到端混合ASR。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 任务1 | 任务2 | 任务3 |
| 注意（147h） | 20.1 | 14.0 | 32.7 |
| 摩尔（147h） | 16.9 | 12.7 | 28.9 |
| 注意（236h） | 16.3 | 12.2 | 24.7 |
| 摩尔（236小时） | 13.4 | 10.1 | 21.5 |
| 注意（581h） | 11.4 | 7.9 | 9.0 |
| 摩尔（581h） | 10.5 | 7.6 | 8.3 |
| MOL+联合解码（重扫描，581h） | 10.1 | 7.1 | 7.8 |
| MOL+联合解码（一次通过，581h） | 10.0 | 7.1 | 7.6 |
| MOL大+联合解码（重扫描，581h） | 8.4 | 6.2 | 6.9 |
| MOL大+联合解码（一次通过，581h） | 8.4 | 6.1 | 6.9 |
| GMM配电盘。[40]（AM为236小时，LM为581小时） | 11.2 | 9.2 | 12.1 |
| HMM/DNN[40]（AM为236小时，LM为581小时） | 9.0 | 7.2 | 9.6 |
| CTC音节[27]（581 h） | 9.4 | 7.3 | 7.5 |

图3显示了在训练期间角色和声音框架之间的注意对齐。我们观察到，我们的MOL在早期的训练阶段，即第5纪元学习到期望的对齐，而注意模型即使在第9纪元也不能学习到期望的对齐。这一结果表明，在我们的MOL方法中，CTC的丢失导致排列是单调的。

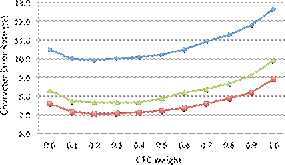
### B、 自发日语语料库

CSJ是一个标准的日语ASR任务，基于一组独白语音数据，包括学术讲座和模拟演讲。它共有581小时的训练数据和三种类型的评估数据，每个评估任务包括10节课（共5小时），如表一所示。实验设置与以前的英语实验相似，我们使用40个mel尺度滤波器组系数及其一阶和二阶时间导数作为输入特征向量。此外，我们分别使用四层BLSTM和一层LSTM作为编码器和解码器网络。我们使用了3315个不同的标签，包括汉字、两种类型的日语音节字符（平假名和片假名）、字母和阿拉伯数字，其中“空白”符号表示CTC，eos/sos符号表示注意。

表四首先比较了不同训练数据量（147、236和581小时）的常规注意和基于MOL的端到端ASR的cer（无联合解码）。在公式（38）中，被设置为0.1。在解码时，我们手动将输出序列的最小和最大长度分别设置为输入序列长度的0.1倍和0.5倍。公式（42）中的长度惩罚设置为0.1。在所有训练数据量的所有评估任务中，MOL都显著优于基于注意的ASR，这证实了MOL在我们的混合CTC/注意结构中的有效性。表IV中的结果还表明，第III-B节中描述的所提议的联合解码在不设置任何搜索参数（最大和最小长度、长度惩罚）的情况下进一步改进了性能，但是在等式（45）中仅设置了权重参数=0.1，类似于MOL情况。图4还比较了CSJ评估任务对CER的依赖性，并显示如果我们围绕MOL使用的值进行设置，它对性能不是太敏感*λγλλλλ*

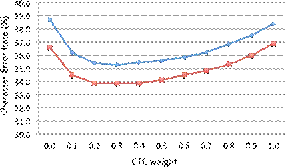
（即0.1）。

我们还比较了所提出的一个更大的网络（一个五层编码网络，MOLlarge）的方法与传统的通过使用语言资源（包括形态学资源）获得的最新技术的性能



                                     任务1任务2任务3

图4：式（45）中CTC权重对CSJ评估任务的影响。*λ*



开发集评估集

图5：式（45）中CTC权重对科大评估任务的影响。*λ*

分析器、发音词典和语言模型。GMM鉴别训练和HMM/DNN sMBR（sMBR：国家级最小贝叶斯风险）系统的最新CER来自Kaldi配方[40]和基于基于音节的CTC和地图解码的系统[27]。Kaldi配方系统使用学术讲座（236小时）进行AM培训，所有培训数据转录用于LM培训。请注意，由于所提出方法和参考方法的训练数据量和实验配置略有不同，因此很难直接比较表中列出的性能。然而，由于所提出方法的核证减排量与最佳参考结果的核证减排量相当或更好，我们可以说，所提出的方法达到了最先进的性能。

### C、 科大普通话电话演讲

科大普通话-汉语会话电话语音识别[37]有5个小时的录音进行评估，我们从训练数据中提取额外的5个小时作为开发集，并将剩余的167个小时作为训练集，正如Talbe I中所总结的，我们在初步调查的基础上，使用=0.5而不是0.1进行训练和解码，80 mel尺度滤波器组系数具有[42]中所建议的基音特征，以及五层表V：传统注意的字符错误率（CER）和建议的混合CTC/注意端到端ASR科大普通话电话语音识别任务。*λ*

|  |  |  |
| --- | --- | --- |
| 模型 | 开发 | 评估 |
| 敬告 | 40.3 | 37.8 |
| 摩尔 | 38.7 | 36.6 |
| 关注+覆盖 | 39.4 | 37.6 |
| 摩尔+覆盖率 | 36.9 | 35.3 |
| MOL+联合解码（重扫描） | 35.9 | 34.2 |
| MOL+联合解码（一次通过） | 35.5 | 33.9 |
| MOL大（速度扰动）+联合解码（重扫描） | 31.1 | 30.1 |
| MOL大（速度扰动）+联合解码（一次通过） | 31.0 | 29.9 |
| MOL+CNN+LSTML（速度扰动）+联合译码（一次通过）[41] | 29.1 | 28.0 |
| 嗯/DNN | – | 35.9 |
| HMM/LSTM（速度扰动） | – | 33.5 |
| 带语言模型的CTC[42] | – | 34.8 |
| HMM/TDNN，无晶格MMI（速度扰动）[24] | – | 28.2 |

BLSTM和两层LSTM分别用于编码器和解码器网络。其余实验条件与第IV-B节和表II中的相同。我们使用3653个带有“空白”的不同标签表示CTC，使用eos/sos表示注意。对于解码，我们还添加了基于覆盖项的解码结果[14]，如第III-B节所述（=1.5，τ=0.5，注意模型=−0.6，MOL=1.0，τ=0.5，MOL=−0.1），由于仅通过调整最大和最小长度以及长度惩罚很难消除解码期间的不规则对齐（我们将输出序列的最小和最大长度分别设置为输入序列长度的0.0和0.1倍，并在表V中设置=0.6）。*ηγηγγ*

表5的结果表明，与基于注意的方法相比，MOL和联合解码的效果更好，尤其是联合CTC/注意解码的效果更为显著。结果还表明，我们的联合译码“MOL+联合译码（一次通过）”比覆盖项“MOL+覆盖”工作得更好，其中CER从35.3%降低到33.9%[2]。与第IV-B节中的CSJ实验类似，我们在联合译码中没有使用长度惩罚项或覆盖项。与需要许多调谐参数的传统方法相比，这是联合解码的优点。此外，图5再次显示，如果我们将MOL的值（即0.5）设置在附近，它对性能不太敏感。*λλ*

最后，我们通过将音频长度线性缩放0.9和1.1（速度扰动）的因子来生成更多的训练数据。最终的模型在不使用语言资源的情况下达到了29.9%，打败了包括基于CTC的方法在内的中等先进系统[3]。

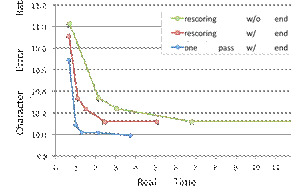


图6:CSJ任务1的一次通过和重排序方法的RTF与CER。

### D、 解码速度

我们评估了第III-B3节中描述的混合CTC/attention架构的联合解码方法的速度，其中ASR解码是在1、3、5、10和20的不同波束宽度下执行的，处理时间和CER是使用配备Intel（R）Xeon（R）处理器、E5-2690 v3、2.6 GHz的计算机测量的。虽然处理器是多核CPU，计算机也有gpu，但我们将解码程序作为单线程进程在CPU上运行，以研究其基本计算开销。

图6和图7分别显示了CSJ和HKUST任务的实时因子（RTF）和CER之间的关系。我们评估了有无端点检测的重定标方法，以及有端点检测的一次通过方法。对于这两个任务，我们可以看到端点检测成功地降低了RTF而没有任何精度下降。此外，一次通过的方法实现更快的解码与较低的CER比重排序的方法。通过单程解码，我们实现了1xRT，精度下降很小，即使它是CPU上的单线程进程。然而，由于CTC和注意机制即使在预测第一个标签时也需要访问输入话语的所有帧，因此解码过程还没有实现实时ASR。这是大多数端到端ASR方法的基本问题，将在未来的工作中得到解决。

## 五、 总结与讨论

该文提出了一种基于混合CTC/注意结构的端到端ASR，解决了基于注意的端到端ASR的失调问题。这种方法不需要语言资源，例如词法分析器、语音词典和语言模型，这些都是传统日语和汉语ASR系统的重要组成部分。然而，该方法在CSJ和香港科技大学的任务中取得了和最先进的传统系统相当的性能。此外，所提出的方法不需要GMM/HMM构造用于初始比对、DNN预训练、用于序列判别训练的格生成、解码过程中的复杂搜索（例如，FST解码器或基于词法树搜索的解码器）。因此，该方法大大简化了计算过程

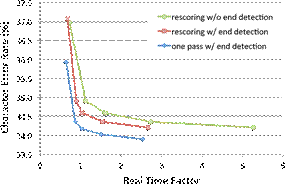


图7:HKUST测试集一次通过和重定标方法的RTF与CER。

ASR构建过程，减少代码大小和复杂性。目前，使用一个GPU对网络进行训练需要7-9天的时间，为CSJ任务提供完整的训练数据（581小时），由于建造过程的简化，这与传统最先进系统的整个训练时间相当。

未来的工作将把这项技术应用到包括英语在内的其他语言中，在这些语言中我们必须解决长序列长度的问题，这需要大量的计算成本，并且使得训练解码器网络变得困难。实际上，最近的序列到序列研究通过使用一个子词单元（将几个字母连接起来形成一个新的子词单元）[13]，[43]来处理这个问题，这将是我们端到端ASR的一个有希望的方向。另一个未来的工作是利用现有的传统HMM/DNN，如果它是可用的，除了一个端到端的概念。在我们的框架中（例如，作为另一个培训目标），将传统的HMM/DNN与CTC结合起来（而不是与CTC结合起来）是很有意思的，因为它们是互补的。进一步研究CTC在训练和解码中的应用也是今后工作的一个有趣方向。我们可以比较不同的CTC使用情况，例如，CTC只用于注意模型编码器的预训练和CTC只用于解码而不用于训练的情况。

## 致谢

我们要感谢Retrieva，Inc.的Jiro Nishitoba先生和Preferred Networks，Inc.的Shohei Hido先生就Chainer在（端到端）语音识别中的应用所作的宝贵讨论和评论。我们还要感谢日立株式会社的Kanda Naoyuki博士提供了日本基准系统的CER结果。

## 参考文献

[1] F.Jelinek，“统计方法的连续语音识别”，《IEEE会议录》，第64卷，第4期，第532-5561976页。

[2] G.Hinton，L.Deng，D.Yu，G.E.Dahl，A.-r.Mohamed，N.Jaitly，A.Senior，V.Vanhoucke，P.Nguyen，T.N.Sainath和B.Kingsbury，“语音识别中声学建模的深度神经网络：四个研究小组的共同观点”，《IEEE信号处理杂志》，第29卷，第6期，第82-972012页。

[3] T.Kudo、K.Yamamoto和Y.Matsumoto，“将条件随机场应用于日语形态学分析”，自然语言处理经验方法会议（EMNLP），第4卷，2004年，第230–237页。

[4] S.Bird，“NLTK:自然语言工具箱”，载于国际计算语言学委员会和计算语言学协会（COLING/ACL）关于交互演示会议的联合会议，2006年，第69-72页。

[5] M.Mohri，“语言和语音处理中的有限状态变换器”，计算语言学，第23卷，第2期，第269-3111997页。

[6] T.Hori和A.Nakamura，使用加权有限状态传感器的语音识别算法。Morgan&Claypool出版社，2013年。

[7] J.Chorowski、D.Bahdanau、K.Cho和Y.Bengio，“使用基于注意的递归神经网络的端到端连续语音识别：第一个结果”，arXiv预印本附件十四：1412.1602, 2014.

[8] A.Graves和N.Jaitly，“使用递归神经网络实现端到端语音识别”，国际机器学习会议（ICML），2014年，第1764-1772页。

[9] J.K.Chorowski、D.Bahdanau、D.Serdyuk、K.Cho和Y.Bengio，“基于注意力的语音识别模型”，神经信息处理系统（NIPS）进展，2015年，第577-585页。

[10] W.Chan、N.Jaitly、Q.V.Le和O.Vinyals，“聆听、参与和拼写”，IEEE声学、语音和信号处理国际会议（ICASSP），2016年，第4960-4964页。

[11] L.Lu，X.Zhang和S.Renals，“关于训练用于大词汇量端到端语音识别的递归神经网络编码器-解码器”，IEEE声学、语音和信号处理国际会议（ICASSP），2016年，第5060-5064页。

[12] D.Bahdanau、K.Cho和Y.Bengio，“通过联合学习对齐和翻译实现神经机器翻译”，arXiv预印本附件十四：1409.0473, 2014.

[13] Y.Wu，M.Schuster，Z.Chen，Q.V.Le，M.Norouzi，W.Macherey，M.Krikun，Y.Cao，Q.Gao，K.Macherey等人，“谷歌的神经机器翻译系统：弥补人类和机器翻译之间的差距”，arXiv预印本附件十四：1609.08144, 2016.

[14] J.Chorowski和N.Jaitly，“在序列到序列模型中实现更好的解码和语言模型集成”，arXiv预印本附件十四：1612.02695, 2016.

[15] A.Graves、S.Fernandez、F.Gomez和J.Schmidhuber，“连接主义者的时间分类：用递归神经网络标记未分段序列数据”，国际机器学习会议（ICML），2006年，第369-376页。

[16] Y.Miao，M.Gowayyed和F.Metze，“EESEN:使用深度RNN模型和基于WFST的解码的端到端语音识别”，IEEE自动语音识别和理解研讨会（ASRU），2015年，第167-174页。

[17] D.Amodei，R.Anubhai，E.Battenberg，C.Case，J.Casper，B.Catanzaro，J.Chen，M.Chrzanowski，A.Coates，G.Diamos等人，“深度语音2：英语和汉语的端到端语音识别”，arXiv预印本附件十四：1512.02595, 2015.

[18] H.Soltau、H.Liao和H.Sak，“神经语音识别器：用于大词汇量语音识别的声词LSTM模型”，arXiv预印本附件十四：1610.09975, 2016.

[19] S.Kim、T.Hori和S.Watanabe，“使用多任务学习的联合CTC基于注意力的端到端语音识别”，IEEE声学、语音和信号处理国际会议（ICASSP），2017年，第4835–4839页。

[20] T.Hori、S.Watanabe和J.R.Hershey，“用于端到端语音识别的联合ctc/注意解码”，计算语言学协会（ACL），2017年。

[21]N.Xue等，“汉语分词作为字符标注”，计算语言学与汉语处理，第8卷，第1期，第29-48页，2003年。

[22]H.Bourlard和N.Morgan，《连接主义语音识别：一种混合方法》。Kluwer学术出版社，1994年。

[23]T.Mikolov、M.Karafiat、L.Burget、J.Cernock´y和S.Khudanpur，“基于递归神经网络的语言模型”，Interspeech，2010，第1045-1048页。

[24]D.Povey、V.Peddinti、D.Galvez、P.Ghahrmani、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，第2751-2755页。

[25]S.Hochreiter和J.Schmidhuber，“长-短期记忆”，《神经计算》，第9卷，第8期，第1735-17801997页。

[26]A.Graves，N.Jaitly和A.-r.Mohamed，“深度双向LSTM混合语音识别”，IEEE自动语音识别与理解研讨会（ASRU），2013年，第273-278页。

[27]N.Kanda，X.Lu和H.Kawai，“CTC声学模型的最大后验概率解码”，Interspeech，2016，第1868-1872页。

[28]S.Tokui，K.Oono，S.Hido和J.Clayton，“Chainer:下一代深度学习开源框架”，NIPS机器学习系统研讨会论文集，2015年。

[29]D.Povey、A.Ghoshal、G.Boulianne、L.Burget、O.Glembek、N.Goel、M.Hannemann、P.Motlicek、Y.Qian、P.Schwarz、J.Silovsky、G.Stemmer和K.Vesely，“Kaldi语音识别工具包”，IEEE自动语音识别和理解研讨会（ASRU），2011年。

[30]L.Lu，L.Kong，C.Dyer和N.A.Smith，“使用CTC和分段CRF进行语音识别的多任务学习”，Interspeech，2017年。

[31]D.Bahdanau、J.Chorowski、D.Serdyuk、P.Brakel和Y.Bengio，“基于端到端注意的大词汇表语音识别”，IEEE声学、语音和信号处理国际会议（ICASSP），2016年，第4945–4949页。

[32]A.Graves，“递归神经网络监督序列标记”，M¨unchen¨理工大学博士论文，2008年。

[33]L.D.Consortium，“CSR-II（wsj1）complete”，语言数据联合会，费城，第LDC94S13A卷，1994年。

[34]J.Garofalo、D.Graff、D.Paul和D.Pallett，“CSR-I（wsj0）complete”，语言数据联盟，费城，第LDC93S6A卷，2007年。

[35]E.Vincent、S.Watanabe、A.A.Nugraha、J.Barker和R.Marxer，“鲁棒语音识别中的环境、麦克风和数据模拟不匹配分析”，《计算机语音和语言》，2017年，第535-557页。

[36]K.Maekawa、H.Koiso、S.Furui和H.Isahara，“日语自发语音语料库”，国际语言资源与评价会议（LREC），第2卷，2000年，第947-952页。

[37]刘彦，冯国平，杨彦，C.Cieri，黄圣贤，D.Graff，“香港科技大学/移动电话系统：一个非常大规模的普通话电话语音语料库”，中文口语处理。斯普林格，2006年，第724-735页。

[38]I.Sutskever、O.Vinyals和Q.V.Le，“神经网络的序列到序列学习”，神经信息处理系统（NIPS）进展，2014年，第3104-3112页。

[39]D.Bahdanau、J.Chorowski、D.Serdyuk、P.Brakel和Y.Bengio，“端到端基于注意力的大词汇量语音识别”，arXiv预印本附件十四：1508.04395, 2015.

[40]T.Moriya、T.Shinozaki和S.Watanabe，“日本自发语音识别的Kaldi配方及其评估”，日本声学学会（ASJ）秋季会议，第3-Q-72015号。

[41]T.Hori，S.Watanabe，Y.Zhang，和W.Chan，“基于深度CNN编码器和RNN-LM的联合CTcatting端到端语音识别的进展”，Interspeech，2017年。

[42]Y.Miao，M.Gowayyed，X.Na，T.Ko，F.Metze和A.Waibel，“CTC声学模型的经验探索”，IEEE声学、语音和信号处理国际会议（ICASSP），2016年，第2623-2627页。

[43]W.Chan，Y.Zhang，Q.Le和N.Jaitly，“潜在序列分解”，国际学习表征会议（ICLR），2017年。

渡边真司是美国马里兰州巴尔的摩市约翰霍普金斯大学的副教授，2006年在日本东京早稻田大学获得博士学位。2001年至2011年，他是日本京都NTT通信科学实验室的研究科学家。2009年1月至3月，他在佐治亚州亚特兰大市乔治亚理工学院（Georgia institute of technology）担任访问学者。2012年至2017年，他在剑桥三菱电气研究实验室（MERL）担任高级首席研究科学家，

妈妈。他的研究兴趣包括-

汉语学习与语音和口语处理。他在期刊和会议上发表了100多篇论文，并获得了包括2003年IEICE最佳论文奖在内的多项奖项。他曾担任《IEEE语音和语言处理学报》的副主编，并且是包括IEEE信号处理学会语音和语言技术委员会在内的多个委员会的成员。

Takaki Hori分别于1994年和1996年获得日本Yonezawa山形大学电气与信息工程学士和硕士学位，1999年获得山形大学系统与信息工程博士学位。1999年至2015年，他在日本电报电话公司的网络空间实验室和通信科学实验室从事语音识别和口语理解研究。

2006年至2007年，他是麻省理工学院（MIT）的访问科学家。自2015年以来，他一直是美国马萨诸塞州剑桥市三菱电机研究实验室（MERL）的高级首席研究科学家，在语音和语言研究领域撰写了90多篇同行评议论文。2009年，他获得了电信进步基金会颁发的第24届电信系统技术奖，2012年获得了日本信息处理学会颁发的IPSJ Kiyasu特别工业成就奖，2013年获得了筑本协会颁发的第58届前岛久冈奖。

2014年以来，金素云是卡内基梅隆大学的博士生。她的研究兴趣包括机器学习、深度学习和口语处理。2016年夏天，她在马萨诸塞州剑桥市三菱电机研究实验室（MERL）的语音和音频实验室实习。她于2014年获得卡内基梅隆大学语言技术硕士学位。她是IEEE的学生会员。

John R.Hershey自2010年以来一直是MERL的高级原则研究科学家和语音和音频团队负责人。在加入MERL之前，John在IBM的T.J.Watson Research工作了5年

并领导了噪声鲁棒性语音识别团队。在获得UCSD博士学位后，他还在微软研究院的演讲小组做了一年的访问研究员。

多年来，他在机器感知、语音处理、语音识别等领域发表了100多篇论文，并获得了30多项专利

识别和自然语言理解。

Tomoki Hayashi分别于2013年和2015年在日本名古屋大学获得工程学学士学位和信息科学硕士学位。他目前是名古屋大学的博士生。他的研究兴趣包括统计语音和音频信号处理。他获得了日本声学学会2014年学生演讲奖。他是日本声学学会的学生会员，也是IEEE的学生会员。

[[1]](" \l "_ftnref1" \o ")在CTC和基于注意的方法中，由于子采样技术[24]，[10]，隐藏状态的序列长度将短于原始输入序列长度（即在CTC情况下，| Z |<T）。然而，为了简单起见，本文中的公式保留了相同的索引和长度。*t型T型*

[[2]](" \l "_ftnref2" \o ")我们进一步进行了CTC和覆盖项联合译码的实验。虽然我们调整了包括长度惩罚在内的解码参数，但其CER为34.2%，略低于联合解码的33.9%

[[3]](" \l "_ftnref3" \o ")尽管所提出的方法没有达到时延神经网络（TDNN）的性能，但该方法充分利用了语言资源，包括语音表示和基于音素的语言模型，在识别训练中[24]。此外，我们最近的工作得分为28.0%，并且在先进的网络架构下优于无晶格MMI结果[41]。