Two-Pass End-to-End Speech Recognition

两遍端到端语音识别

*塔拉·N·赛纳*[[1]](" \l "_ftn1" \o ")*，彭若明*∗*，大卫·赖巴赫，何延章，罗希特·普拉巴瓦尔卡，李伟，*

*米尔科·维桑泰、乔亮、特雷弗·斯特罗曼、吴永辉、伊恩·麦格劳、钟正秋*

美国谷歌公司

{tsainath，rpang}@谷歌网站

# 摘要

The requirements for many applications of state-of-the-art speech recognition systems include not only low word error rate (WER) but also low latency. Specifically, for many use-cases, the system must be able to decode utterances in a streaming fashion and faster than real-time. Recently, a streaming recurrent neural network transducer (RNN-T) end-to-end (E2E) model has shown to be a good candidate for on-device speech recognition, with improved WER and latency metrics compared to conventional on-device models [1]. However, this model still lags behind a large state-of-the-art conventional model in quality [2]. On the other hand, a non-streaming E2E Listen, Attend and Spell (LAS) model has shown comparable quality to large conventional models [3]. This work aims to bring the quality of an E2E streaming model closer to that of a conventional system by incorporating a LAS network as a second-pass component, while still abiding by latency constraints. Our proposed two-pass model achieves a 17%-22% relative reduction in WER compared to RNN-T alone and increases latency by a small fraction over RNN-T.

最先进的语音识别系统的许多应用要求不仅包括低字错误率（WER），而且还包括低延迟。具体来说，对于许多用例，系统必须能够以流的方式解码话语，并且比实时更快。最近，流式递归神经网络传感器（RNN-T）端到端（E2E）模型被证明是设备上语音识别的一个很好的候选模型，与传统的设备上模型相比，该模型具有更好的WER和延迟度量[1]。然而，该模型在质量上仍然落后于大型最先进的传统模型[2]。另一方面，非流E2E侦听、参与和拼写（LAS）模型显示出与大型传统模型相当的质量[3]。这项工作的目的是使E2E流模型的质量更接近于传统系统的质量，通过将LAS网络作为第二通过组件，同时仍然遵守延迟约束。我们提出的双通道模型达到了预期的效果

与单纯RNN-T相比，WER相对减少17%-22%，并且比RNN-T增加一小部分潜伏期。

# 1.     介绍

语音识别的端到端模型（E2E）越来越流行[1、3、4、5、6、7、8、9]。这些模型将声学、语音和语言模型（AM、PM、LMs）合并成一个单一的网络，与具有单独AM、PM和LMs的传统ASR系统相比，显示出了具有竞争力的结果。E2E型号对设备上的ASR特别有吸引力，因为它们可以比同等尺寸的设备上的传统型号[10]表现更好。

在具有直接用户交互的设备上运行ASR带来了许多挑战。首先，识别结果必须是流式的。也就是说，单词一开口就应该出现在屏幕上。第二，模型必须有一个小的延迟（即，用户说话和文本出现之间的延迟），因此在移动设备上以实时或比实时更快的速度运行。第三，该模型必须能够利用用户上下文[11]（例如联系人列表、歌曲名称等）来提高识别率。最近，我们提出了一个满足这些约束条件的RNN-T E2E模型[1]。然而，RNN-T模型的质量仍然落后于传统的大型模型[2]。

非流E2E模型，如Listen、attent和Spell（LAS）[7]，已显示出与大型传统模型[3]的竞争性能。但是，LAS模型不能流式传输，因为它们必须处理整个音频段，因此在交互式应用程序中使用它们很有挑战性。

在双通译码中，第二通模型通常用于通过使用晶格重排序[12]或n-最佳重排序[13]来改善第一通模型的初始输出。在获得质量增益的同时保持用户感知的低延迟是应用二次传递模型的主要挑战。语言模型重排序通常用于多通解码[14、15、16、17]，但最近已与LAS模型一起用于重排序第一通传统模型的假设[18]。我们可以认为LAS解码器，它从编码器中获取声学信息，从先前的预测中获取语言模型信息，严格地说，它比二次通过语言模型更强。因此，在这项工作中，我们探讨使用LAS模型进行二次通过处理。

具体地说，我们探讨了一种双通道结构，其中RNN-T解码器和LAS解码器共享一个编码器网络。与RNN-T和LAS的专用编码器相比，共享编码器可以减少模型尺寸和计算成本。在推理过程中，RNN-T模型产生流预测，而LAS解码器最终完成预测。我们通过运行LAS解码器作为波束搜索与从RNN-T重新筛选假设来探索折衷。

我们的实验是在一个30000小时的语音搜索任务上进行的。我们发现，对于较短的话语（SU）测试集，使用LAS第二遍波束搜索，我们可以比第一遍RNN-T获得15%的相对改进，但是对于较长的话语（LU），模型会退化，这是注意力模型的一个常见问题[19]。相比之下，第二遍重扫描给了我们一个更好的折衷为SU和LU相比，波束搜索。接下来，我们尝试通过改变训练目标函数，使之更接近于重分类，来改进重分类模型。具体而言，我们采用最小字错误率（MWER）训练策略[20]，其中来自RNN-T的假设被用作LAS解码器的输入，并且LAS解码器被训练以最小化预期字错误率。此外，我们通过使用自适应波束运行第一通RNN-T模型[21]并在重新排序之前修剪第一通晶格来降低计算成本。总的来说，我们发现我们提出的LAS重定标模型与一次通过RNN-T模型相比，提供了17%到22%的相对改善，而不降低偏置精度。此外，第二遍LAS解码器将终结延迟增加不到200ms，这被认为是可接受的交互延迟的限制[22]。

本文的其余部分组织如下。第2节描述了本文研究的双通道结构和各种推理策略。第3节介绍了实验，第4节讨论了结果。最后，第五部分对全文进行了总结，并对今后的工作进行了展望。

# 2.     二通E2E ASR

## 2.1.    模型体系结构

图1显示了建议的双通道体系结构。

我们将参数化的输入声帧表示为=**十**

（x1…xT），其中∈本工作中的稀有对数梅尔滤波器组能量（=80），并表示中的帧数。在第一过程中，每个声学帧通过由多层LSTM组成的共享编码器来获得输出，该输出被传递到RNN-T解码器以流式方式在每个时间步产生。在第二个过程中，所有帧的共享编码器的输出=（e1…eT）被传递到LAS解码器。在训练过程中，LAS解码器根据训练结果计算输出。在解码期间，其可另外使用如下所述。**xxxeye是***t型d级d级T型t型t型右我右*

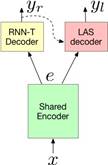


图1：双通道架构

## 2.2.    解码

我们探讨在两种不同的解码模式下使用LAS解码器。明确地，

•在“第二波束搜索”模式下，它仅产生RNN-T解码器的输出，忽略其输出。**是的***我右*

•在“重新排序”模式下，我们首先从RNN-T解码器中选取top-K假设。然后，我们在教师强制模式下，在注意力开启的情况下，对每个序列运行LAS解码器，以计算分数，该分数结合了序列的对数概率和注意力覆盖惩罚[23]。选择LAS得分最高的序列作为输出序列。**电子**

## 2.3.    培训

在本节中，我们将描述双通道模型的训练策略。

### 2.3.1.    综合损失

理论上，我们可以通过随机初始化训练一个双过程模型，其中包含以下组合损失，其中表示地面真实记录：**是的**∗

**我**组合（x，y∗）=λLRNNT（x，y∗）+（1−λ）LLAS（x，y∗）（1）

在上面的等式中，是一个超参数，我们在设置中将其设置为0.5，以使RNN-T和LAS损失相等。在实践中，我们发现直接从头开始的训练是不稳定的，这主要是因为在从头开始的训练中，RNN-T和LAS的损失在很大的不同范围内。因此，我们采取一个多步骤的过程来训练模型：*λ*

1.    按照[1]训练RNN-T模型；

2.    将在步骤（1）中训练的编码器冻结，并训练LAS解码器，如[3]所示。

3.    “深度微调”：以综合损耗同时训练共享编码器和两个解码器。

### 2.3.2.    MWER培训

等式1中的损耗的缺点之一是二次通过LAS解码器独立于RNN-T解码器而优化。这意味着在第2.2节中概述的训练和解码策略之间存在不匹配。

为了解决这个问题，我们使用了一个额外的训练步骤来进一步完善LAS解码器，以最小化错误，遵循了[20]中介绍的MWER训练过程。具体地说，给定的输入，基本事实记录，由LAS（ym | x）计算出的在教师强迫下任何给定目标序列的概率**xyy公司**∗*第米*

（其中，=r if由RNN-T给出，=l if由LAS给出），我们将预先训练的双过程模型细化如下。*米***yy年***米米米*

首先，我们使用两遍模型中的一个解码器进行波束搜索，得到一组假设{h1，…，hb}，其中是波束大小。为了使MWER训练匹配译码，目标译码方式决定了训练码的生成。对于要在“第二波束搜索”模式下使用的LAS解码器，我们通过在LAS解码器自身打开（=l）的情况下运行波束搜索来计算。另一方面，对于要在“重扫描”模式下使用的LAS解码器，我们通过使用第一通RNN-T解码器（=r）运行波束搜索来计算（x）。*米嗯b类嗯嗯***十***米嗯米*

对于每个序列∈Hm**yy年***米*，设为*米*，let（y，Hm）=*W型*

是单词er的平均数-

的ROR*嗯*，然后让

是中的相对字错误率。我们还让**是的***米嗯*

*第*ˆ（y | x，H）=P*米米***yy年***我*其中∈（y | x（）y | x）表示假设的条件概率。MWER损失定义为*第小时毫米第我米嗯*

**我**兆瓦（xy，∗）=十*第*ˆ（yxm |，Hm）宽ˆ（年\*月）(2)

**是的***米*∈H（x）*米*

我们训练LAS解码器以最小化MWER损失和最大似然交叉熵损失的组合：

**我**兆瓦（xy，∗）+λ最大似然估计logP（yx |）(3)

其中是一个超参数，我们在[20]之后通过实验将其设置为=0.01。*λ*最大似然估计*λ*最大似然估计

# 3.     实验细节

## 3.1.    数据集

我们的实验是在一个30000小时的训练集上进行的，这个训练集由4300万个英语单词组成。这些训练话语都是匿名和手写的，是谷歌在美国语音搜索流量的代表。多风格训练（MTR）数据是通过使用室内模拟器人为地破坏干净的话语，添加不同程度的噪声和混响，平均信噪比为12dB[24]。噪声源来自YouTube和日常生活中嘈杂的环境录音。我们报告的主要测试集包括长度小于5.5秒的∼14K短句（SU）和长度大于5.5秒的∼16K长句（LU）

5.5秒，都是从谷歌流量中提取的。

为了评估上下文偏向的性能，我们报告了contacts测试集的性能，该测试集由呼叫/文本联系人请求组成。这个集合是通过从web上挖掘联系人姓名，并使用一个语音的串联TTS方法合成每个类别中的TTS语句而创建的[25]。然后将噪声人工添加到TTS数据中，类似于上述过程[24]。为了使模型预测偏向于接触，我们在接触短语列表上构造了一个偏置FST，并在推理过程中在偏置FST和E2E模型之间进行浅层融合。关于E2E浅熔合偏压的更多细节，请参阅[26]。

## 3.2.    模型架构细节

所有实验都使用80维log-Mel特征，用25ms窗口计算，每10ms移动一次。与[2]类似，在当前帧，这些特征向左叠加2帧，并以30ms的帧速率下采样。*t型*

所有实验都使用了文献[1]中描述的相同编码器网络。它由8个LSTM层组成，每个层有2048个隐藏单元，后面是一个640维的投影层。我们在编码器的第二个LSTM层之后插入一个时间缩减层，缩减因子为2。*不*

RNN-T解码器包括预测网络和联合网络。该预测网络具有2层共2048个隐单元的LSTM层和每层640维的投影以及128个单元的嵌入层。编码器和预测网络的输出被馈送到具有640个隐藏单元的联合网络。LAS解码器由具有四个注意头的多头注意[27]组成，该注意头被馈入2048个隐藏单元的2个LSTM层和640维投影层。它有96个单元的嵌入层。这两个译码器都经过训练，可以预测4096个单词的片段[28]，这些片段是使用大量文本转录本得到的。

RNN-T模型的总尺寸为114M参数，附加的二通LAS解码器为33M参数。所有模型都在Tensorflow[29]中使用Lingvo[30]工具箱在8×8张量处理单元（TPU）切片上进行训练，全局批量大小为4096。

## 3.3.    测量延迟

由于计算设备可能不同，我们使用简化的计算模型来估计延迟。首先，我们假设CPU上的带宽是10GB/秒；这个数字在现代移动CPU的范围内。我们还假设每个假设是独立计算的，这意味着主要的运算是矩阵/向量乘法，其时间将由加载到CPU的矩阵参数的速度决定。*K公司*

假设没有中断或跨波束搜索假设的批处理，当使用令牌上的假设进行固定波束解码/重扫描时，根据等式4计算延迟。当使用自适应波束生成晶格时，我们假设在计算延迟时，·N现在由晶格弧的数量代替。*小时不小时*

                        延迟解码器(4)

其中表示模型解码器部分中的字节数。*米*解码器

我们报告了90%的tile LU集的延迟，该集有较长的话语。我们假设90%的tile包含大约295个音频帧和=28个令牌的目标序列。最后，假设量化后的LAS解码器的33M参数为33MB，我们发现其精度下降可以忽略不计[1]。我们的目标是确保90%磁贴上的第二次传递延迟低于200ms，以使用户感知的延迟最小化[22]。*不米*解码器

# 4.     结果

## 4.1.    第二波束搜索

表1显示了在第二波束搜索模式下运行LAS解码器的结果。为了进行比较，该表还显示了两条基线B0-B1，即仅RNN-T和仅LAS模型，从零开始分别训练。所有结果都是在固定光束尺寸=8的情况下得到的。*小时*

实验E0表明，当编码器从RNN-T模型初始化并保持固定时，LAS解码器的性能比具有专用编码器（B1）的仅LAS模型差，说明了通过单独调整LAS解码器来将单个编码器与不同类型的解码器共享的挑战。当我们在从E0初始化的模型中联合训练编码器和两个解码器时，SU和LU的模型质量（E1）都比E0有所提高。总的来说，我们发现第二波束搜索在SU上优于RNN-T（B0），但在LU上有所下降，这是长话语注意模型的一个常见问题[19]。

表1:WER结果，LAS波束搜索。

|  |  |  |  |
| --- | --- | --- | --- |
| 实验ID | 模型 | 苏 | 卢 |
| *B0级* | RNN-T公司 | 6.9 | 4.5 |
| *地下一层* | 仅限LAS | 5.4 | 4.5 |
| *E0级* | 冻结共享Enc | 6.4 | 5.3 |
| *E1级* | 深度微调 | 6.1 | 4.8 |

## 4.2.    重新扫描

我们注意到RNN-T-only模型（B0）的oracle WER远低于其解码WER。这促使我们探索用LAS解码器重新编码RNN-T假说。表2比较了使用波束搜索（E1）运行LAS和使用重扫描（E2）运行LAS的性能。该表显示，与波束搜索相比，重扫描对SU上的WER产生的影响很小，这可能是因为具有更高SUR 6.9（B0）的第一通RNN-T解码器生成的一组假设的质量略低于LAS解码器在波束搜索期间生成的假设。然而，LU上的重扫描质量要比beam搜索好得多，这可能是因为RNN-T（B0）比LAS在长话语上表现得更好。总的来说，LAS重扫描不仅在第一次通过RNN-T时显著提高了suwer，而且还提高了LU的WER，证明了重扫描能够结合RNN-T和LAS的优点。由于重新排序为我们提供了SU和LU质量之间的最佳折衷，因此我们将只关注重新排序，并在下一节中介绍进一步的改进。

表2:WER结果，LAS重新扫描。

|  |  |  |  |
| --- | --- | --- | --- |
| 实验ID | 解码 | 苏 | 卢 |
| *B0级* | RNN-T公司 | 6.9 | 4.5 |
| *地下一层* | 仅限LAS | 5.4 | 4.5 |
| *E1级* | 波束搜索 | 6.1 | 4.8 |
| *E2级* | 重新扫描 | 6.2 | 4.1 |

## 4.3.    进一步的重新排序改进

### 4.3.1.    自适应波束

为了弥补二次波束搜索与SU重扫描之间的差距，我们首先探索了用更大的一次通过RNN-T波束来增加重扫描候选的多样性。表3显示，随着光束尺寸的增加（E2-E4），WER得到改善，但自然会以成比例增加的第一次通过计算成本为代价。为了解决这个问题，我们研究了一种自适应波束搜索策略[21]。具体地说，如果第一次通过的波束候选距离当前最佳候选的阈值太远，我们会对它们进行修剪，其中阈值会在[1]之后优化第一次通过的延迟。该表显示，使用自适应波束（E5），我们可以获得与固定但较大的波束（E3）类似的功率。

表3：第一次通过固定波束与自适应波束的重新扫描。

|  |  |  |  |
| --- | --- | --- | --- |
| 实验ID | 首次通过最大光束尺寸 | 苏 | 卢 |
| *E2级* | 固定，8 | 6.2 | 4.1 |
| *E3型* | 固定，10 | 6.2 | 4.1 |
| *E4类* | 固定，16 | 6.1 | 4.1 |
| *E5级* | 自适应，10 | 6.2 | 4.0 |

自适应波束还有一个额外的优点，即它生成一个用于重排序的晶格，而不是一个N-最佳列表。重排序格比重排序N-最佳列表更有效，因为它避免了对候选序列之间的公共前缀进行重复计算，因此应该减少延迟。作为提醒，等式4中的延迟现在是通过观察晶格中的总弧来计算的。表4比较了自适应波束和固定波束的N-最佳重排序，其中我们重排序了所有首过假设。从表中可以看出，使用自适应波束和晶格重频，我们可以比使用N最佳重频的固定波束减少延迟。然而，延迟仍然高于我们的预算。

表4：延迟与重排序方法。

|  |  |
| --- | --- |
| 战略 | 延迟（毫秒） |
| 第一次通过固定，N-最佳重新扫描 | 369.6 |
| 第一次通过自适应，晶格重扫描 | 247.5 |

为了进一步减少延迟，我们探索了在每一步中减少重新扫描的晶格中最大弧的数量。表5显示，我们可以将重新扫描的假设限制为4，我们发现这不会降低准确性，也会减少延迟。总的来说，重新调整自适应波束格的二程解码器符合我们的200ms延迟预算。

表5：双通道性能与Las重新扫描光束尺寸的关系。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 光束尺寸 | 苏 | 卢 | 联络 | 延迟（毫秒） |
| 2 | 6.2 | 4.0 | 7.5 | - |
| 4 | 6.2 | 4.0 | 7.1 | 171.6 |
| 8 | 6.2 | 4.0 | 7.1 | 247.5 |

### 4.3.2.    兆瓦

最后，我们报告两个通过MWER训练后的结果。由于LAS解码器将被用于重排序，因此我们使用RNN-T来提供LAS解码器的候选假设

MWER培训。表6显示，MWER改进了重排序

苏、鲁的相对增长率均为8%。总的来说，两遍重排序模型给出的SU和LU的相对减少率分别为17%和22%。

表6:MWER训练后的两次复测结果。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验ID | 模型 | 苏 | 卢 | 联络 |
| *B0级* | 仅RNN-T | 6.9 | 4.5 | 7.0 |
| *E6型* | 无MWER | 6.2 | 4.0 | 7.1 |
| *E7型* | 兆瓦 | 5.7 | 3.5 | 7.0 |

## 4.4.    与传统大型模型的比较

我们工作的目标是在E2E系统中实现与大型传统模型相当的性能[2]。在此基础上，我们通过一个“并排”（SxS）的评估来比较我们提出的两遍重排序模型与一个大型的传统模型的性能。在这个实验中，每一个话语都被常规模式和双过程模式所转录。我们收集了两个模型中转录不同的500个话语，并将这些话语发送给两个人类转录者进行评分。每一份成绩单都被评为两次超过传统模式的胜利（只有两次通过是正确的），两次超过传统模式的失败（只有传统模式是正确的），或中立（两种模式都是正确的或不正确的）。与自动WER评估不同，这种并排评估允许评分员确定两个不同的成绩单都是正确的；这有时会导致与自动评估不同的结论。我们报告以下统计数据来定量评估SxS：

•改变了两种模型产生不同假设的话语百分比

•Wins:#在utts中，双程假设是正确的，传统模型是错误的

•损失：两通假设不正确，传统模型正确

•中性点：#两通和传统型号的UTT都正确或不正确

•p值：与传统模型相比，双通道下功率变化的静态意义

表7显示，双通道模型改变了约13%的流量。与传统模型相比，双通模型的损失（61）略高于wins（48），但大多数假设在两个系统之间具有中性评级（391）。总体而言，p值表明两个模型之间的性能差异在统计学上不显著。

表7:SxS结果（常规与双通道）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变化（%） | 赢 | 损失 | 中立的 | p值 |
| 13.2 | 48 | 61 | 391 | 10.0%-20.0% |

误差的进一步分析如表8所示。与传统模型相比，双过程模型训练的纯文本数据少了一个数量级，因此在专有名词（PN）和弱语言模型（wLM）上丢失。相反，由于双过程模型是在书面领域训练的，并且隐含地学习了文本规范化（TN），因此与传统的基于规则的文本规范步骤的模型相比，双过程模型在这方面是成功的。

表8：传统模型与双通道模型的误差分析。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 类型 | 传统的 | 两通 |
| 损失 | 零件号 | 爱丽丝餐厅 | 艾莉森餐厅 |
|  | wLM公司 | 索尼等离子电视47 | 索尼定价电视中的47 |
| 赢 | 田纳西州 | 万维网纽约时报网 | www.nytimes.com |
|  | 田纳西州 | 约翰·史密斯办公室 | 约翰·史密斯的办公室 |

# 5.     结论

在本文中，我们提出了一个两程E2E解。具体地说，我们使用二通LAS解码器从一通RNN-T系统中重新存储假设。我们发现，与仅RNN-T相比，这种方法的WER降低了17%到22%，并且延迟增加了不到200ms。

# 6.     工具书类

[1] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y.Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, “Streaming End-to-end Speech Recognition For Mobile Devices,” in Proc. ICASSP, 2019.

[2] G. Pundak and T. N. Sainath, “Lower frame rate neural network acoustic models,” in Proc. Interspeech, 2016.

[3] C. Chiu, T. N. Sainath, Y.Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, N. Jaitly, B. Li, and J. Chorowski, “State-of-the-art speech recognition with sequence-to-sequence models,” in Proc. ICASSP, 2018.

[4] A. Graves, “Sequence transduction with recurrent neural networks,” CoRR, vol. abs/1211.3711, 2012.

[5] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep neural networks,” in Proc. ICASSP, 2012.

[6] K. Rao, H. Sak, and R. Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer,” in Proc. ASRU, 2017, pp. 193–199.

[7] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

[8] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in Proc. ICASSP, 2017, pp. 4835–4839.

[9] C.-C. Chiu and C. Raffel, “Monotonic chunkwise alignments,” in Proc. ICLR, 2017.

[10] I. McGraw, R. Prabhabalkar, R. Alvarez, M. Gonzalez, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, “Personalized speech recognition on mobile devices,” in Proc. ICASSP, 2016.

[11] P. Aleksic, M. Ghodsi, A.Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach, and P. Moreno, “Bringing contextual information to Google speech recognition,” in Proc. Interspeech, 2015.

[12] S. Ortmanns, H. Ney, and X. Aubert, “A word graph algorithm for large vocabulary continuous speech recognition,” Computer Speech and Language, vol. 11, no. 1, pp. 43–72, Jan. 1997.

[13] R. Schwartz and S. Austin, “A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses,” in Proc. ICASSP, 1991, pp. 701–704.

[14] M. Sundermeyer, H. Ney, and R. Schl¨uter, “From Feedforward to Recurrent LSTM Neural Networks for Language Models,” IEEE Trans. Audio, Speech and Language Processing, vol. 23, no. 3, pp. 517–528, 2015.

[15] X. Liu, X. Chen, Y. Wang, M. J. F. Gales, and P. C. Woodland, “Two Efficient Lattice Rescoring Methods Using Recurrent Neural Network Language Models,” IEEE Trans. Audio, Speech and Language Processing, vol. 24, no. 8, pp. 1438–1449, 2016.

[16] S. Kumar, M. Nirschl, D. Holtmann-Rice, H. Liao, A. T. Suresh, and F. Yu, “Lattice Rescoring Strategies for Long Short Term Memory Language Models in Speech Recognition,” in Proc. ASRU, 2017.

[17] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in Proc. ICASSP, 2018.

[18] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, and N. Jaitly, “An analysis of ”attention” in sequence-to-sequence models,” in Proc. Interspeech, 2017.

[19] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in Proc. NIPS, 2015.

[20] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C. C. Chiu, and A. Kannan, “Minimum Word Error Rate Training for Attention-based Sequence-to-sequence Models,” in Proc. ICASSP, 2018.

[21] B. T. Lowerre, The Harpy Speech Recognition System., Ph.D. thesis, Pittsburgh, PA, USA, 1976.

[22] R. B. Miller, “Response time in man-computer conversational transactions,” in Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, New York, NY, USA, 1968, AFIPS ’68 (Fall, part I), pp. 267–277, ACM.

[23] J. K. Chorowski and N. Jaitly, “Towards Better Decoding and Language Model Integration in Sequence to Sequence Models,” in Proc. Interspeech, 2017.

[24] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, “Generated of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home,” in Proc. Interspeech, 2017.

[25] X. Gonzalvo, S. Tazari, C. Chan, M. Becker, A. Gutkin, and H. Silen, “Recent Advances in Google Real-time HMM-driven Unit Selection Synthesizer,” in Proc. Interspeech, 2016.

[26] G. Pundak, T. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, “Deep Context: End-to-End Contextual Speech Recognition,” in Proc. SLT, 2018.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” CoRR, vol. abs/1706.03762, 2017.

[28] M. Schuster and K. Nakajima, “Japanese and Korean voice search,” in Proc. ICASSP, 2012.

[29] M. Abadi et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” Available online: http://download.tensorflow.org/paper/whitepaper2015.pdf, 2015.

[30] Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, et al., “Lingvo: a modular and scalable framework for sequenceto- sequence modeling,” 2019.