SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition

SpecAugment：一种用于自动语音识别的简单数据增强方法

*Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu,*

*Barret Zoph, Ekin D. Cubuk, Quoc V. Le*

谷歌大脑

{danielspark, williamchan, ngyuzh, chungchengc, barretzoph, cubuk, qvl}@google.com

# 摘要

We present SpecAugment, a simple data augmentation method for speech recognition. SpecAugment is applied directly to the feature inputs of a neural network (i.e., filter bank coefficients). The augmentation policy consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps. We apply SpecAugment on Listen, Attend and Spell networks for end-to-end speech recognition tasks. We achieve state-of-the-art performance on the LibriSpeech 960h and Swichboard 300h tasks, outperforming all prior work. On LibriSpeech, we achieve 6.8% WER on test-other without the use of a language model, and 5.8% WER with shallow fusion with a language model. This compares to the previous stateof- the-art hybrid system of 7.5% WER. For Switchboard, we achieve 7.2%/14.6% on the Switchboard/CallHome portion of the Hub5’00 test set without the use of a language model, and 6.8%/14.1% with shallow fusion, which compares to the previous state-of-the-art hybrid system at 8.3%/17.3% WER.

Index Terms: end-to-end speech recognition, data augmentation

提出了一种简单的语音识别数据增强方法SpecAugment。SpecAugment直接应用于神经网络的特征输入（即滤波器组系数）。增强策略包括特征扭曲、频率通道掩蔽块和时间步长掩蔽块。我们将SpecAugment应用于端到端语音识别任务的Listen, Attendand Speell网络。我们在LibriSpeech 960h和Swichboard 300h任务上实现了最先进的性能，优于以前的所有工作。在LibriSpeech上，在不使用语言模型的情况下，我们在测试其他语言时获得了6.8%的WER，在使用语言模型的浅层融合时，我们获得了5.8%的WER。这与之前最先进的7.5%的混合系统具有可比性。对于Switchboard，我们在Hub5’00测试集的Switchboard/CallHome部分实现了7.2%/14.6%，而没有使用语言模型，而在浅层融合中实现了6.8%/14.1%，这与之前最先进的混合系统的8.3%/17.3%相比。

索引词：端到端语音识别、数据扩充

# 1.     介绍

深度学习已经成功地应用于自动语音识别（ASR）[1]，其中研究的重点是设计更好的网络架构，例如DNNs[2]、CNNs[3]、RNNs[4]和端到端模型[5、6、7]。然而，这些模型往往容易拟合过度，需要大量的训练数据[8]。

数据增强被提出作为一种方法来为ASR生成额外的训练数据。例如，在[9,10]中，为低资源语音识别任务增加了人工数据。在[11]中，声道长度标准化已被用于数据增强。在[12]中，噪声音频是通过将干净的音频与噪声音频信号叠加而合成的。在[13]中，速度扰动已应用于LVSCR任务的原始音频。[14]探讨了声学室模拟器的使用。文献[15,16]研究了关键词识别的数据扩充。特征drop-outs已被用于训练多流ASR系统[17]。更普遍地说，所学的增强技术已经探索了不同的增强变换序列，这些序列已经在图像领域实现了最先进的性能[18]。

受最近在语音和视觉领域的成功增强的启发，我们提出了SpecAugment，一种基于输入音频的对数mel谱图而不是原始音频本身的增强方法。这种方法简单，计算成本低，因为它直接作用于log-mel谱图，就像它是一幅图像一样，并且不需要任何额外的数据。因此，我们可以在训练期间在线应用SpecAugment。SpecAugment由对数mel谱图的三种变形组成。第一种是时间扭曲，时间序列在时间方向上的变形。另外两个增强，受计算机视觉[19]中提出的“剪切”的启发，是时间和频率掩蔽，我们掩蔽了一块连续的时间步长或mel频率通道。

这种方法虽然还很初级，但非常有效，可以让我们训练端到端的ASR网络，称为Listen-attent-and-Spell（LAS）[6]，从而超越更复杂的混合系统，甚至在不使用语言模型（LMs）的情况下也能获得最先进的结果。在LibriSpeech[20]上，我们在testclean集上实现了2.8%的字错误率（WER），在testother集上实现了6.8%的字错误率（WER），而不使用LM。通过对LibriSpeech LM语料库中的LM进行浅层融合[21]，我们能够提高我们的性能（test clean和testother的性能分别为2.5%和5.8%），相对而言，testother的当前技术水平提高了22%。在Switchboard 300h（LDC97S62）[22]上，在不使用LM的情况下，我们在Hub5’00（LDC2002S09，LDC2003T02）测试集的Switchboard部分获得7.2%的WER，在CallHome部分获得14.6%的WER。在对Switchboard和Fisher（LDC200{4,5}T19）[23]语料库的联合转录本进行浅层融合后，我们在Switchboard/CallHome部分获得了6.8%/14.1%的WER。

# 2.     扩充政策

我们的目标是构造一个直接作用于log-mel谱图的增广策略，帮助网络学习有用的特征。出于这些特征对时间方向上的变形、频率信息的部分丢失和小片段语音的部分丢失具有鲁棒性的目标，我们选择了以下变形作为策略：

1.   时间扭曲是通过tensorflow的sparse\_image\_warp函数来实现的。给定一个带有*τ*时间步长的对数mel谱图，我们将其视为时间轴水平，频率轴垂直的图像。在时间步长（W，τ−W）内，沿穿过图像中心的水平线的随机点向左或向右扭曲，扭曲距离从0到沿该线的时间扭曲参数的均匀分布中选择。我们在边界的四个角和垂直边的中点上固定六个锚定点。*w型W型*

2.   应用频率掩蔽使得连续的mel频率信道[f0，f0+f]被掩蔽，其中*f*首先从0到频率掩码参数*F*的均匀分布中选择，然后*f*0从[0，ν−f）中选择。*ν*是mel频率通道数。

3.   应用时间掩蔽使得连续时间步[t0，t0+t）被掩蔽，其中*t*首先从0到时间掩蔽参数*T*的均匀分布中选择，*t*0从[0，τ-t]中选择。我们在时间掩码上引入了一个上限，使得时间掩码的宽度不能超过时间步数的p倍。

图1显示了应用于单个输入的单个增强的示例。对数mel谱图被归一化为具有零平均值，因此将屏蔽值设置为零相当于将其设置为平均值。

我们可以考虑应用多个频率和时间掩码的策略。多个遮罩可能重叠。在这项工作中，我们主要考虑了一系列手工制作的策略，LibriSpeech basic（LB）、LibriSpeech double（LD）、switchboard mild（SM）和switchboard strong（SS），其参数总结在表1中。在图2中，我们展示了一个使用LB和LD策略增强的log-mel谱图的示例。

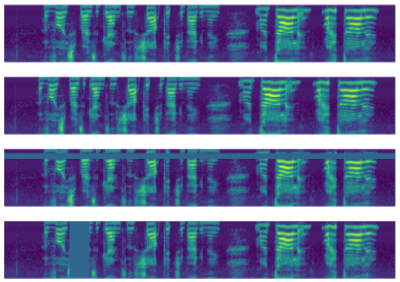
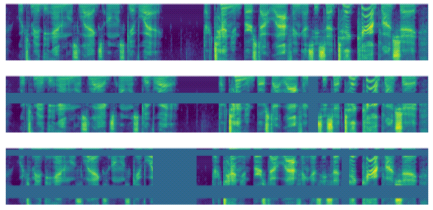
 

图1：从上到下，没有应用增强、时间扭曲、频率掩蔽和时间掩蔽。

图2：应用于基本输入的扩充策略。从上到下，这些图描绘了None、LB和LD的基本输入的对数mel谱图。

表1：策略的扩充参数。*mF和mT表示应用的频率和时间掩码的数量。*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 策略 | *W* | *F* | *mF* | *T* | *p* | *mT* |
| None | 0 | 0 | - | 0 | - | - |
| LB | 80 | 27 | 1 | 100 | 1.0 | 1 |
| LD | 80 | 27 | 2 | 100 | 1.0 | 2 |
| SM | 40 | 15 | 2 | 70 | 0.2 | 2 |
| SS | 40 | 27 | 2 | 70 | 0.2 | 2 |

# 3.     模型

我们使用LAS网络[6]来完成ASR任务。这些模型是端到端的，易于训练，而且还有一个额外的好处，那就是我们有充分记录的基准[24，25]，我们可以在此基础上得到我们的结果。在本节中，我们将回顾LAS网络并介绍一些参数化它们的符号。我们还介绍了我们用来训练网络的学习速率计划，因为它们是决定性能的一个重要因素。最后，我们回顾了shallowfusion[21]，我们使用它来合并语言模型以进一步提高性能。

## 3.1.    LAS网络体系结构

我们使用Listen，attent和Spell（LAS）网络[6]来表示在[25]中学习的端到端ASR，我们使用符号LAS-d-w。输入的对数mel谱图被传递到一个2层卷积神经网络（CNN），使用max-pooling且步长为2。CNN的输出通过d个单元大小为w的双向lstm组成的编码器，产生一系列的注意向量。注意向量被输入到w细胞维度的2层RNN解码器中，该解码器产生转录本的标记。文本使用Word Piece Model（WPM）[26]进行标记，LibriSpeech的词条大小为16k，switchboard的词条大小为1k。librispeech960h的WPM是使用训练集转录本构建的。对于switchboard300小时任务，将训练集的文本与Fisher语料库的文本相结合，构建WPM。最终的转录本通过波束大小为8的波束搜索获得。为了与[25]进行比较，我们注意到它们在我们的符号中的“大模型”是LAS-4-1024。

## 3.2.    学习率表

学习速率调度是决定ASR网络性能的一个重要因素，尤其是当存在增广时。这里，我们介绍的培训计划有两个目的。首先，我们使用这些调度来验证更长的调度是否可以提高网络的最终性能，而在增强时更是如此（表2）。其次，基于此，我们引入了非常长的调度，用于最大化网络的性能。

我们使用一个学习率计划，在这个计划中，我们提高，保持，然后指数衰减学习率，直到它达到最大值。学习速度保持不变，超过这一点。此时间表由三个时间戳（sr、si、sf）参数化–爬升（从零学习率）完成的步骤、指数衰减开始的步骤和指数衰减停止的步骤。1*/*100 *锶硅平方英尺*

在我们的实验中，还有两个因素引入了时间尺度。首先，我们在步骤中打开标准偏差为0.075的可变权重噪声[27]，并在整个训练过程中保持不变。在步长间隔（sr，si）中引入权值噪声，即在学习率的高平稳期。*s码*噪音

其次，我们引入了不确定性为0.1的均匀标签平滑[28]，即正确的类标签被赋予置信度0.9，而其他标签的置信度也相应增加。正如后面再次讨论的，标签平滑可以破坏较小学习率的训练，我们有时选择只在训练开始时打开它，在学习率开始衰减时关闭它。

我们使用的两个基本时间表如下所示：

1.      B（asic）：（sr，s）=（0.5k，10k，20k，80k）噪音*，平方英尺，平方英尺*

2.      D（双倍）：（sr，s）=（1k，20k，40k，160k）噪音*，平方英尺，平方英尺*

如第5节进一步讨论的，我们可以通过使用更长的时间表来提高训练网络的性能。因此，我们推出以下时间表：

3.      L（ong）：（sr，s）=（1k，20k，140k，320k）我们用来训练最大的模型来提高性能。当使用调度L时，对于LibriSpeech 960h，时间步长=140k引入不确定度为0.1的标签平滑，随后关闭。对于交换机300小时，标签平滑在整个培训过程中启用。噪音*，平方英尺，平方英尺<硅*

## 3.3.    语言模型的浅层融合

虽然我们可以通过扩充获得最先进的结果，但是我们可以通过使用语言模型得到进一步的改进。因此，我们通过两个任务的浅层融合合并了RNN语言模型。在浅层融合中，解码过程中的“下一个令牌”由**是的**∗

i、 例如，通过使用基本ASR模型和语言模型对令牌进行联合评分。我们还使用覆盖率惩罚[29]。*c级*

对于LibriSpeech，我们使用了一个嵌入维数为1024的两层RNN，在[25]中用于LM，它是在LibriSpeech LM语料库上训练的。我们始终使用[25]中使用的相同融合参数（=0.35和=0.05）。*λc级*

对于交换机，我们使用嵌入维数为256的两层RNN，它是在Fisher和交换机数据集的组合转录本上训练的。通过在RT-03（ldc207s10）上的性能测试，我们通过网格搜索找到融合参数。我们将在第4.2节讨论个别实验中使用的融合参数。

# 4.     实验

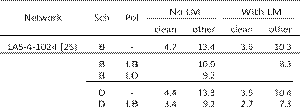
在这一节中，我们将用SpecAugment描述我们在LibriSpeech和switchember上的实验。我们报告了最新的研究结果，其性能超过了大量设计的混合动力系统。

## 4.1.    藏语960h

对于LibriSpeech，我们使用与[25]相同的设置，其中我们使用80维滤波器组，具有delta和delta-delta加速，以及16k字块模型[26]。

三个网络LAS-4-1024、LAS-6-1024和LAS-61280在LibriSpeech 960h上进行训练，并结合了增强策略（None、LB、LD）和训练计划（B/D）。在这些实验中没有应用标签平滑。实验在32个Google云TPU芯片上进行，峰值学习率为0.001，批量大小为512，持续7天。除了增广策略和学习速率计划之外，所有其他超参数都是固定的，没有应用额外的调整。我们在表2中报告了由dev-other集验证的测试集编号。我们发现，扩充可以持续地提高训练网络的性能，而且更大的网络和更长的学习速率时间表的好处在更严厉的扩充中更为明显。

我们采用最大的网络LAS-6-1280，使用schedule L（训练时间∼24天）和policy LD对网络进行训练，以最大限度地提高性能。如前所述，我们为时间步140k启用标签平滑。测试集性能是通过评估具有最佳dev-other性能的检查点来报告的。最先进的性能是通过LAS-6-1280模型实现的，即使没有语言模型。我们在表2：LibriSpeech测试WER（%）中评估了不同的网络、时间表和策略。[25]的第一行。*<*

                                                       3.7                                3.4

                                                       3.6                                2.8                  7.5

拉斯维加斯4-1024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | D级 | LD公司 | 3.4 | 8.3 | 2.8 | 6.8 |
| 拉斯维加斯6-1024 | D级  D级 | -  磅 | 4.5  3.4 | 13.1 8.6 | 3.6  2.6 | 10.3 6.7 |
|  | D级 | LD公司 | 3.2 | 8.0 | 2.6 | 6.5 |
| 拉斯维加斯6-1280 | D级  D级 | -  磅 | 4.3  3.4 | 12.9 8.7 | 3.5  2.8 | 10.5 7.1 |
|  | D级 | LD公司 | 3.2 | 7.7 | 2.7 | 6.5 |

可以使用浅层融合合并LM以进一步提高性能。结果见表3。

表3:LibriSpeech 960h WERs（%）。



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 清洁的 | 其他 | 清洁的 | 其他 |
| 隐马尔可夫模型  Panayotov等人，（2015）[20] |  |  | 5.51 | 13.97 |
| Povey等人，（2016）[30] |  |  | 4.28 |  |
| Han等人，（2017）[31] |  |  | 3.51 | 8.58 |
| Yang等人（2018）[32] |  |  | 2.97 | 7.50 |
| CTC/ASG公司  Collobert等人，（2016）[33] | 7.2 |  |  |  |
| Liptchinsky等人，（2017）[34] | 6.7 | 20.8 | 4.8 | 14.5 |
| 周等，（2018）[35] |  |  | 5.42 | 14.70 |
| Zeghidour等人，（2018）[36] |  |  | 3.44 | 11.24 |
| 李等，（2019）[37] | 3.86 | 11.95 | 2.95 | 8.79 |
| 拉斯维加斯  Zeyer等人，（2018）[24] | 4.87 | 15.39 | 3.82 | 12.76 |
| Zeyer等人，（2018）[38] | 4.70 | 15.20 |  |  |
| Irie等人，（2019）[25] | 4.7 | 13.4 | 3.6 | 10.3 |
| Sabour等人，（2019）[39] | 4.5 | 13.3 |  |  |
| 我们的工作是 | 4.1 | 12.5 | 3.2 | 9.8 |
| LAS+规格 | 2.8 | 6.8 | 2.5 | 5.8 |

## 4.2.    配电盘300h

对于交换机300h，我们使用Kaldi[40]“s5c”方法来处理数据，但是我们调整了方法，使用80维滤波器组，具有delta和delta加速。我们使用1k WPM[26]来标记输出，它是使用switched和Fisher转录本的组合词汇构建的。

我们使用策略（None、SM、SS）和时间表B训练LAS-4-1024。和以前一样，我们将峰值学习率设置为0.001，总批量大小设置为512，并使用32个Google Cloud TPU芯片进行训练。在这里的实验是运行和不标签平滑。由于没有规范的开发集，我们选择在训练计划的终点评估检查点，对于计划B，我们选择10万步。我们注意到，在衰退计划完成后，训练曲线松弛（step），网络的性能变化不大。表4显示了交换台300h在有标签平滑和无标签平滑的情况下各种扩充策略的性能。我们看到标签平滑和增强对这个语料库有加性的影响。*平方英尺*

表4:LAS-4-1024的交换机300h功率（%）评估，采用不同的增强和标签平滑（LS）策略，按照附表B进行培训。没有使用LMs。

|  |  |  |  |
| --- | --- | --- | --- |
| 政策 | 长征 | 开关柜 | 中国 |
|  |  | 9.7 | 18.2 |
|  |  | 8.6 | 16.3 |

与LibriSpeech 960h一样，我们在交换机300小时训练集上训练LAS-6-1280，并安排L（训练时间∼24天），以获得最先进的性能。在这种情况下，我们发现在整个训练过程中启用标签平滑有利于最终性能。我们报告了340k步在训练结束时的表现。我们在表5中的其他工作中给出了我们的结果。我们也使用Fisher交换台上训练的LM进行浅层融合，其融合参数是通过在RT-03语料库上评估性能得到的。与LibriSpeech的情况不同，融合参数在不同训练的网络之间传递不好。表5中的三个条目分别使用融合参数（λ，c）=（0.3，0.05），（0.2，0.0125）和（0.1，0.025）获得。

表5：交换机300h WERs（%）。



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 开关柜 | 中国 | 开关柜 | 中国 |
| 隐马尔可夫模型  Vesely等人，（2013）[41]&apos; |  |  | 12.9 | 24.5 |
| Povey等人，（2016）[30] |  |  | 9.6 | 19.3 |
| Hadian等人，（2018）[42] |  |  | 9.3 | 18.9 |
| Zeyer等人，（2018）[24] |  |  | 8.3 | 17.3 |
| 反恐委员会  Zweig等人，（2017）[43] | 24.7 | 37.1 | 14.0 | 25.3 |
| Audhkhasi等人，（2018）[44] | 20.8 | 30.4 |  |  |
| Audhkhasi等人，（2018）[45] | 14.6 | 23.6 |  |  |
| 拉斯维加斯  吕等，（2016）[46] | 26.8 | 48.2 | 25.8 | 46.0 |
| Toshniwal等人，（2017）[47] | 23.1 | 40.8 |  |  |
| Zeyer等人，（2018）[24] | 13.1 | 26.1 | 11.8 | 25.7 |
| 翁等，（2018）[48] | 12.2 | 23.3 |  |  |
| Zeyer等人，（2018）[38] | 11.9 | 23.7 | 11.0 | 23.1 |
| 我们的工作是 | 11.2 | 21.6 | 10.9 | 19.4 |
| LAS+规格（SM） | 7.2 | 14.6 | 6.8 | 14.1 |
| LAS+规格（SS） | 7.3 | 14.4 | 7.1 | 14.0 |

# 5.     讨论

时间扭曲有助于提高性能，但不是提高性能的主要因素。在表6中，我们给出了分别关闭时间扭曲、时间掩蔽和频率掩蔽的三个训练结果。我们看到时间扭曲效应虽然很小，但仍然存在。考虑到预算的限制，时间扭曲（Time warping）应该是第一个被放弃的扩充，它是本研究中讨论的扩充中最昂贵也是影响最小的。

表6：在没有LM的情况下评估的测试集WER（%），用于按照附表B培训的网络LAS-4-1024。







标签平滑会给训练带来不稳定性。我们注意到，当标签平滑应用于增广时，LibriSpeech的不稳定训练运行的比例增加。当学习率下降时，这一点变得更加明显，因此我们引入了一个标签平滑计划来训练LibriSpeech，其中标签仅在学习率计划的初始阶段平滑。

增广将过拟合问题转化为欠拟合问题。从图3中网络的训练曲线可以看出，训练过程中的网络不仅在增广训练集上的loss和WER下拟合，而且在对增广数据进行训练时也在训练集本身上拟合。这与通常的情况形成了鲜明的对比，在这种情况下，网络往往过于适合于训练数据。这是增强训练的主要好处，如下所述。

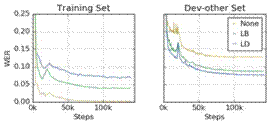


图3:LAS-6-1280关于LibriSpeech的附表D。

处理欠拟合产量改进的常用方法。我们能够通过标准的方法来减轻不适合的情况，从而在性能上取得显著的进步——建立更大的网络和延长培训时间。当前报告的性能是通过递归过程获得的，即应用严格的增强策略，然后生成更宽、更深的网络，并使用更长的时间表对其进行训练，以解决不适合的问题。

对相关工作的评论。我们注意到，在[49]中，CNN声学模型的上下文中研究了类似于频率掩蔽的增强。在那里，相邻频率的块被预先分组到箱子中，每个小批量随机清零。另一方面，SpecAugment中频率掩模的大小和位置都是随机选择的，并且每个输入都不同。文献[50]讨论了在结构上省略频谱图频率数据的更多想法。

# 6.     结论

SpecAugment极大地提高了ASR网络的性能。通过使用简单的手工策略扩充训练集，我们能够在端到端的LAS网络上获得LibriSpeech 960h和switchember 300h任务的最新结果，即使没有语言模型的帮助，也超过了混合系统的性能。SpecAugment将ASR从过拟合问题转化为欠拟合问题，我们能够通过使用更大的网络和更长的训练时间来获得性能。

致谢：我们要感谢曹元元、切尔巴、伊莉、叶佳、坎南、帕特里克

Nguyen、Vijay Peddinti、Rohit Prabhavalkar、Wu Yonghui和Shuyuan Zhang进行了有益的讨论。我们还要感谢久尔基·科瓦茨为我们介绍了这些作品

# 7.     工具书类

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury et al., “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal processing magazine, vol. 29, 2012.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, Jan 2012.

[3] T. Sainath, A. rahman Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep Convolutional Neural Networks for LVCSR,” in ICASSP, 2013.

[4] A. Graves, A. rahman Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in ICASSP, 2013.

[5] A. Graves and N. Jaitly, “Towards End-to-End Speech Recognition with Recurrent Neural Networks,” in ICML, 2014.

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition,” in ICASSP, 2016.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-End Attention-based Large Vocabulary Speech Recognition,” in ICASSP, 2016.

[8] C.-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, and M. Bacchiani, “State-of-the-art Speech Recognition With Sequence-to-Sequence Models,” in ICASSP, 2018.

[9] N. Kanda, R. Takeda, and Y. Obuchi, “Elastic spectral distortion for low resource speech recognition with deep neural networks,” in ASRU, 2013.

[10] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, “Data augmentation for low resource languages,” in INTERSPEECH, 2014.

[11] N. Jaitly and G. Hinton, “Vocal Tract Length Perturbation (VTLP) improves speech recognition,” in ICMLWorkshop on Deep Learning for Audio, Speech and Language Processing, 2013.

[12] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, “Deep Speech: Scaling up end-to-end speech recognition,” in arXiv, 2014.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio Augmentation for Speech Recognition,” in INTERSPEECH, 2015.

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

[15] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, “Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks,” in ICASSP, 2015.

[16] A. Raju, S. Panchapagesan, X. Liu, A. Mandal, and N. Strom, “Data Augmentation for Robust Keyword Spotting under Playback Interference,” in arXiv, 2018.

[17] S. H. Mallidi and H. Hermansky, “Novel neural network based fusion for Multistream ASR,” in ICASSP, 2016.

[18] E. D. Cubuk, B. Zoph, D. Man´e, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation policies from data,” in CVPR, 2019.

[19] T. DeVries and G. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout,” in arXiv, 2017.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in ICASSP, 2015.

[21] C¸ . G¨ulc¸ehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, “On using monolingual corpora in neural machine translation,” in arxiv, 2015.

[22] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: telephone speech corpus for research and development,” in ICASSP, 1992.

[23] C. Cieri, D. Miller, and K. Walker, “The fisher corpus: a resource for the next generations of speech-to-text,” in LREC, 2004.

[24] A. Zeyer, K. Irie, R. Schl¨uter, and H. Ney, “Improved training of end-to-end attention models for speech recognition,” in INTERSPEECH, 2018.

[25] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen, “Model Unit Exploration for Sequence-to-Sequence Speech Recognition,” in arXiv, 2019.

[26] M. Schuster and K. Nakajima, “Japanese and korean voice search,” in ICASSP, 2012.

[27] A. Graves, “Practical Variational Inference for Neural Networks,” in NIPS, 2011.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z.Wojna, “Rethinking the inception architecture for computer vision,” in CVPR, 2016.

[29] J. Chorowski and N. Jaitly, “Towards better decoding and language model integration in sequence to sequence models,” in INTERSPEECH, 2017.

[30] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in INTERSPEECH, 2016.

[31] K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane, “The CAPIO 2017 Conversational Speech Recognition System,” in arXiv, 2017.

[32] X. Yang, J. Li, and X. Zhou, “A novel pyramidal-FSMN architecture with lattice-free MMI for speech recognition,” in arXiv, 2018.

[33] R. Collobert, C. Puhrsch, and G. Synnaeve, “Wav2Letter: an Endto- End ConvNet-based Speech Recognition System,” in arXiv, 2016.

[34] V. Liptchinsky, G. Synnaeve, and R. Collobert, “Letter-Based Speech Recognition with Gated ConvNets,” in arXiv, 2017.

[35] Y. Zhou, C. Xiong, and R. Socher, “Improving End-to-End Speech Recognition with Policy Learning,” in ICASSP, 2018.

[36] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, “Fully Convolutional Speech Recognition,” in arXiv, 2018.

[37] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, “Jasper: An End-to-End Convolutional Neural Acoustic Model,” in arXiv, 2019.

[38] A. Zeyer, A. Merboldt, R. Schl¨uter, and H. Ney, “A comprehensive analysis on attention models,” in NIPS: Workshop IRASL, 2018.

[39] S. Sabour, W. Chan, and M. Norouzi, “Optimal Completion Distillation for Sequence Learning,” in ICLR, 2019.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi Speech Recognition Toolkit,” in ASRU, 2011.

[41] K. Vesely, A. Ghoshal, L. Burger, and D. Povey, “Sequencediscriminative training of deep neural networks,” in INTERSPEECH, 2013.

[42] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, “End-to-end speech recognition using lattice-free MMI,” in INTERSPEECH, 2018.

[43] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, “Advances in All- Neural Speech Recognition,” in ICASSP, 2017.

[44] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, “Direct Acoustics-to-Word Models for English Conversational Speech Recognition,” in INTERSPEECH, 2018.

[45] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. Picheny, “Building competitive direct acoustics-to-word models for english conversational speech recognition,” in ICASSP, 2018.

[46] L. Lu, X. Zhang, and S. Renals, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition,” in ICASSP, 2016.

[47] S. Toshniwal, H. Tang, L. Lu, and K. Livescu, “Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition,” in INTERSPEECH, 2017.

[48] C.Weng, J. Cui, G.Wang, J.Wang, C. Yu, D. Su, and D. Yu, “Improving Attention Based Sequence-to-Sequence Models for Endto- End English Conversational Speech Recognition,” in INTERSPEECH, 2018.

[49] G. Kov´acs, L. T´oth, D. Van Compernolle, and S. Ganapathy, “Increasing the robustness of cnn acoustic models using autoregressive moving average spectrogram features and channel dropout,” Pattern Recognition Letters, vol. 100, pp. 44–50, 2017.

[50] L. T´oth, G. Kov´acs, and D. Van Compernolle, “A perceptually inspired data augmentation method for noise robust cnn acoustic models,” in SPECOM, 2018.

[[1]](" \l "_ftnref1" \o ")作为谷歌人工智能住院医师项目的成员所做的工作。