规格放大数据集

*丹尼尔·S·帕克*∗*，张宇*∗*，钟正秋，陈友政，李波，陈伟霆，郭伟乐，吴永辉*

谷歌公司。

# 摘要

最近，SpecAugment，一种直接作用于输入语音谱图的自动语音识别增强方案，在提高端到端网络在公共数据集上的性能方面显示出了非常有效的效果。在本文中，我们通过研究其在Google多域数据集（Narayanan等人，2018）中的应用，证明了其在大规模数据集任务中的有效性。在训练声学模型时，我们将原始训练数据与SpecAugment和噪声扰动的训练数据混合在一起，从而在所有的测试域中实现改进。我们还引入了SpecAugment的一个修改，它根据话语的长度调整时间掩码的大小和/或多重性，这可能有利于大规模任务。通过使用自适应掩蔽，我们能够进一步提高LibriSpeech上的Listen，attent和Spell模型的性能，在test clean上提高2.2%，在test other上提高5.2%。

*索引扩展*-端到端语音识别，数据增强，多域训练

# 1导言

在自动语音识别（ASR）中，数据增强是提高泛化性能的一种成功方法。最近，SpecAugment[1]作为一种直接增强输入语音谱图的增强方案，在960h Librispeech和300h交换台数据集上显示出惊人的提高ASR网络性能的效果。出现的一个自然问题是SpecAugment的有效性是否适用于大规模任务。

在本文中，我们将SpecAugment应用于[2]中介绍的Google多域数据集来解决这个问题。Google多域数据集是一个大规模的多域数据集，包含来自不同域的多个测试集。数据集中的所有数据都是匿名的。我们比较了经过训练的网络在各种形式的数据扩充方面的性能，其结果总结在表1中。在[2]中，多风格训练（MTR）[3]，其中混合室模拟器用于将干净的音频与大量噪声音频库相结合，用于增强输入数据。在研究SpecAugment的性能时，我们以此为基准。

表1。各种形式扩充的结果。

|  |  |
| --- | --- |
| 数据扩充 | 性能 |
| 多方式培训（MTR） | 基线 |
| 没有 | 更糟 |
| 规格增大 | 更好[[1]](" \l "_ftn1" \o ") |
| 规格+地铁 | 更糟 |
| 混合规格和MTR | 更好 |

如表1所示，我们比较了在干净数据、应用MTR的数据、应用SpecAugment的数据、同时应用SpecAugment和MTR的数据以及混合SpecAugment和MTR数据时网络的性能。我们发现，SpecAugment应用于干净数据时，在所有自然测试集上的性能都优于基线，而在通过MTR测试话语获得的合成测试集上的性能更差。令我们惊讶的是，在MTR之上应用SpecAugment会降低大多数域的性能。同时，通过混合spec增强数据和MTR数据，我们能够在所有领域实现改进。

SpecAugment需要的额外计算资源可以忽略不计，不需要额外的音频数据，可以在线应用，因此随着训练集变得越来越大，它具有很高的可伸缩性。因此，我们的结果表明，SpecAugment可以被认为是一个严重的替代更复杂的资源密集型扩增方法。

具体策略包括频率询问、时间掩蔽和时间扭曲。文[1]中考虑的增广策略具有固定数量的时间掩码，而与话语的长度无关。在跨多个领域的大规模任务中，我们期望话语的长度会有很大的差异。因此，我们引入了自适应时间掩蔽，其中时间掩蔽的数量和/或时间掩蔽的大小根据输入的长度而变化。我们在googlemultido上试验了几种自适应策略-

|  |
| --- |
| 2020年IEEE。允许个人使用此材料。在任何当前或未来的媒体中，所有其他用途必须获得IEEE的许可，包括为广告或促销目的重新印刷/重新发布本材料，创建新的集体作品，转售或重新分发到服务器或列表，或在其他作品中重新使用本作品的任何受版权保护的部分。 |

主数据集和LibriSpeech 960h[4]。到目前为止，我们还没有

在Google多域数据集上发现了性能优于vanilla SpecAugment的自适应时间策略。同时，我们发现，相对于[1]，LibriSpeech上的自适应策略可以提高性能，因为我们能够训练一个Listen、attent和Spell[5]网络，使其在测试clean时的性能提高2.2%，在测试other时的性能提高5.2%。

1.1. 相关工作

关于ASR中的增强有大量的文献，我们在这里只调查了其中的一部分。文献[6,7]研究了低资源语音识别任务中的人工数据增强。[8]在ASR数据增强的背景下引入了声道长度扰动，并在[9]中作了进一步探讨。在[10]中，噪声音频信号被用于增强。速度扰动[11]是语音数据增强的一个组成部分。工作[3]研究了使用声学室内模拟器的效果。论文[12，13]研究了数据增强在关键词识别中的应用。[14,15]中的特征退出已用于训练多流ASR系统。在CNN ASR网络的背景下，对输入频谱图的频率通道的系统性遗漏进行了研究[16,17]。我们在引言中对SpecAugment[1]进行了评论。

数据扩充也成功地应用于大规模工业数据集。如前所述，多风格训练（MTR）是一种流行的技术，使用室内模拟器将干净的音频与背景噪声相结合[3]。MTR已成功应用于基于HMM的系统[18,19]和端到端LAS模型[5,20,21]。一个自然的问题是SpecAugment如何与MTR这样的现有数据增强技术进行比较或补充，尤其是在大规模数据集上。

我们在本文中的贡献有三个方面：

1.    我们将spec扩展到大型工业数据集。我们比较了现有的地铁数据增强，并提出了我们可以如何改进它。

2.    我们证明SpecAugment提高了流模型的性能。

3.    提出了一种自适应的SpecAugment算法，其时间掩蔽度与输入序列长度相适应。

2spec增强与自适应掩蔽

在本节中，我们将简要回顾SpecAugment，并介绍其自适应变体。通过时间扭曲、频率掩蔽和时间掩蔽三种基本增强策略的组合，得到了一种spec增强策略。我们将频谱图的时间和频率维度表示为和。*τν*

1.   带参数的时间扭曲：位移从−W到的均匀分布中选择。从时间间隔[W，τ−W]中选择一个起点。定义线性翘曲函数W（t），使起点映射到点+W，且边界点=0和=τ−1固定：*W型w型W型w型*0 *w型*0 *w型*0 *t型t型*

*,*

扭曲的定义使得扭曲特征（t）（在我们的例子中，对数mel频率系数）在时间上与原始特征（t）通过扭曲（W（t））=xorig（t）相关。**xxx号**弯曲*t型*原始

我们注意到，在[1]中提出的时间扭曲的原始实现，出于所有实际目的，等同于这个替代定义。

2.   带参数的频率掩模：掩模尺寸从0到0的均匀分布中选择。然后屏蔽连续的对数mel频率通道[f0，f0+f），其中从[0，ν−f]中选择。*F级f级F级f级*0

3.   带参数的时间掩模：掩模尺寸从0到0的均匀分布中选择。连续时间步[t0，t0+t）被掩蔽，其中从[0，τ−t）中选择。*T型t型T型t型*0

[1]中的SpecAugment策略由应用这三个扩展的固定次数组成。

在包含不同输入域的大规模数据集中，我们期望输入音频的长度会有很大的差异。因此，固定数量的时间掩蔽对于此类任务可能是不够的，因为时间掩蔽对于较长的话语可能太弱，或者对于较短的话语可能太严重。因此，我们介绍了两种不同的方法，时间掩蔽可以根据光谱图的长度进行自适应：*τ*

1.   自适应多重性：时间掩码的个数或多重性设置为＝⌊pM·τ⌋，表示多重性比率。*米*t型面罩*米*t型面罩*下午*

2.   自适应大小：时间掩码参数设置为=⌊pS·τ⌋大小比。*T型pS公司*

在本文中，当使用自适应时间掩蔽时，我们将时间掩蔽的数量限制为20个，这是由*米*t型面罩

*米*t型面罩=最小值（20，⌊pM·τ⌋）。

# 三。实验

3.1. 藏语960h

## 3.1.1. 设置

我们对librispeech960h的设置基于[1]。我们使用的是LAS-6-1280型号的那项工作，训练计划是“L”（ong）。我们将浅层融合[22]与LSTM语言模型（LM）结合使用，该模型具有两个融合参数——LM权重和覆盖惩罚[23]。在这项工作中，我们使用了一个宽度为4096的三层LSTM，在dev set转录本上产生了63.6的wordlevel困惑。我们使用网格搜索调整dev集上的融合参数，并将其应用于测试集以报告最终结果。

## 3.1.2. 自适应扩展策略

我们比较了三种增强策略。基准策略是[1]中创造的“LibriSpeech Double”策略。此策略有两个频率掩码=27，两个时间掩码=100，在时间扭曲=80后应用。让我们介绍一个手工构建的自适应策略，我们称之为LibriFullAdapt。此策略有两个频率掩码应用，分别为=27和时间掩码，其中自适应多重性和大小分别为=0.04和=0.04，时间扭曲应用于=80。*F级T型W型F级下午pS公司W型*

## 3.1.3. 结果

我们在表2中列出了培训的结果。我们发现自适应策略的性能优于固定策略，并且在与语言模型进行浅层融合之前和之后都观察到了性能的提高。

表2。LibriSpeech 960h WERs（%）。



方法无LM带LM



                                                         清洁其他清洁其他

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 近期工作说明书[1] | 2.8 | 6.8 | 2.5 | 5.8 |
| Lu¨scher等人，（2019）[24] |  |  | 2.3 | 5.0 |
| Kim等人，（2019）[9] |  |  | 2.4 | 8.3 |
| Karita等人，（2019）[25] |  |  | 2.6 | 5.7 |
| Han等人，（2019）[26] |  |  | 2.2 | 5.8 |
| 这项工作  LAS+基线规格 | 2.8 | 6.8 | 2.4 | 5.7 |
| LAS+LibriFullAdapt公司 | 2.6 | 6.0 | 2.2 | 5.2 |

3.2. Google多域数据集

## 3.2.1. 数据和扩充

我们研究了在Google多域数据集上训练时SpecAugment的效果[2]。我们考虑了五个测试集Search、Search noise、TTS Audiobook、Telephony和YouTube来衡量网络的性能。所有训练和测试数据都是匿名的。

作为我们实验的基线，我们通过使用[3]中描述的房间模拟器来增加输入数据。训练时，随机选择房间模拟器的各种因素，包括房间大小、混响时间、麦克风位置、语音和噪声源、信噪比，并应用于所有输入语音。注入的噪声是从匿名的YouTube音频或现实生活中的噪声集合中采样的。通过将这些扰动应用于搜索测试集，构造了测试集搜索噪声。

网络输入是一个对数mel频谱图，使用32毫秒的帧窗口和10毫秒的位移从音频中获取。log-mel频率系数有128个维度，高度为512，步长为3。文本使用词汇量为4k的词块模型（WPM）[27]进行标记化。

我们考虑了五种不同的输入配置：MTR数据、clean数据、应用SpecAugment的MTR数据、应用SpecAugment的clean数据，最后将clean数据与应用SpecAugment的MTR数据以8:2的比例混合得到数据。在分解特征后对谱图进行增强，得到128维特征的阵列。然后将增强后的频谱图重新压缩为原始形式并输入声学模型。

我们提出了一个香草SpecAugment政策，我们表示SpecAugBasic培训的结果。此策略有两个频率掩码和两个时间掩码，其中=50。没有使用时间扭曲。作为一个控制实验，我们还训练了仅使用两个=27的掩模的频率掩模增强数据的网络。*T型F级*

## 3.2.2. RNN传感器上的规范（RNN-T）

我们训练了[28]中描述的RNN-T模型。编码器是

一个单元大小为2048的8层单向LSTM，而解码器是具有相同单元大小的2层LSTM。没有使用语言模型。

我们注意到，由于其流的性质，该模型产生较弱的上下文信息。尽管如此，我们还是从时间掩蔽中获得了收益，正如我们很快演示的那样。

如[28]所述，我们的RNN-T模型严重依赖于层标准化[29]。需要注意的是，时间掩码的应用使得隐藏激活的方差消失，这在存在层规范化的情况下破坏了训练的稳定性。即使使用积极的方差下限，当网络变得更深时，这仍然会导致巨大的梯度。为了缓解这种不稳定性，我们在时间掩蔽区域中加入高斯噪声，从而稳定训练。

## 3.2.3. 结果

使用不同增强方法训练声学模型的结果如表3所示。请注意，当SpecAugment应用于MTR之上时，所有测试集的性能都会降低到基线以下。

同时，我们发现，当SpecAugBasic应用于干净的话语时，它在所有“自然测试集”上的表现都优于基线，而在用MTR搜索领域话语得到的合成测试集上，它的表现更差。然而，这种退化可以通过en来解决-

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 表3。各种扩充方案在Google多域数据集上的性能。   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 增强 | 搜索 | 搜索噪音 | TTS有声读物 | 电话 | YouTube网站 | | MTR（基线） | 6.6 | 9.6 | 4.6 | 7.8 | 11.8 | | 对照实验  清除数据 | 6.7 | 14.3 | 4.5 | 10.3 | 11.3 | | 基本+地铁 | 6.9 | 9.7 | 4.5 | 8.2 | 10.8 | | 仅频率屏蔽 | 6.4 | 13.4 | 4.8 | 8.0 | 11.4 | | 规格增大  SpecAugBasic清洁 | 6.2 | 12.9 | 4.2 | 7.2 | 10.3 | | SpecAugBasic和MTR（20%）混合 | 6.3 | 9.4 | 4.2 | 7.2 | 10.4 | |

将spec扩展数据与MTR数据合并，如表的最后一行所示。

我们注意到，虽然我们已经用自适应时间掩蔽策略进行了实验，但还没有发现比固定策略性能更好的策略。自适应时间掩蔽对这个数据集的好处还有待观察。

我们强调训练后的模型是一个流模型，其性能仍有显著提高。此外，我们还发现时间掩蔽在提高网络性能方面起着重要作用，这一点从YouTube数据集上的评估结果可以明显看出。

# 4总结与讨论

我们发现，SpecAugment尽管简单，但与经过时间检验和更复杂的增强方法相比，它在大规模数据集上产生更好的增益。鉴于SpecAugment的计算优势，我们发现它在工业规模任务的数据管道中有着巨大的潜力。

我们为SpecAugment引入了自适应时间掩蔽。在Google多域数据集上，我们还没有找到一种比非自适应策略更有效的自适应策略，但是我们已经在librispeech960h上证明了自适应掩蔽的有效性，我们期望在SpecAugment应用于大规模任务时，自适应掩蔽的进一步探索能够带来改进。

# 5确认

我们感谢曹元元、黄燕平、梅茨、纳拉亚南、彭若明、赛纳特、谢启哲和佐夫对我们实验的有益讨论和帮助。

# 6参考文献

[1] Daniel S.Park、William Chan、Yu Zhang、Chung Cheng Chiu、Barret Zoph、Ekin D.Cubuk和Quoc V.Le，“Specaugment:自动语音识别的简单数据增强方法”，Interspeech，2019年。

[2] Arun Narayanan、Ananya Misra、Khe Chai Sim、Goland Pundak、Anshuman Tripathi、Mohamed Elfeky、Parisa Haghani、Trevor Strohman和Michiel Bacchiani，“通过大规模培训实现域不变语音识别”，arXiv，2018年。

[3] Chanwoo Kim、Ananya Misra、Kean Chin、Thad Hughes、Arun Narayanan、Tara Sainath和Michiel Bacchiani，“在虚拟房间中生成大规模模拟话语，以便在Google Home中训练用于远场语音识别的深度神经网络”，Interspeech，2017年。

[4] Vassil Panayotov、Guogouchen、Daniel Povey和Sanjeev Khudanpur，“Librispeech:基于公共领域有声图书的ASR语料库”，ICASSP，2015年。

[5] William Chan、Navdeep Jaitly、Quoc V.Le和Oriol Vinyals，“倾听、参与和拼写：用于大词汇量会话语音识别的神经网络”，ICASSP，2016年。

[6] Naoyuki Kanda、Ryu Takeda和Yasunari Obuchi，“深度神经网络用于低资源语音识别的弹性频谱失真”，ASRU，2013年。

[7] Anton Ragni、Kate M.Knill、Shakti P.Rath和Mark J.F.Gales，“低资源语言的数据增强”，Interspeech，2014年。

[8] Navdeep Jaitly和Geoffrey Hinton，“声道长度扰动（VTLP）改善语音识别”，ICML音频、语音和语言处理深度学习研讨会，2013年。

[9] Chanwoo Kim、Minkyu Shin、Abhinav Garg和Dhananjaya Gowda，“为最先进的端到端语音识别系统改进声道长度扰动”，Interspeech，2019年。

〔10〕Awni Hannun、卡尔案、Jared Casper、Bryan Catanzaro、Greg Diamos、Erich Elsen、Ryan Prenger、Pig、Y、Y、和“2014”。

[11] Tom Ko、Vijayaditya Peddinti、Daniel Povey和Sanjeev Khudanpur，《语音识别的音频增强》，Interspeech，2015年。

[12] R.Prabhavalkar、R.Alvarez、C.Parada、P.Nakkiran和T.N.Sainath，“利用深度神经网络进行鲁棒小足迹关键词识别的自动增益控制和多方式训练”，ICASSP，2015年。

[13] Anirudh Raju、Sankaran Panchapagesan、Xing Liu、Arindam Mandal和Nikko Strom，“回放干扰下鲁棒关键词定位的数据增强”，arXiv，2018年。

[14] Sri Harish Mallidi和Hynek Hermansky，“基于神经网络的多流ASR融合”，ICASSP，2016年。

[15] Gyo¨rgy Kova´cs、La´szlo´To´th、Dirk Van Compernolle和Marcus Liwicki，“检验多频带处理和信道丢失的组合以实现鲁棒语音识别”，Interspeech，2019年。

[16] Gyo¨rgy Kova´cs、La´szlo´To´th、Dirk Van Compernolle和Sriram Ganapathy，“使用自回归移动平均频谱特征和信道丢失增强cnn声学模型的鲁棒性”，《模式识别快报》，第100卷，第44–50页，2017年。

[17] La´szlo´To´th、Gyo¨rgy Kova´cs和Dirk Van Compernolle，“噪声鲁棒cnn声学模型的感知启发数据增强方法”，SPECOM，2018年。

[18] Tara Sainath、Ron Weiss、Kevin Wilson、Andrew Senior和Oriol Vinyals，“使用原始波形CLDNS学习演讲前端”，Interspeech，2015年。

[19] Bo Li、Tara Sainath、Arun Narayanan、Joe Caroselli、Michiel Bacchiani、Ananya Misra、Izhak Shafran、Hasim Sak、Goland Pundak、Kean Chin、Khe Chai Sim、Ron Weiss、Kevin Wilson、Ehsan Variani、Chanwoo Kim、Olivier Siohan、Mitchel Weintraub、Erik McDermott、Rick Rose和Matt Shannon，“谷歌主页的声学建模”，Interspeech，2017

[20] Chung Cheng Chiu、Tara N.Sainath、Yonghui Wu、Rohit Prabhavalkar、Patrick Nguyen、Zhifeng Chen、Anjuli Kannan、Ron J.Weiss、Kanishka Rao、Ekaterina Gonina、Navdeep Jaitly、Bo Li、Jan Chorowski和Michiel Bacchiani，“序列到序列模型的最先进语音识别”，ICASSP，2018年。

[21]Chung Cheng Chiu、Anshuman Tripathi、Katherine Chou、Chris Co、Navdeep Jaitly、Diana Jaunzeikare、Anjuli Kannan、Patrick Nguyen、Hasim Sak、Ananth Sankar、Justin Tansuwan、Nathan Wan、Yonghui Wu和Xuedong Zhang，“医疗对话的语音识别”，Interspeech，2018年。

[22]C¸aglar Gu–lc¸ehre、Orhan Firat、Kelvin Xu、Kyunghyun Cho、LoıC Barrault、Huei Chi Lin、Fethi Bogares、Holger Schwenk和Yoshua Bengio，“关于在神经机器翻译中使用单语语料库”，arxiv，2015年。

[23]Jan Chorowski和Navdeep Jaitly，“在序列到序列模型中实现更好的解码和语言模型集成”，Interspeech，2017年。

[24]克里斯托夫·卢舍尔、尤金·贝克、卡祖基·伊里、马库斯

Kitza、Wilfried Michel、Albert Zeyer、Ralf Schlter和Hermann Ney，“图书馆演讲的RWTH ASR系统：混合vs注意力-无数据增强”，Interspeech，2019年。

[25]Shigeki Karita、Nanxin Chen、Tomoki Hayashi、Takaki Hori、Hirofumi Inaguma、Ziyan Jiang、Masao Someki、Nelson Enrique Yalta Soplin、Ryuichi Yamamoto、Xiaofie Wang、Shinji Watanabe、Takenori Yoshimura和Wangyou Zhang，“语音应用中变压器与rnn的比较研究”，arXiv，2019年。

[26]Kyu J.Han、Ramon Prieto、Kaixing Wu和Tao Ma，

“利用多流自我注意和扩展1d卷积的最新语音识别技术”，arXiv，2019年。

[27]Mike Schuster和Kaisuke Nakajima，“日语和韩语语音搜索”，ICASSP，2012年。

[28]何延章、塔拉·N·赛纳、罗希特·普拉巴瓦尔卡、伊恩·麦格劳、拉齐尔·阿尔瓦雷斯、丁昭、大卫·瑞巴赫、安居里·坎南、吴永辉、彭若明、乔

Liang、Deepti Bhatia、Yuan Shangguan、Bo Li、Golan Pundak、Khe Chai Sim、Tom Bagby、Shuo Yin Chang、Kanishka Rao和Alexander Grunstein，“移动设备的流式端到端语音识别”，ICASSP，2019年。

[29]Jimmy Lei Ba、Jamie Ryan Kiros和Geoffrey E Hinton，“层标准化”，arXiv，2016年。

[[1]](" \l "_ftnref1" \o ")平等贡献。