arXiv:2305.01170v1 [cs.SD] 2 May 2023

对比语音混合 面向低资源关键词检出的对比语音混合

Dianwen Ng[1,2], *Ruixi Zhang, Jia Qi Yip*[1,2],*Chong Zhang*[1],*Yukun Ma*[1],
Trung Hieu Nguyen[1], *Chongjia Ni*[1],*Eng Siong Chng*[2],*Bin Ma*[1]

[1]阿里巴巴集团[2]南洋理工大学计算机科学与工程学院，新加坡

摘要

现有的大多数智能设备关键词检出（KWS）基于神经网络的模型，都需要数千个训练样本才能学得良好的音频表示。然而，随着用户对智能设备个性化需求的不断增长，关键词检出模型需要能够快速适应数量较少的用户样本。为应对这一挑战，我们提出了一种面向低资源关键词语音识别的对比语音混合（CosMix）学习算法。CosMix 在现有的 Mixup 数据增强技术中引入辅助对比损失，以最大化原始预混合样本与增强样本之间的相对相似度。其目标是注入增强约束，引导模型从两个增强视图（即含噪混合语音片段和干净的预混合语音）中学习更简单但内容更丰富的语音表示。我们在 Google Speech Command 数据集上开展实验，并将训练集缩减至每个关键词仅 2.5 分钟，以模拟低资源条件。实验结果表明，该方法能够持续提升多个模型的性能，展现了其有效性。

索引词— 关键词检出，数据增强，对比学习，低资源

1. 引言

许多智能设备内置的语音命令系统为我们的生活带来了更多便利。例如，我们只需说出一个唤醒命令，如“嘿，Siri”，就能轻松激活设备，安排下午稍晚些时候的重要会议。要构建能够识别这类命令的系统，需要建立关键词识别（KWS）模型，以便在连续语音中检测预定义词语。在实际运行中，系统会先将原始音频转换为时频特征，再将其输入声学神经网络，预测能够使错误率最低的关键词类别。

近年来，开发高精度关键词识别网络取得了多项成功 [1, 2, 3, 4]。这些网络架构包括采用 Transformer 模块 [5,6], 、卷积模块 [7, 8], 、自回归层 [9]以及由这些结构组成的混合结构 [9, 10]，以构建深层的基于神经网络的声学模型。这些基于神经网络的模型大多面临标注训练数据不足的挑战。对于预定义关键词，学习良好的音频表示通常需要数千个训练样本，以避免过拟合 [11], ，深层网络尤其如此。然而，随着对个性化智能设备的需求不断增长，定制化关键词检出系统需要能够利用有限的用户样本快速适应。

为克服这一挑战，研究人员探索了许多数据增强技术，例如混合微小噪声失真、时间平移、时间拉伸和 SpecAugment[12, 13], ，以提升深度神经网络在低资源条件下的泛化能力。这些数据增强为数据实例注入少量变化，防止模型记住数据集，从而降低过拟合倾向。然而，语音预处理中可用的扰动类型有限，这限制了数据增强的多样性。

在本文中，我们尝试提升 KWS 模型在低资源条件下的鲁棒性。具体而言，我们旨在训练一个深度网络，使其在极小的训练集上（即 2.5 分钟、5 分钟、10 分钟）取得更好的性能。为此，我们提出了一种面向低资源关键词语音识别的对比语音混合（Cos-Mix）学习算法，该算法受到输入混合算法 [14]的启发。Mixup 算法是一种正则化技术，它使用两个线性插值样本及其对应的软标签来拟合网络。尽管已有多项研究证明这一基础方法有效，但 [15, 16] 表明，混合样本有时会产生空间上模糊且不自然的实例，进而使模型困惑，尤其是在模型试图定位并不存在的空间信息时。因此，我们引入了一种辅助对比损失，施加额外约束，以最大化原始单个样本与增强版本之间的相对相似度（我们将其称为正样本对）。借助对比损失，

本项工作获阿里巴巴集团通过阿里巴巴创新研究（AIR）计划和阿里巴巴-南洋理工大学新加坡联合研究院（JRI）资助，南洋理工大学，新加坡。

损失中，预混合样本表示为模型提供监督信号，使模型拉近正向混合样本之间相一致的属性。这种监督通过让模型感知混合语音片段来减少歧义。此外，它也带来了一定的增强效果，类似于 [17, 18], ，可生成信息丰富的语音编码。实验结果表明，在不同规模的训练数据上，CosMix 在各种关键词检出模型中的表现均优于原始 Mixup 增强方法。在参与比较的模型中，Keyword ConvMixer 表现最佳，在 5% 的训练数据（每个关键词 2.5 分钟）上达到了 90% 的准确率。总体结果表明，尽管所提出的 CosMix 学习算法设计简单且几乎不增加计算开销，仍然十分有效。

2. 相关工作

人们已在不同领域进行了多种尝试，以改进原始 Mixup 算法。在图像分类中，Manifold Mixup [19]利用语义插值作为额外的训练信号，从而在多个表示层级上获得更加平滑的决策边界。CutMix [15]用从另一张训练图像中采样的图像块替换图像中的某个区域，以保持训练图像的自然性，并增强模型抵御输入损坏的鲁棒性以及分布外检测性能。在语音信号处理领域，MixSpeech [20] 与 Mixup 有相似之处，它混合两段输入语音信号序列，并结合针对文本输出的两个损失函数，从而促进混合算法在自动语音识别等序列任务中的应用。此外，L-mix [21] 利用实例混合（i-mix）正则化来训练自监督说话人嵌入系统，以提升训练稳定性和说话人验证性能。相比之下，我们的工作聚焦于另一种变体，并已证明该变体对于低资源关键词检出是有效的。

3. 方法

3.1. Mixup 数据增强

原始 Mixup 数据增强采用邻域风险最小化原则，促使分类器在训练样本之间呈现线性行为。这一特性能够减少在未见过的实例上进行推理时出现的不期望变化。在数据预处理期间，我们从训练集中随机抽取两个音频实例，通过以下方式构造一个虚拟训练样本

$$\begin{aligned}\tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j\end{aligned} \tag{1}$$

其中 $i, j \in 1, ..N$ 表示训练集的索引， x 和 y 分别表示原始输入波形和独热标签编码。 $\lambda \sim$ 贝塔分布(α, α)，对于 $\alpha \in (0, \infty)$ 是 the

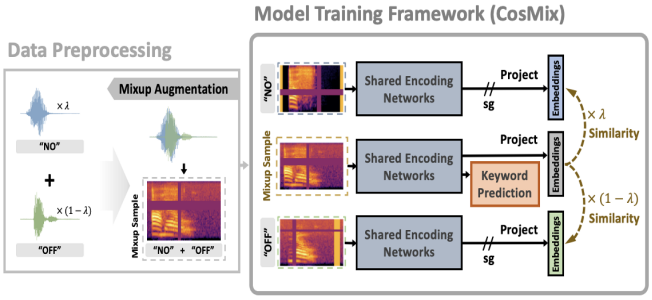


图 1。对比语音 Mixup（CosMix）模型架构示意图。所提出的方法包括音频 Mixup 数据增强，以及对 Mixup 实例进行对比学习，其中投影器 指将潜在向量嵌入映射到投影维度为 128 的投影器，而 sg表示停止梯度函数。

用于确定线性混合内容量的插值参数。给定虚拟输入-标签对$\tilde{}$ $\tilde{}(X, \tilde{y})$，其中 X 是$\tilde{x}$的 STFT 时频特征，我们按如下方式计算损失

$$\mathcal{L}_{\text{mix}} = \lambda \cdot \text{CE}(f(\tilde{X}), y_i) + (1-\lambda) \cdot \text{CE}(f(\tilde{X}), y_j) \tag{2}$$

其中 $f(.)$ 是声学编码模型，CE 指标准交叉熵损失。

3.2. CosMix 学习算法

尽管 Mixup 数据增强在广泛的语音应用中取得了成功，但它可能会将两段重叠语音混合，产生高度失真的信号，这些信号既不自然，也不利于任务。噪声输入可能造成混淆，导致模型误差出现异常尖峰；这些尖峰可能主导有效梯度，损害网络收敛[15, 16]。为解决这一问题，我们引入一个辅助组件，利用对比学习最大化原始混合样本与增强样本之间的相对相似度。我们的方法受到 [17] 的启发，该方法利用两个增强（正）视图之间一致的属性，来增强信息丰富的语音编码。在此，我们采用上述训练方法处理两个混合语句，从而促成简约（即复杂度较低）且充分（即保真度较高）的嵌入。这使模型能够生成更有效的表征，并在低资源条件下获得更强的泛化能力。

具体而言，训练过程使用两段语音片段的并行采样$\tilde{}$ X_i X_j 及其 Mixup X。我们对三个并行样本分别独立地应用随机增强。随后，将它们输入共享编码网络，在网络中，瓶颈表示将被投影，并用于与 Mixup 表示及各自未经混合的表示进行对比。与此同时，仅将 Mixup 表示输入关键词预测模块。训练框架示意图见

图 1。接着，为计算对比学习的损失，我们对所有投影嵌入进行 L2范数归一化，再利用均方误差衡量归一化投影之间的相似度。损失定义如下：

$$\mathcal{L}_{\cos}(\tilde{X}, X_r) = -\frac{\langle f_p(\tilde{X}), f_p(X_r)\rangle}{\|f_p(\tilde{X})\|_2 \|f_p(X_r)\|_2} \tag{3}$$

其中， $f_p(.)$是模型的投影器， $r \in \{i, j\}$。

此外，本研究采用了 50% 的 Mixup 比率，这意味着模型有一半的概率在不使用 Mixup 的情况下进行学习。尽管如此，我们还是将对比损失与其增强视图结合起来，以降低受扰动样本带来的变化，从而获得更简单的嵌入。因此，相对对比损失的加权方式如下

$$\Lambda_r = \begin{cases} \lambda, & \text{if } r = i \;\&\; r \neq j \\ 1-\lambda, & \text{if } r = j \;\&\; r \neq i \\ 1, & \text{if } r = i = j \end{cases} \tag{4}$$

完整的训练损失定义如下

$$\mathcal{L} = \mathcal{L}_{\text{mix}} + \beta \sum_{r \in \{i,j\}} \Big(\Lambda_r \cdot \mathcal{L}_{\cos}(\tilde{X}, X_r) \Big) \tag{5}$$

其中， β 是惩罚参数，用于衡量对比损失的贡献权重；本研究将其设为 0.5。

4. 实验

4.1. 数据集

在本研究中，我们使用 Google Speech Command V2 数据集 [22]。该数据集共包含 105,000 条语音片段，涵盖 35 个不同的单词。每个音频样本均以 WAV格式 文件存储，时长不超过一秒，采样率为 16 kHz。为确保代码的可复现性，我们采用官方提供的训练集、验证集和测试集划分，并选取其中包含 10 个关键词类别的子集，涵盖以下单词：“up”“down”“left”“right”“yes”“no”“on”“off”“go”和“stop”。为模拟低资源关键词语音识别场景，我们按说话人对每个单词对应的语音片段进行划分，并缩减相应训练集的规模。在本研究中，我们分别使用训练集的 5%、10%、20%、30% 和 50%。缩减后，每个单词对应的训练数据平均时长分别约为 2.5 分钟、5 分钟、10 分钟、15 分钟和 25 分钟。通过缩减说话人划分中的数据量，我们降低了训练数据的多样性，也增加了模型适应实际部署中更广泛人群的难度。这样的训练数据规模接近实际应用中训练个性化关键词语音识别模型所能获得的数据量。

4.2. 实验设置

输入特征 - 我们将所有 wav 文件转换为 64 维对数梅尔滤波器组（FBank），窗口大小为 25ms，

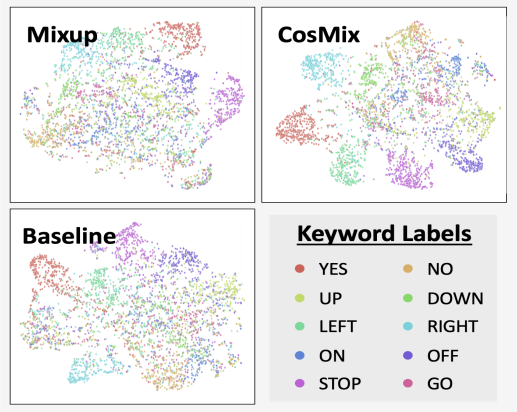


图 2。比较使用 KWT-3 的 20% 训练集时（训练 200 个周期后）不同技术所生成嵌入的 t-SNE 图。

移位量为 10ms。我们将 FBank 的分辨率固定为 98×64（即相当于 1s 的语音片段）。短于 1s 的指令将在右侧补零。训练期间，我们会对样本进行增强，包括在 -100 至 100ms 范围内随机时间移位，以及以 0.9 至 1.1 的比例随机时间拉伸[23]。此外，我们应用 SpecAugment，时间维度和频谱维度的掩蔽尺寸分别为 13 和 7。最后，表1中主实验的 Mixup 比率为 0.5，服从 贝塔分布(10,10) 分布。此设置适用于原始 mixup 和 CosMix。

训练细节- 在实验中，我们选取了业界常用的两类热门基于神经网络的模型，即基于 Transformer 的（KWT-1 和 KWT-3 [5]）以及基于卷积的（Keyword Con- vMixer [7]和 ResNet18 [24]）关键词识别网络。这些模型代表了适用于不同条件环境、具有不同模型大小和复杂度的近期先进水平。我们发现，最轻量的模型 ConvMixer [7],仅包含 0.1M 个参数。此外，我们为每个模型添加了一个投影器，将潜在向量嵌入映射到大小为 128 的投影维度。投影器由一个带 ReLU 激活的线性全连接块构成。所有模型的批量大小均设为 128。初始学习率为 5e-3，并从第 5 个周期到第 70 个周期，每隔四个周期以 0.85 的比例进行阶梯衰减。最后，我们使用 Adam 优化器和二元交叉熵损失进行优化。

5. 实验结果

我们使用 Google Speech Command 官方测试集（10 个类别）来评估模型性能。实验中的基线模型在不使用任何 Mixup 数据增强的情况下进行训练。尽管如此，我们也将 CosMix 与原始 Mixup 数据增强进行比较，以确定 CosMix 模型训练中的对比损失带来了多少性能提升。表1展示了我们的关键词识别模型在四种不同架构配置下的准确率。我们发现，所有模型在小型数据集上训练时，性能都会下降。

表1。Google Speech Command-V2（10类）测试数据集上的实验结果。我们在官方测试集上评估了不同训练数据规模（即 5%、10%、20%、30%、50%）下的关键词分类准确率。

模型	模型大小（M）	数据增强	训练数据规模（占训练集的百分比）				
			5%（2.5分钟）	10%（5分钟）	20%（10分钟）	30%（15分钟）	50%（25分钟）
ResNet18 [24]	11.9	基线	0.828	0.902	0.934	0.963	0.971
		原始 Mixup	0.822	0.909	0.947	0.968	0.978
		CosMix（我们的方法）	**0.841**	**0.919**	**0.955**	**0.970**	**0.981**
KWT-3 [5]	5.4	基线	0.465	0.613	0.635	0.677	0.732
		原始 Mixup	0.524	0.607	0.643	0.715	0.766
		CosMix（我们的方法）	**0.566**	**0.687**	**0.765**	**0.804**	**0.841**
KWT-1 [5]	0.6	基线	0.683	0.836	0.882	0.927	0.950
		原始 Mixup	0.625	0.853	0.892	0.933	0.951
		CosMix（我们的方法）	**0.701**	**0.863**	**0.911**	**0.935**	**0.960**
Keyword ConvMixer [7]	0.1	基线	0.887	0.926	0.954	0.965	0.971
		原始 Mixup	0.897	0.939	**0.963**	**0.973**	**0.978**
		CosMix（我们的方法）	**0.902**	**0.940**	0.962	**0.973**	0.976

在训练数据仅占 5% 时，KWT-3 的性能下降幅度最大，采用基线训练时准确率仅为 46.5%。尽管如此，无论使用哪种模型，CosMix 在不同训练集规模下总体上都取得了最佳性能。此外，训练数据越少（规模为 5%），性能提升越显著；KWT-3 的相对性能提升最高可达 21.7%。分析各个模型后，我们注意到，在低资源场景下，基于 Transformer 的模型更容易出现性能下降。这可能是因为训练样本不足以学习高度复杂的注意力机制。即便如此，CosMix 通过更好的正则化帮助缓解了这一问题。最后，基于卷积的模型表现最佳，其中 Keyword ConvMixer 在 5% 的训练集上使用 CosMix 时，准确率最高，达到 90%。

为考察不同技术所产生的声学表征的质量，我们在图 2 中使用 t-SNE 图可视化嵌入，实验采用 KWT-3，并使用 20% 的训练数据。基线设置能够区分“right”和“yes”与其他命令，但模型无法区分其余命令。随着 Mixup 数据增强改善嵌入，各簇之间的间隔略有增大。CosMix 中的簇最易分离，只有一些较短的辅音词仍然聚在一起，这表明我们的方法能够有效学习精确且内容丰富的表征。

5.1. 消融研究

在本节中，我们考察使用不同参数对两种 Mixup 算法性能的影响。具体而言，我们研究了 Mixup 比率以及从 Beta分布中得出的插值权重的变化。当α Beta分布中的值小于 1 时，贝塔分布(α, α) 会呈现凸形曲线，音频混合量往往集中在一侧。当 α 该值大于 1 时，曲线则变得更凹，两个音频样本更有可能按比例混合。表 2 展示了我们的研究结果。首先，贝塔分布(10,10) 通常优于 贝塔分布(0.5,0.5)，这表明对于关键词识别任务，模型使用等比例混合的音频样本能够更有效地学习。其次，Mixup 和 CosMix 的最佳混合比例有所不同。Mixup 在混合比例为 30% 时性能最佳，而 CosMix 在 50% 时效果最好。最后，用户在调整超参数时应谨慎，因为我们观察到模型性能出现了两个峰值。这似乎在很大程度上受到 Beta 的双峰分布影响。

表 2。基于 ResNet18 在不同 Mixup 混合比例下，使用 20% 训练集在官方测试集上的准确率。

混合比例（%）	贝塔分布(0.5, 0.5)		贝塔分布(10, 10)	
	Mixup	CosMix	Mixup	CosMix
10	**0.940**	0.932	0.936	**0.946**
30	**0.949**	0.940	**0.953**	**0.953**
50	0.941	**0.958**	0.947	**0.955**
70	**0.946**	0.945	0.945	**0.952**
100	0.938	**0.949**	**0.929**	0.924

6. 结论

本文提出了 CosMix，这是一种面向低资源关键词语音识别的新型数据增强策略。CosMix 的模型训练方法利用对比损失，减轻传统 Mixup 训练所产生的噪声训练信号带来的不良副作用。在多种模型大小下，CosMix 都能够提升低资源条件下的模型性能。因此，对于智能设备上的个性化关键词识别系统等应用，CosMix 作为一种提升低资源条件下性能的通用方法是有效的。

7. 参考文献

[1] Tara N Sainath 和 Carolina Parada，《用于小型关键词检出的卷积神经网络》，*Proc. Inter- speech 2015*，2015。[2] Yuan Gong、Yu-An Chung 和 James Glass，《Ast：音频频谱图 Transformer》，*arXivpreprint arXiv:210401778*，2021。。[3] Dianwen Ng、Jin Hui Pang、Yang Xiao、Biao Tian、Qiang Fu 和 Eng Siong Chng，《用于基于质心感知的关键词检出的轻量级多通道 ConvMixer》，*arXiv preprintarXiv:2204.05445*，2022。[4] David Peter、Wolfgang Roth 和 Franz Pernkopf，《使用神经架构搜索和量化的端到端关键词检出》，载于*ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*。IEEE，2022，第 3423–3427 页。[5] Axel Berg、Mark O'Connor 和 Miguel Tairum Cruz，《关键词 Transformer：一种用于关键词检出的自注意力模型》，载于 *Interspeech 2021*。ISCA，2021，第 4249–4253 页。[6] Kevin Ding、Martin Zong、Jiakui Li 和 Baoxiang Li，《LETR：一种轻量高效的关键词检出 Transformer》，载于*ICASSP 2022-2022 IEEE International Conference on Acous- tics, Speech and Signal Processing(ICASSP)*。IEEE，2022，第 7987–7991 页。[7] Dianwen Ng、Yunqi Chen、Biao Tian、Qiang Fu 和 Eng Siong Chng，《ConvMixer：结合课程学习的特征交互卷积，用于小型化及噪声远场关键词检出》，载于 *2022 IEEE International Conference on Acoustics, Speech andSignal Processing (ICASSP)*，2022，第 3603–3607 页。[8] Seungwoo Choi、Seokjun Seo、Beomjun Shin、Hyeongmin Byun、Martin Kersner、Beomsu Kim、Dongyoung Kim 和 Sungjoo Ha，《移动设备上的实时关键词检出时序卷积》，*arXiv preprint arXiv:190403814*，2019。。[9] Oleg Rybakov、Natasha Kononenko、Niranjan Subrahmanya、Mirk´o Visontai 和 Stella Laurenzo，《移动设备上的流式关键词检出》，*Proc. Interspeech 2020*，第 2277–2281 页，2020。[10] Mengjun Zeng 和 Nanfeng Xiao，《DenseNet 与 BiLSTM 在关键词检出中的有效结合》，*IEEE Access*，第 7 卷，第 10767–10775 页，2019。[11] Chen Chen、Nana Hou、Yuchen Hu、Shashank Shirol 和 Eng Siong Chng，《使用 10 分钟无与伦比的域内数据实现抗噪语音识别》，载于 *ICASSP 2022-2022IEEE International Conference on Acoustics, Speech and Sig- nal Processing (ICASSP)*。IEEE，2022，第 4298–4302 页。[12] Shengyun Wei、Shun Zou、Feifan Liao 等，《基于深度学习的音频分类数据增强方法比较》，载于 *Journal of Physics: Conference Series*。IOP Publishing，2020，第 1453 卷，第 012085 页。[13] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le，《SpecAugment：一种用于自动语音识别的简单数据增强方法》，*arXiv preprintarXiv:1904.08779*，2019。[14] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz，《Mixup：超越经验风险最小化》，载于 国际学习表征会议，2018。

[15] Sangdoo Yun、Dongyoon Han、Seong Joon Oh、Sanghyuk Chun、Junsuk Choe 和 Youngjoon Yoo，《Cutmix：利用可定位特征训练强大分类器的正则化策略》，载于*IEEE/CVF* 国际计算机视觉会议论文集，2019 年，第 6023–6032 页。[16] JangHyun Kim、Wonho Choo、Hosan Jeong 和 Hyun Oh Song，《Co-mixup：由显著性引导、具有超模多样性的联合 Mixup》，载于国际学习表征会议，2020 年。[17] Dianwen Ng、Jia Qi Yip、Tanmay Surana、Zhao Yang、Chong Zhang、Yukun Ma、Chongjia Ni、Eng Siong Chng 和 Bin Ma，《I2cr：使用内外对比正则化提升关键词检出的抗噪能力》，*arXiv* 预印本 *arXiv:2209.06360*，2022 年。[18] Yuchen Hu、Nana Hou、Chen Chen 和 Eng Siong Chng，《面向端到端抗噪语音识别的交互式特征融合》，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022 年，第 6292–6296 页。[19] Vikas Verma、Alex Lamb、Christopher Beckham、A mir Najafi、Ioannis Mitliagkas、David Lopez-Paz 和 Yoshua Bengio，《Manifold Mixup：通过插值隐藏状态获得更好的表征》，载于国际机器学习会议。PMLR，2019 年，第 6438–6447 页。[20] Linghui Meng、Jin Xu、Xu Tan、Jindong Wang、Tao Qin 和 Bo Xu，《MixSpeech：面向低资源自动语音识别的数据增强》，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021 年，第 7008–7012 页。[21] Woo Hyun Kang、Jahangir Alam 和 Abderrahim Fathan，《L-mix：用于鲁棒自监督说话人表征学习的潜在层级实例 Mixup 正则化》，*IEEE* 信号处理专题期刊，第 16 卷，第 6 期，第 1263–1272 页，2022 年。[22]Pete Warden，《Speech commands：面向有限词汇语音识别的数据集》，*arXiv* 预印本 *ar Xiv:1804.03209*，2018 年。[23] Tom Ko、Vijayaditya Peddinti、Daniel Povey 和 Sanjeev Khudanpur，《面向语音识别的音频数据增强》，载于国际语音通信协会第十六届年会，2015 年。[24] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，《用于图像识别的深度残差学习》，载于*IEEE* 计算机视觉与模式识别会议论文集，2016 年，第 770–778 页。