



# 【有Github】 Pre-training on high-resource speech recognition improves low-resource speech-to-text translation

提出了一种在源语言资源不足时改进直接语音到文本翻译 (ST) 的简单方法：我们在高资源自动语音识别 (ASR) 任务上对模型进行预训练，然后进行微调 ST 的参数。

1. 通过对 300 小时的英语 ASR 数据进行预训练，当只有 20 小时的西班牙语–英语 ST 训练数据可用时，将 Spanish–English ST 从 10.8 提高到 20.2 BLEU。预训练的编码器（声学模型）改进明显，尽管这些任务中的共享语言是目标语言文本，而不是源语言音频。
2. 即使 ASR 语言与源和目标 ST 语言都不同，对 ASR 的预训练也有助于 ST：对法语 ASR 的预训练也可以提高西班牙语–英语 ST。
3. 对英语 ASR 和法语 ASR 组合的预训练提高了 Mboshi–French ST，其中只有 4 小时的数据可用，从 3.5 到 7.1 BLEU。

(题图)

标题：预训练高资源的语音识别进行改善低资源的语音翻译

Github 地址：

github.com

<https://github.com/0xSameer/ast>

## 为什么要读？

- 解决具体问题

## 作者写这个的目的是什么？

- 报告新发现
- 表达特别的观点

## 和我的研究相关之处

对于低资源的 ST 训练语料，即使 ASR 语言与源和目标 ST 语言都不同，对 ASR 的预训练也有助于 ST

## 研究有多可信

### 1. 文章潜在假设是什么？

低资源 ST 可以利用高资源目标语言的转录音频，甚至完全不同的语言，只需为高资源 ASR 任务预训练模型，然后转移和为低资源 ST 微调部分或全部模型参数。

### 2. 研究方法是什么？

这个方法的每一步操作怎么做，每一步什么目的，每一步为了解决什么样的小问题，有什么难点，作者是如何解决的？

#### 2.1 关于语音和文本的语料处理：

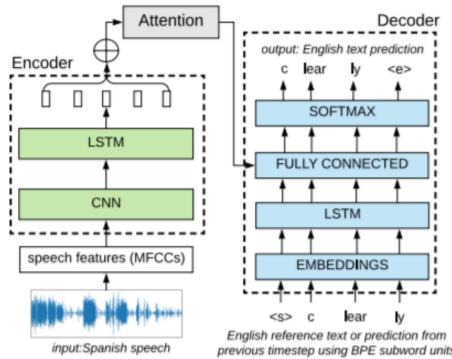
- 语音处理 Speech:

我们使用 Kaldi (Povey et al., 2011) 将原始语音输入转换为 13 维 MFCC。我们还执行说话人级别的均值和方差归一化。

- 文本处理 Text:

使用字节对编码 (BPE; Sennrich et al., 2016) 将每个词分割成子词，每个子词是一个字符或字符的高频序列——我们使用了 1000 个这些高频序列。由于子词集包含了全集字符，模型仍然是开放词汇；但它产生的文本只有 190 万个标记和超过 1K 种类型，训练速度几乎与词级模型一样快。BPE 的词汇取决于字符序列的频率，因此必须针对特定的语料库进行计算。

## 2.2 模型结构：



- **Speech Encoder:**

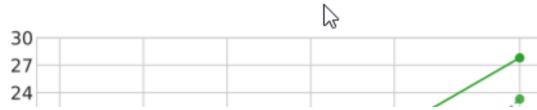
1. 使用 25 ms 的窗口大小和 10 ms 的步长提取的 MFCC 特征向量被馈送到两个 CNN 层的堆栈中，具有 128 和 512 个滤波器，每个滤波器宽度为 9 帧。
2. 在每个 CNN 层中，我们沿时间跨度为 2，应用 ReLU 激活 (Nair 和 Hinton, 2010)，并应用批量归一化 (Ioffe 和 Szegedy, 2015)。
3. CNN 层的输出被馈送到三层双向长期短期记忆网络 (LSTM; Hochreiter 和 Schmidhuber, 1997)；每个隐藏层有 512 个维度。

- **Text decoder:**

1. 在每个时间步，解码器从全连接层产生的 softmax 层的输出中选择最可能的标记，然后接收从先前时间步计算的循环层的当前状态和输入。
2. 使用具有一般评分函数和输入馈送的全局 attention 模型计算注意力。
3. 将预测的标记输入 128 维嵌入层 (EMBEDDINGS)
4. 然后是三层 LSTM 以更新循环状态；每个隐藏状态有 256 个维度。在训练时，我们有 20% 的时间使用预测标记作为下一个解码器步骤的输入，其余 80% 的时间使用训练标记。
5. 在测试时，我们使用束大小为 5 的束解码和权重为 0.6 的长度归一化。

## 3.研究结论如何？

- **Spanish–English ST**



对于 Low-resource (20–50H) / very low-resource (<10H), 迁移学习预训练的作用对于 BLEU/精度 prec/召回度 recall, 都有显著提升. 这主要是由于子词而不是词的更好的正则化和建模

<i>Spanish</i>	super caliente pero muy bonito
<i>English</i>	super hot but very nice
20h	you support it <u>but</u> it was <u>very</u> nice
20h+asr	you can get alright <u>but</u> it's <u>very</u> nice
50h	<u>super</u> expensive but <u>very</u> nice
50h+asr	super hot but it's <u>very</u> nice
<i>Spanish</i>	sí y usted hace mucho tiempo que que vive aquí
<i>English</i>	yes and have you been living here a long time
20h	yes i've <u>been</u> a long time what did <u>you</u> come <u>here</u>
20h+asr	yes and <u>you</u> have a long time that <u>you</u> live <u>here</u>
50h	yes <u>you</u> are a long time that <u>you</u> live <u>here</u>
50h+asr	yes and have <u>you</u> been <u>here</u> long

表3示例：对于基线 sp-en-50h 模型，在翻译句子的时候可以正确翻译单词，但不能正确理解英语单词顺序。通过添加 ASR 的预训练的迁移学习，该模型以正确的词序生成更短但仍然准确的翻译。结论：预训练对于翻译语句中词序质量有提升

- Using French ASR to improve Spanish–English ST

假设：

研究发现来自英语 ASR 模型的迁移学习相比其他语言对于模型更有正面影响,为了将其与语言无关语音特征的迁移学习隔离开来,进行了进一步的实验

方法：

使用法语 ASR 数据对西班牙语–英语 ST 任务进行预训练。因为法语数据集 (20 小时) 远小于我们的英语数据集 (300 小时) , 为了公平比较, 我们在本实验中使用了 20 小时的英语数据子集进行预训练对比实验。对于英语和法语模型, 只传输编码器参数。

结论：

	baseline	+fr-20h	+en-20h
sp-en-20h	10.8	12.5	13.2

Table 4: Fisher dev set BLEU scores for *sp-en-20h*.  
 baseline: model without transfer learning. Last two columns: Using encoder parameters from French ASR (+fr-20h), and English ASR (+en-20h).

model	pretrain	BLEU	Pr.	Rec.
fr-top-8w	–	0	23.5	22.2
fr-top-10w	–	0	20.6	24.5
en-300h	–	0	0.2	5.7
fr-20h	–	0	4.1	3.2
mb-fr-4h	–	3.5	18.6	19.4
	fr-20h	5.9	23.6	20.9
	en-300h	5.3	23.5	22.6
	<b>en + fr</b>	<b>7.1</b>	<b>26.7</b>	<b>23.1</b>

英语和法语的 20 小时预训练模型都提高了西班牙语–英语 ST 的性能。英文模型的效果稍好一些, 但法语模型也很有用, 将 BLEU 从 10.8 提高到 12.5。

在完全不同的第三语言上进行 ASR 预训练可以帮助资源不足的 ST。若使用更大的 ASR 数据集, 收益会更大,

- Mboshi–French ST

方法：

将只有 4 小时的 ST 训练数据：Mboshi 语音输入与法语文本输出配对用于训练。

结论：

上图中的下表，通过对比实验，尝试 en-300h 模型的编码参数和 fr-20h 模型的解码参数。最终的 (en+fr) 高资源 ASR 预训练结合 提供了所有指标的最佳评估分数，但是结果还是不优秀。

**选定不同参数对于实验效果的印象：**

假设

ASR 模型的所有参数，包括语音编码器 CNN 和 LSTM；文本解码器嵌入、LSTM 和输出层参数；和 attention 参数。查看哪组参数的影响最大

方法

通过仅传输来自 en-300h 的选定参数并随机初始化其余参数来训练 sp-en20h 模型。

结果

1. 传输所有参数是最有效的，并且语音编码器参数占了大部分增益。
2. 仅传输解码器参数并不能提高准确性。由于解码器参数在与编码器参数一起使用时会有所帮助，因此传输的解码器参数已经被训练为期望来自编码器的特定输入表示，仅传输解码器参数而不传输编码器可能没用
3. 当只有几十小时的训练数据可用于 ST 时，迁移学习最有用。随着 ST 训练数据量的增加，迁移学习的好处逐渐减少。

## 对我有什么用

- 回答了我为什么要读的问题了吗？

实现了在极低平行语料库资源的情况下，ST 模型的训练。其模型以及数据处理方式具有参考意义

- 准备正面/负面的引用吗？

参考代码的模型结构进行构建。