ST数据增强方法：通过转换现有数据来创建合成数据

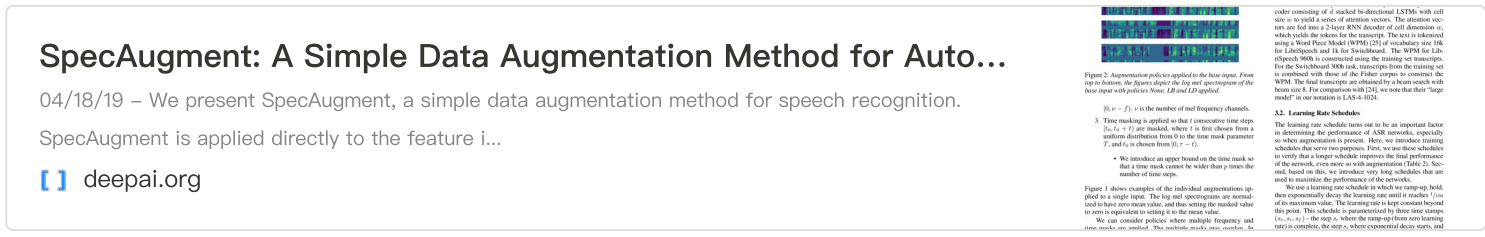
模型可以通过增加数据大大改善其质量。经典的方法是改变输入样本，同时保持其类别标签不变。

1 音频扰动

此方法类似于处理图像的方法,改变输入但保留标签

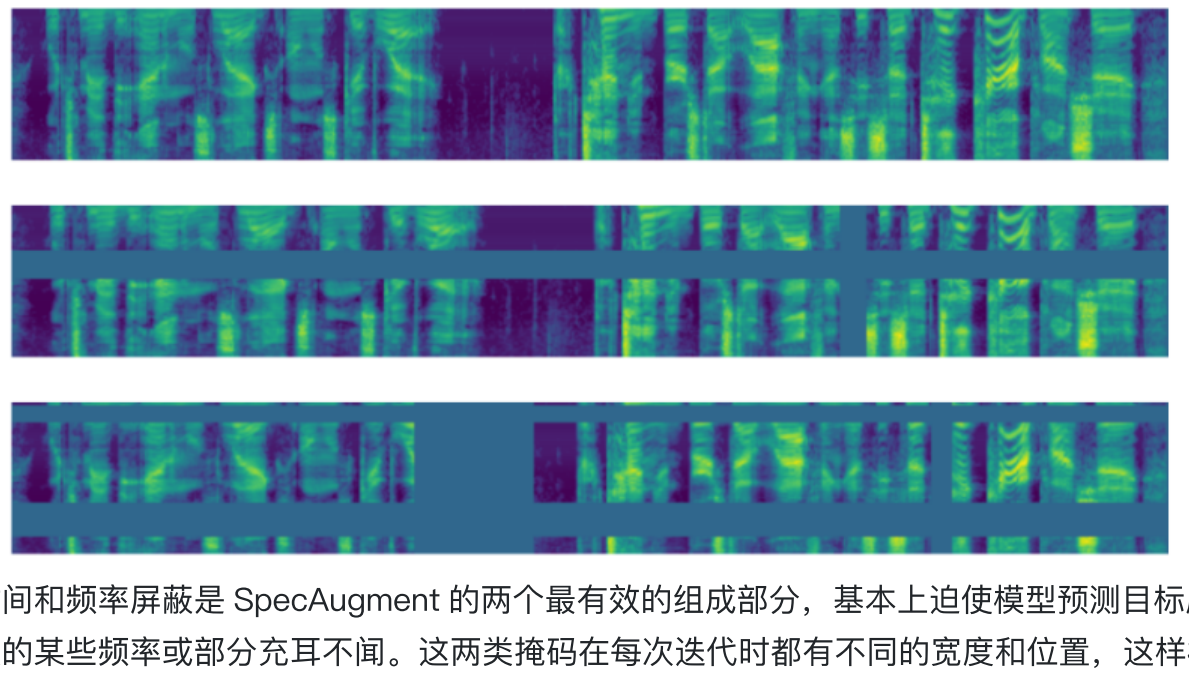
1.1 ASR 界广泛使用的一种方法是**速度扰动**(speed perturbation)

> 它包括使用像 SoX 这样的工具来扰动音频速度，同时保持翻译的固定。例如，在 IWSLT 2019[1]的获奖作品中就使用了这种方法，一种常见的做法是将速度（或时间长度，对我们的目的来说是等效的）乘以一个范围为[0.9–1.1]的随机因素，这通常会产生一个仍然像人一样的音频，但听起来与原声不同。这种转换通常是在建立数据集的时候离线应用。



SpecAugment 是一种流行的频谱图增强技术，去年被提出并获得了巨大的成功。它由三个步骤组成:

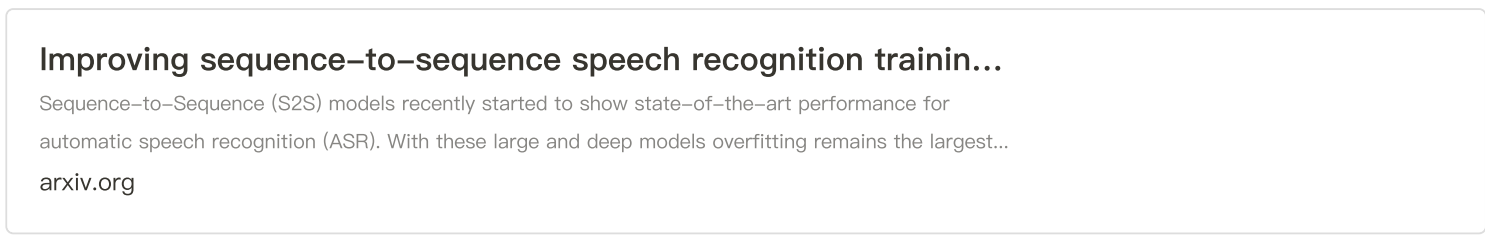
1. 时间扭曲是一种复杂的、计算成本很高的算法，它将频谱图的一部分沿时间轴移动；
2. 频率屏蔽应用一个水平屏蔽，覆盖整个时间维度的一些频率；
3. 时间屏蔽应用一个垂直屏蔽，覆盖一些相邻时间步骤的所有频率。



> 时间和频率屏蔽是 SpecAugment 的两个最有效的组成部分，基本上迫使模型预测目标序列，同时对音频的某些频率或部分充耳不闻。这两类掩码在每次迭代时都有不同的宽度和位置，这样模型就能学会使用它们。

1.2 **在线扰动**

> 只有某些类型的转换可以在线应用，因为目前的语音翻译系统接收的输入是频谱图，而不是波形图。为了在训练过程中占用更少的磁盘空间并节省计算时间，数据集通常以频谱图的形式存储，这需要计算的功率，而且比波形图更紧凑。



> 他们的方法为时间拉伸，将输入的频谱划分为固定宽度的窗口，并以不同的随机因素（0.8–1.25）收缩或拉伸每个窗口。有了这种技术，每个频谱图都会有不同的时间扰动窗口，而且是对比性的，根据作者的说法，它可以取代离线速度扰动。

2 Transfer Learning——迁移学习

首先，在资源大的任务上训练一个模型，然后对第二个任务**使用相同的深度学习架构**，并用从第一个任务中学到的权重进行初始化。

1. 训练一个 ASR 系统，使用将用于直接 ST 的相同架构（至少在编码器方面）。
2. 训练一个 MT 系统，其解码器结构与直接 ST 所用的相同。
3. 最后，用在 ASR 中学习的编码器权重和在 MT 中学习的解码器权重初始化直接 ST 系统。

这种方法会导致更快的收敛和更好的翻译质量，然而，解码器的预训练似乎不如编码器的预训练有效。

> **A Comparative Study on End-to-end Speech to Text Translation**
> Recent advances in deep learning show that end-to-end speech to text translation model is a promising approach to direct the speech translation field. In this work, we provide an overview of...
> arxiv.org

为了克服这个问题，如果在预训练的编码器上再加一个 "适配器 "编码器层，那么解码器预训练会更有效。

3 Two-Stage Decoding——两级解码

> **Structured-based Curriculum Learning for End-to-end...**
> Sequence-to-sequence attentional-based neural network architectures have been shown to provide a powerful model for machine translation and speech recognition. Recently, several works have...
> arxiv.org

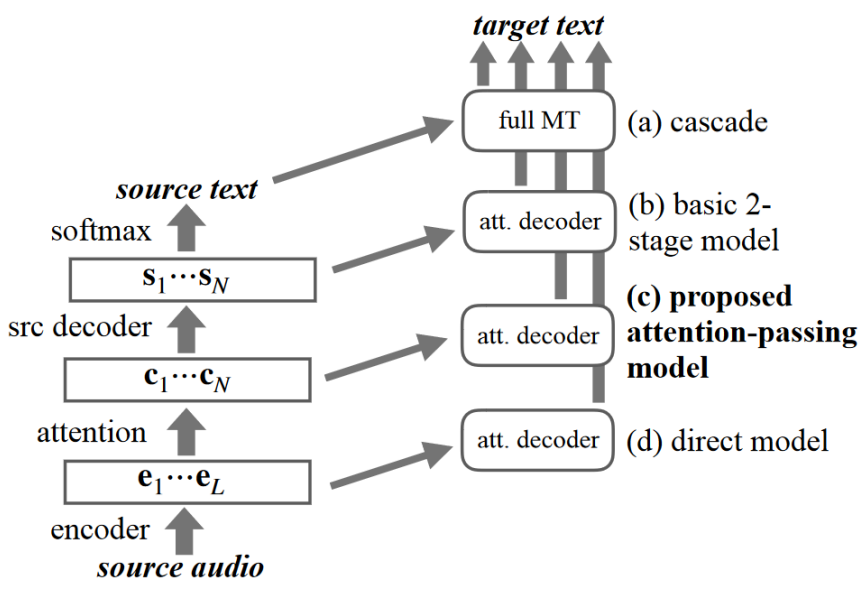
两级解码，作为一种更好地使用预训练组件的方法。它是一个由三个组件组成的网络，一个编码器和两个级联解码器。

1. 第一个解码器生成源语言的句子，类似于 ASR 系统，
2. 第二个解码器生成目标语言的文本。用第一个解码器的状态计算其注意力，就在它们被用来选择输出符号之前。

在这个意义上，两级解码类似于级联系统，但它没有将源序列离散化，以防止错误。两阶段解码模型是由预先训练好的 ASR 和 MT 模型初始化的，以提高有效性。

> **Attention-Passing Models for Robust and Data-Efficient End-...**
> Speech translation has traditionally been approached through cascaded models consisting of a speech recognizer trained on a corpus of transcribed speech, and a machine translation system trained...
> arxiv.org

Sperber 提出了一个 "注意力传递 "模型，该模型改进了两阶段解码模型，以实现更有效的预训练。解码器的状态已经包含了选择生成下一个词的信息，那么模型中又出现了一种错误传播的形式。为了克服这个问题，他们建议使用**由第一个解码器和编码器之间的注意力产生的上下文向量作为第二个解码器注意力的输入**。这样一来，第二个解码器直接与相关的音频部分相连。



4 弱监督

利用现有的系统为其他任务创建合成的平行音译数据:**当只有音频-转录数据时，使用 MT 系统将转录本翻译成目标语言**。通过这种方法，直接 ST 的训练数据可以增加几个数量级，而且这些任务的系统质量可以保证数据的高质量。他们还表明，使用 MT 比 TTS 更有效，也许是因为合成的数据与真实条件下的人声不是很相似。这种数据增强的效果优于所有其他提出的方法，在某些情况下还使预训练变得无用。

目前总结最有效的方法(三合一):

- 经过强大 ASR 模型预训练的编码器(encoder)
- 通过使用强大的 MT 系统生成合成翻译来增强训练集
- 在训练期间使用 SpecAugment