论文标题

Permutation invariant training of deep models for speaker-independent multi-talker speech separation

要点

本文提出了一种 speaker independent multitalker speech separation 的训练方法，称为 permutation invariant training (PIT)，PIT 首先决定最好的输出目标（多个音频源）分配任务，然后在这一任务下最小化误差。这一策略直接在网络结构中实现，高雅地解决了长且持续的标签序列问题（阻碍深度学习在语音分离上的进度），我们在 WJS0 和 Danish mixed-speech separation tasks 上评估了 PIT。实验结果显示 PIT 比 NMF，CASA 和 DPCL 上对没有见过的说话人和语言具有更好的泛化效果。

如果在时域上有 S 段源信号序列，$\mathbf{x}_s(t), s = 1, \cdots, S$，那么混合信号序列为：

$$\mathbf{y}(t) = \sum_{s=1}^{S} \mathbf{x}_s(t)$$

相应的，用 STFT 转换到时频域上则有：

$$\mathbf{Y}(t,f) = \sum_{s=1}^{S} \mathbf{X}_s(t,f)$$

单声道语音分离的目标是从 $\mathbf{Y}(t,f)$ 中复原出 $\mathbf{X}_s(t,f)$。

在典型的做法中，一般会假设只有 STFT 幅度谱信息是有用的，而相位信息会在分离的过程中被忽略，仅在复原出信号源在时域上的波形图时会用到相位信息（即使用 ISTFT 的时候）。

很明显，仅仅给出 mixed spectrum 的幅度谱 $|\mathbf{Y}(t,f)|$，复原出 $|\mathbf{X}_s(t,f)|$ 是一个“病态”问题，因为 $|\mathbf{X}_s(t,f)|$ 的组合有无穷多种可能。为了解决这个问题，需要从一个包含了 pairs（$|\mathbf{Y}(t,f)|$，$|\mathbf{X}_s(t,f)|$）的训练集 $\mathbb{S}$ 中寻找规律，更确切地说，可以通过训练一个深度学习模型 $g(\cdot)$ 例如：

$$g\left(f(|\mathbf{Y}|);\theta\right) = |\mathbf{X}_s|, s = 1, \cdots, S$$

θ 是模型的参数，$f(|\mathbf{Y}|)$ 是 $|\mathbf{Y}|$ 的一些特征表示。在语音分离问题中，比较常见的作法并不是直接估计 $|\mathbf{X}_s|$，而是首先使用模型 $h(f(|\mathbf{Y}|);\theta) = \tilde{\mathbf{M}}_s(t,f)$ 估计一系列的掩码 $\mathbf{M}_s(t,f)$，注意到约束 $\tilde{\mathbf{M}}_s(t,f) \geq 0$ 和 $\sum_{s=1}^{S} \tilde{\mathbf{M}}_s(t,f) = 1$ 在所有的 T-F 单元上都是成立的，这一约束可以使用 softmax 轻易得到满足。如果我们要估计 $|\mathbf{X}_s|$，直接进行如下计算就可以了：

$$|\tilde{\mathbf{X}}_s| = \tilde{\mathbf{M}}_s \circ |\mathbf{Y}|,$$

估计掩码时，模型参数可以通过优化估计的掩码 $\tilde{\mathbf{M}}_s$ 和真实的掩码（IRM）$\mathbf{M}_s = \frac{|\mathbf{X}_s|}{|\mathbf{Y}|}$ 之间的 MSE 来获得：

$$J_m = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \|\tilde{\mathbf{M}}_s - \mathbf{M}_s\|^2$$

该模型存在两个主要问题，其一，在安静的片段上 $|\mathbf{X}_s| = 0$ 和 $|\mathbf{Y}| = 0$，此时 $\mathbf{M}_s$ 由于发生除 0 操作而无法明确定义。我们真正关心的是每个源估计的幅度谱和真实的幅度谱之间的误差，掩码上误差较小不代表幅度谱上的误差较小。为了克服这一缺陷，近期的研究提出优化 MSE 直接在幅度谱上进行：

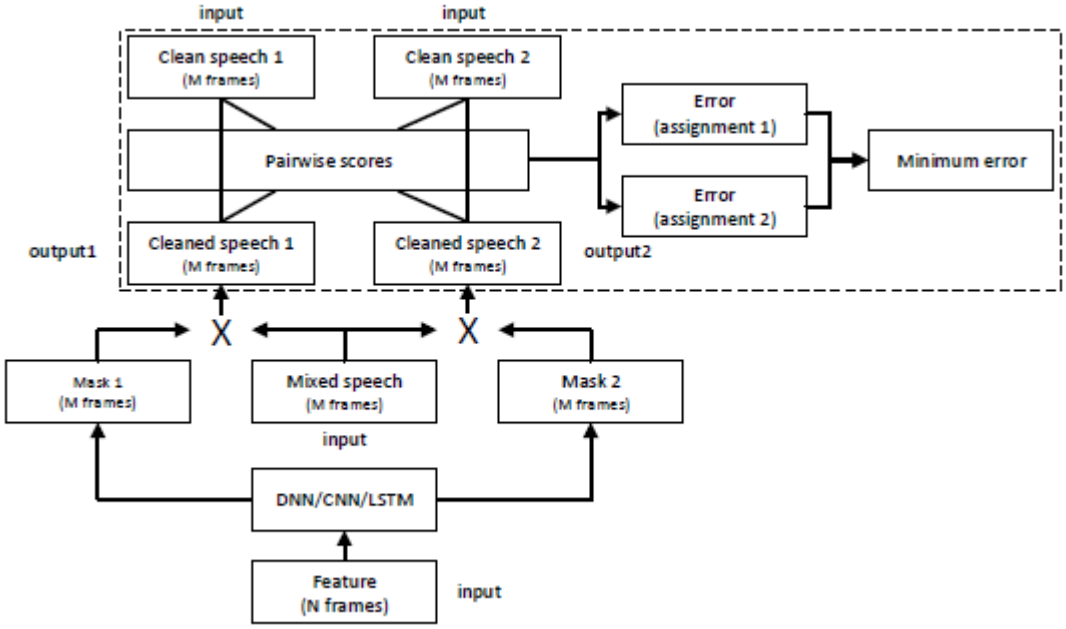
$$J_x = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \||\tilde{\mathbf{X}}_s| - |\mathbf{X}_s|\|^2$$

由于安静的片段上 $|\mathbf{X}_s| = 0$ 和 $|\mathbf{Y}| = 0$，因此掩码估计的准确率不会影响这些片段上的训练准则。不过这篇文章依然采用前一种 MSE 进行优化。

多数时候，语音的分离问题被视为一个多分类的回归问题，混合音频信号 $|\mathbf{Y}|$ 中的 N 帧（不是全部帧）特征向量被用作深度学习模型（DNN，CNN，RNN，LSTM）的输入来为每一位说话者生成一帧掩码（通常位于中间，大小为 N 的窗的中间），这些掩码分别被用于复原出单源语音 $|\tilde{\mathbf{X}}_1|$ 和 $|\tilde{\mathbf{X}}_2|$，它们分别代表信号源 1 和 2。在训练的过程中，我们需要提供正确的参考（目标）频谱 $|\mathbf{X}_1|$ 和 $|\mathbf{X}_2|$ 给相关的输出层作为监督，而且它们都**取决于相同的输入---同一段混合音频**。参考目标的分配是很复杂的特别是当训练集中包含了很多说话人的发言时，这一问题被称为标

签含糊（或序列）问题。由于这个问题，许多现有的模型在 speaker-independent 的 multitalker speech separation 问题上表现较差。解决这一问题的方案如图所示，有两个关键的概念：序列不变训练(PIT)和基于片段(segment)的决策。

在我们的模型中，混合音频中用于参考的源音频流是一个集合而不是一个有序列表。换句话说，无论这些源的顺序是怎么样的，我们总会得到相同的训练结果。这一行为如下图中虚线矩形框所示。



为了将参考目标与输出层关联起来，我们首先在参考目标和估计的信号源之间决定出 S!中可能的任务。然后我们为每一个分配的任务计算总的 MSE，即组合（加法）起各个信号源的参考值$|\mathbf{X}_s|$与估计值$|\tilde{\mathbf{X}}_s|$的 MSE。接着从各大任务的 MSE 中挑选出最小的一个进行优化。换句话说，我们在进行标签分配任务的同时也在进行误差评估。同样，PIT 使用连续的 N 帧特征（元数据帧）作为输入以利用上下文的信息，与已经提出的模型不同，PIT 的输出也是一个窗大小的帧（N）。在 PIT 的帮助下，我们直接最小化元数据帧级别的分离误差。注意到尽管任务的数量是说话人的阶乘，但计算 MSE 的代价却是二次方级别，更重要的是 MSE 的计算在评估时几乎可以忽略不计。

在前向传播的时候，唯一可用的信息就是混合音频，语音的分离可以直接由每个输入元数据帧得到，从而为每个流输出具有 M 帧的元数据帧。输入的源数据帧在切换时可以一帧一帧移动也可以移动多帧。所有的输出源数据帧都要被考虑，

例如取平均。

实验数据集

我们在 PIT 上评估 WSJ0-2mix 和 Danish-2mix 数据集，WSJ0-2mix 是由 WSJ0 corpus 派生而来的，训练集一共有 30h，验证集为 10h，它们包含了来自 WSJ0 的训练集（si_tr_s）中随机选择的说话人和发言所生成的双说话人混合音频，然后将它们均匀混合到 0dB 到 5dB 各个级别的 SNR 下。5h 的测试集由 WSJ0 的验证集 si_dt_05 和评估集 si_et_05 的 16 位说话人中以类似的方式生成。

Danish-2mix 数据集是从 Danish corpus 中生成的，大约包含了 560 位说话人的的 312 段发言，平均发言停顿大约为 5 秒。数据集由 corpus 中随机选取的 45 位男性和 45 位女性组成，并从每位发言者中分配 232，40，40 段发言来生成混用于训练集，验证集和 CC（seen speaker）测试集。另外 45 位男性和 45 位女性被随机选取出来构建 OC（unseen speaker）测试集。混合音频的构建方式与 WSJ0-2mix 数据集类似，但它们全部混合着 0dB SNR 的噪音---最极端的情况下，因此，我们总共创建了 10k 混合音频的训练集和 1k 混合音频的验证集，此外，还有 1k 混合音频的 CC 测试集 1k 混合音频的 OC 测试集，Danish-3mix 也采用类似的方式创建。这篇文章主要关注 WSJ0-2mix 数据集，因为可以和已发表的模型结果比较。

实验结果分析

模型实现用 CNTK，DNN 共有三个隐藏层，每层 1024 个 ReLU 单元，在（inChannel, outChannel）-（strideW, strideH）格式，CNN 模型的结构如下：

卷积层：

1*（1，64）-（2，2）

4*（64，64）-（1，1）

1*（64，128）-（2，2）

2*（128，128）-（1，1）

1*（128，256）-（2，2）

2*（256，256）-（1，1）

卷积核 3*3,还有一个池化层和一个 1024 的 ReLU 层。模型的输入是混合音频的 257-dim 的 STFT 幅度谱的堆叠(多帧),是由 32ms 的帧长和 16ms 的帧移的 STFT 计算而来的。模型一共使用 S 个输出（假设一共是 S 个说话人的混合音频），每个输出流的维度为 257*M，M 代表输出的元数据帧中的帧的数目。研究中，验证集被用于控制学习率。

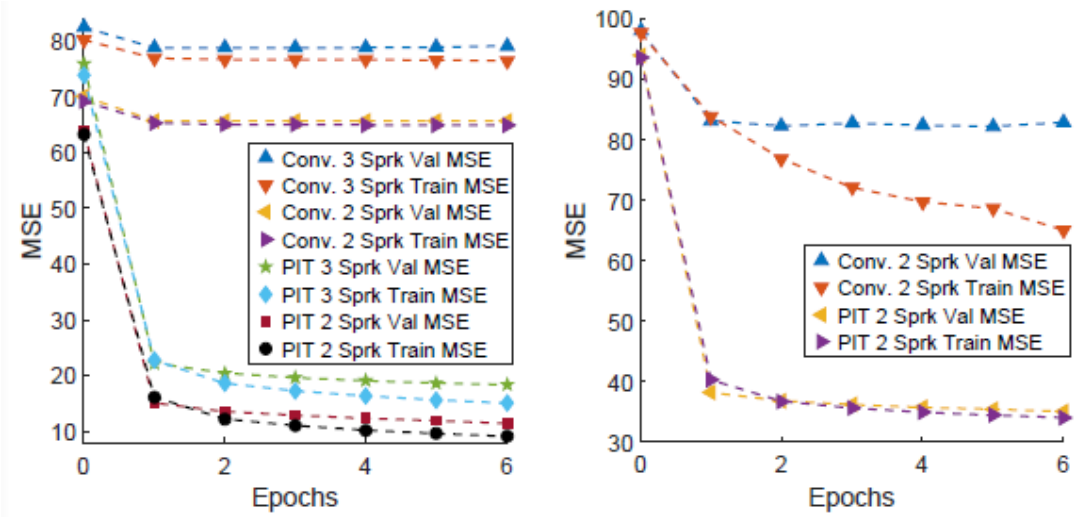


Fig. 2. MSE over epochs on the Danish (left) and WSJ0 (right) training and validation sets with conventional training and PIT.

可以看出，PIT 比起传统的方法（CNN），验证集上的 MSE 下降更快。

Table 1. SDR improvements (dB) for different separation methods on the WSJ0-2mix dataset.

Method	Input\Output window	Opt. Assign CC	Opt. Assign OC	Def. Assign CC	Def. Assign OC
Oracle NMF [23]	-	-	-	5.1	-
CASA [23]	-	-	-	2.9	3.1
DPCL [23]	100\100	6.5	6.5	5.9	5.8
DPCL+ [24]	100\100	-	-	-	10.3
PIT-DNN	101\101	6.2	6.0	5.3	5.2
PIT-DNN	51\51	7.3	7.2	**5.7**	**5.6**
PIT-DNN	41\7	10.1	10.0	-0.3	-0.6
PIT-DNN	41\5	**10.5**	**10.4**	-0.6	-0.8
PIT-CNN	101\101	8.4	8.6	**7.7**	**7.8**
PIT-CNN	51\51	9.6	9.7	7.5	7.7
PIT-CNN	41\7	10.7	10.7	-0.6	-0.7
PIT-CNN	41\5	**10.9**	**10.9**	-0.8	-0.9
IRM	-	12.3	12.5	12.3	12.5

Table 2. SDR improvements (dB) based on optimal assignment for DNNs trained with Danish-2mix.

Method	Input\Output window	CC	OC	WSJ0 OC
IRM	-	17.2	17.3	13.2
PIT-DNN	101\101	9.00	8.61	4.29
PIT-DNN	61\61	9.87	9.44	5.17
PIT-DNN	31\31	11.1	10.7	6.18
PIT-DNN	31\7	14.0	13.8	9.03
PIT-DNN	31\5	14.1	13.9	9.29

从上面的第一个表中可以看出，首先，如果没有说话人追踪（def. assign），PIT 可以达到和原始的 DPCL 相似或者比它更好的表现性能，但稍差于 DPCL+。其次，减小通过减小输出的窗的大小，分离的性能可以得到明显提升（如果使用说话人追踪 opt. assign）。然后，PIT 在 unseen speakers 的泛化效果比较好，因为 OC 和 CC 测试集的效果非常接近。

第二个表中的报告的虽然是 Danish-2mix，但却基于 WSJ0 的训练数据 si_yr_s 进行训练。一个有趣的现象是，尽管系统没听过英语语音，但却能良好分离，这一结果显示了 PIT 不仅可以跨说话人泛化还可以跨语言泛化。