论文标题

Recurrent deep stacking networks for supervised speech separation

要点

本文基于 DNN/CNN/RNN+ mask estimation 等模型继续改进，提出了前驱帧的掩码作为额外输入来预测当前帧的办法，这样一来，输出中的全文信息将会被显式利用，这种模型可以使用 RNN 来建模，但优化过程会很困难，并很可能引起“梯度消失”的问题。我们因此提出了一种循环的深度堆栈策略，在这种策略下，前驱帧的掩码会在每一个周期的训练中得到更新，并且更新的掩码会作为额外的输入用于下个周期的 DNN 训练。在测试阶段，我们使用若干前驱帧来预测当前帧的掩码（通过将上述的堆栈式 DNN 转化为序列式的 RNN），recurrent connection 从前驱帧的输出单元连接到用于当前帧的输入（类似于语言模型），此外，我们使用 L1 损失作为损失函数，使用 SDR 作为评价标准，数据集则是 CHiME-2（task 2）数据集。

监督语音分离的关键点是使用一个学习机器（DNNs，CNNs，LSTMs），从混合的音频（发言）中估计出 IRM（软掩码），从而使用 IRM 及点乘计算复原出源音频的时频域特征。

传统的 IRM 实际上是 Wiener Filer 的平方根，在这篇文章的研究中，提出了如下一种不一样的掩码作为训练的目标：

$$M_{t,f} = min\left(1, \frac{S_{t,f}^2}{Y_{t,f}^2}\right)$$

$S_{t,f}^2$ 和 $Y_{t,f}^2$ 分别代表语音和混合音频在某个 T-F 单元上的能量，M 中每个元素的取值范围在 0 和 1 之间，这样一来神经网络的输出层就可以使用 **sigmoid** 来与作为训练目标的 M 配合起来，并且该掩码使用起来比 IRM 能更接近地还原出干净语音的能量谱。

传统上 mask estimation 会使用 L2 损失，但是这篇文章认为 L1 损失会更好，下面给出 L1 损失的定义和它的误差梯度：

$$Loss = \frac{1}{T}\sum_{t}\sum_{f}\left|M^{*}_{t,f} - M_{t,f}\right|$$

$$\frac{\partial Loss}{\partial M^{*}_{t,f}} = \frac{1}{T}\left(1\left[M^{*}_{t,f} > M_{t,f}\right] - 1\left[M^{*}_{t,f} \le M_{t,f}\right]\right)$$

在传统的理想二值掩码中有一个假设是：众多的 T-F 单元中，每一个单元仅被一个信号源占有，在该假设下，如果绘制理想掩码矩阵的直方图，很多掩码单元的值会集中到 0 和 1 上，并且以指数下降的速度分别从 0 和 1 下降到 0.5（在不考虑室内回声的情况下），因此，可以认为，误差项的分布也呈现出 Laplacian 分布的形式。而实际上，在实验中，如果用 L1 损失训练，误差，验证集上的误差直方图确实接近于 Laplacian，如果用 L2 损失训练，验证集上的误差直方图并不近似于高斯分布。复原出源音频的的方法如下：

$$\hat{Y}^{2} = M^{*} \otimes Y^{2}$$

下面介绍循环**深度堆栈网络（Recurrent Deep Stacking Networks）**，其输入时混合音频特征与若干前驱帧的掩码值的组合：

$$< M^{*}_{t-w}, \ldots, M^{*}_{t-1}, x_{t-w}, \ldots, x_{t}, \ldots, x_{t+w} >$$

W 是半个窗的大小，x_t 和 M^{*}_{t} 分别代表提取 t 时刻处的混合音频特征以及掩码值，输出是理想掩码的中心帧，使用 $M^{*}_{t-w}, \ldots, M^{*}_{t-1}$ 作为额外输入来预测 M_t，那么输出的全文信息可以被显式利用起来。

整个训练的过程如下图所示，我们在每个训练周期结束时更新**所有**的 M^{*}_{t}（t=1,2,3...所有都更新），并使用更新过的 M^{*}_{t} 作为额外的输入特征用于 DNN 的下个周期的训练。这个效果和隐式地堆起 N 层的 DNNs 是类似的，N 代表训练的周期数，每个周期中的 DNN 模型可以看成栈中的一个模块。DNN 堆得更多，更多输入的全文信息可以被利用。

尽管我们在训练阶段堆了很多层 DNN，但测试时没有必要全部保存，仅需白村最后一个训练周期的 DNN，并在测试阶段将它表示为一个 RNN（从输出连接到输入，类似于语言模型），因此预测时可以按序列模型进行预测。，更具体的说，由于我们的策略是仅使用过去估计出来的掩码，因此在预测当前帧时所有的输入

特征是可用的。特别的，在预测第一帧的掩码时，我们需要将前驱帧的掩码值设置为 0。

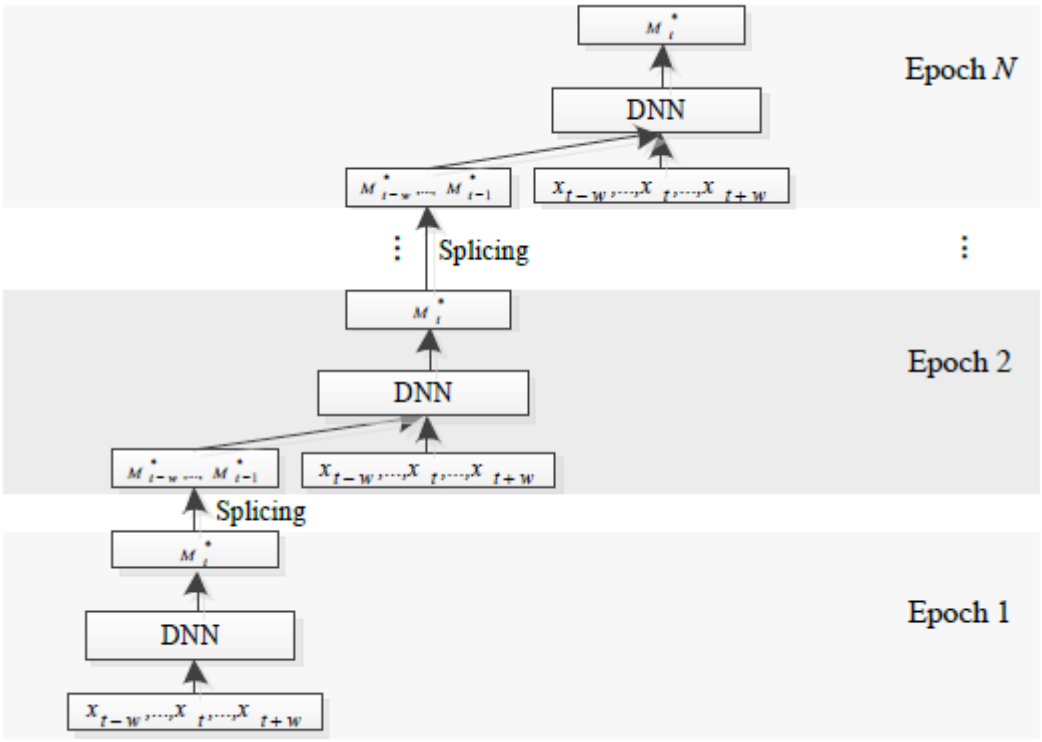

Fig. 1. Illustration of the training process of the proposed approach

在这个模型中，我们设置了 4 个隐藏层，每一层都有 2048ELUs。ELUs 比 RELUs 的收敛速度更快，此外，dropout 的比率设为 0.05，上下文窗大小 w=9。此外，STFT 中的窗长度设置为 25ms，帧位移 10ms，n_fft=512，因此，对数能量谱特征的维度是 257，同时也是 DNN 输出的维度，所有的特征都经过 Z 分数标准化，我们在每次更新中重新计算了估计的掩码的均值和方差。注意到每次更新我们都要使用**所有的训练数据**，梯度下降算法使用 AdaGrad 并带上 Momentnum 项，一共训练 30 个周期，学习率在前面 10 个周期固定为 0.005，剩余周期线性下降到 0.0001，momentnum 项在前 5 个周期从 0.1 线性提升到 0.9，后续固定围殴 0.9。

实验数据集

CHiME-2 dataset (task-2)，回响，混合音频是通过混合 WSJ0-5k corpus，binaural room impulse responses (BRIRs)以及噪音（SNR 等级在-6-9dB）形成的。训练数据中一共有 7138 段发言（14.5h），验证数据集中每个 SNR 等级上有 409 段发言

（4.5h），测试数据集中每个 SNR 等级上有 330 段发言。（4h）实验使用的评估标准有 Short-Time Objective Intelligibility (STOI), Perceptual Estimation of Speech Quality (PESQ)和 Signal-to-Distortion Ratio (SDR).其中，STOI 和 PESQ 是语音可理解性和质量的常见度量指标。

实验结果分析

从下面几个表可以看出，使用 L1 损失进行训练可以带来更好的 STOI，PESQ 和 SDR 分数（在各个 SNR 等级上）。图 2 显示了掩码值主要分布在 0 和 1 之间，而图 3 显示了误差项在使用 L1 损失时服从拉普拉斯分布（尖端和后尾），而使用 L2 损失时误差项并没有很像高斯分布，因此使用 L1 损失会更好。

TABLE I. COMPARISON OF SDR SCORES ON TEST SET

Approaches	Loss functions	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
Unprocessed	-	-2.55	-1.12	1.11	2.78	4.48	5.78	1.75
DNN	L_2	8.94	10.42	12.28	13.90	15.60	17.51	13.11
DNN	L_1	9.76	11.12	12.88	14.43	16.05	17.89	13.69
Recurrent Deep Stacking Networks	L_1	10.35	11.70	13.43	14.91	16.46	18.25	14.18
Recurrent Deep Stacking Networks	+Signal Approximation	**10.76**	**12.06**	**13.69**	**15.08**	**16.57**	**18.33**	**14.41**
LSTM [16]	Signal Approximation	10.46	11.85	13.40	14.86	16.34	18.07	14.17

TABLE II. COMPARISON OF PESQ SCORES ON TEST SET

Approaches	Loss functions	-6dB	-3dB	0dB	3dB	6dB	9dB
Unprocessed	-	2.138	2.327	2.492	2.662	2.854	3.049
DNN	L_2	2.791	2.940	3.076	3.217	3.356	3.506
DNN	L_1	2.888	3.049	3.186	3.321	3.449	3.586
Recurrent Deep Stacking Networks	L_1	2.996	3.162	3.295	3.432	3.533	3.663
Recurrent Deep Stacking Networks	+Signal Approximation	**3.014**	**3.181**	**3.315**	**3.448**	**3.559**	**3.685**
Phoneme-specific Speech Separation [30]	Signal Approximation	2.731	2.884	3.011	3.146	3.284	3.430

TABLE III. COMPARISON OF STOI SCORES ON TEST SET

Approaches	Loss functions	-6dB	-3dB	0dB	3dB	6dB	9dB
Unprocessed	-	0.737	0.778	0.813	0.852	0.881	0.909
DNN	L_2	0.871	0.895	0.914	0.932	0.944	0.957
DNN	L_1	0.878	0.901	0.919	0.936	0.946	0.959
Recurrent Deep Stacking Networks	L_1	**0.886**	**0.909**	**0.925**	**0.940**	**0.950**	**0.961**
Recurrent Deep Stacking Networks	+Signal Approximation	0.884	0.907	0.924	0.939	0.948	0.959
Phoneme-specific Speech Separation [30]	Signal Approximation	0.861	0.886	0.905	0.922	0.935	0.949

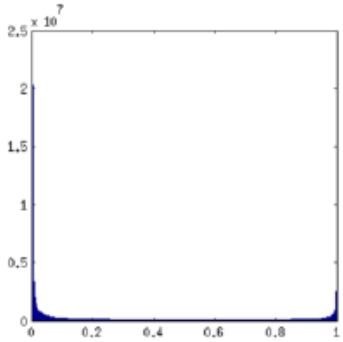

Fig. 2. The histogram of all the values in the ideal masks on the -6 dB subset of the validation set.

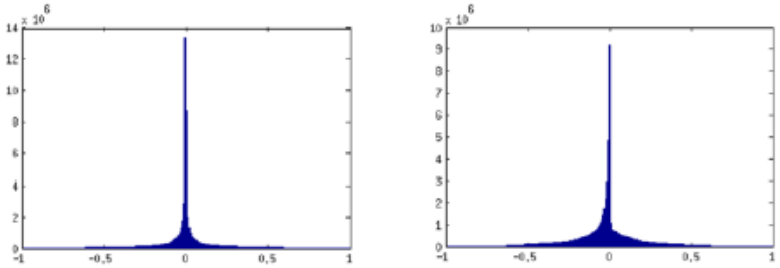

Fig. 3. Error histograms on the -6 dB subset of the validation set. The left histogram is obtained using the DNN trained with the L_1 loss, and the right histogram is obtained using the DNN trained with the L_2 loss.

我们的模型在训练的时候不完全使用 L1 损失，而是一开始使用 L1 损失，后来使用 signal approximation 损失，可以看在表格第 3 和 4 个 entry 中看出，这样训练比单独使用 signal approximation 损失要好，而且在这种训练方式下，SDR 和 PESQ 比单独使用 L1 训练时高，但 STOI 反而轻微下降。