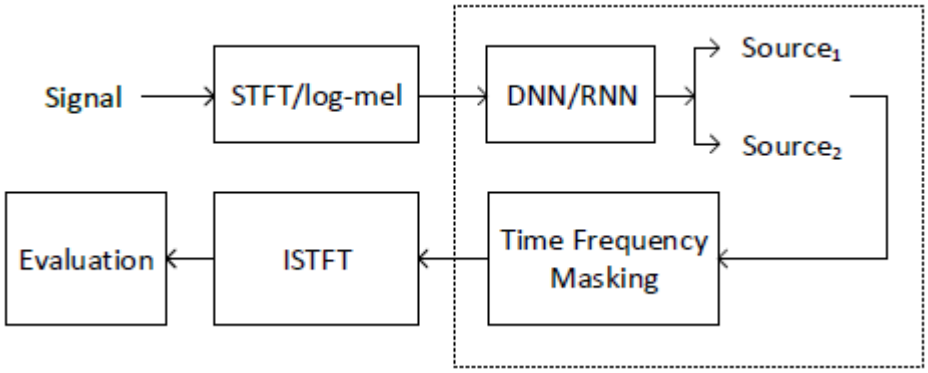
论文标题

DEEP LEARNING FOR MONAURAL SPEECH SEPARATION

要点

在语音分离任务中，一般的 NMF 模型属于线性的模型，非线性的模型如 DNNs 和 RNNs 会表现更优。



基本的处理思路是：给出 noisy signal x，使用 DNN/RNN 得到 clean speech y。通过探索一种使用 DNN 和 RNN 来学习最优的隐藏层表示来复原出 target spectra。网络的输入x_t代表混合音频在一定窗大小内的 spectra 或者 log-mel filterbank features 的拼接。输出$\hat{\mathbf{y}}_{1_t}$和$\hat{\mathbf{y}}_{2_t}$分别代表不同源的 spectra。RNN 将基于当前的输入x_t和前一个状态的隐藏层激活值来得到当前的输出：

$$h^l(\mathbf{x}_t) = f\left(\mathbf{W}^l h^{l-1}(\mathbf{x}_t) + \mathbf{b}^l + \mathbf{U}^l h^l(\mathbf{x}_{t-1})\right)$$

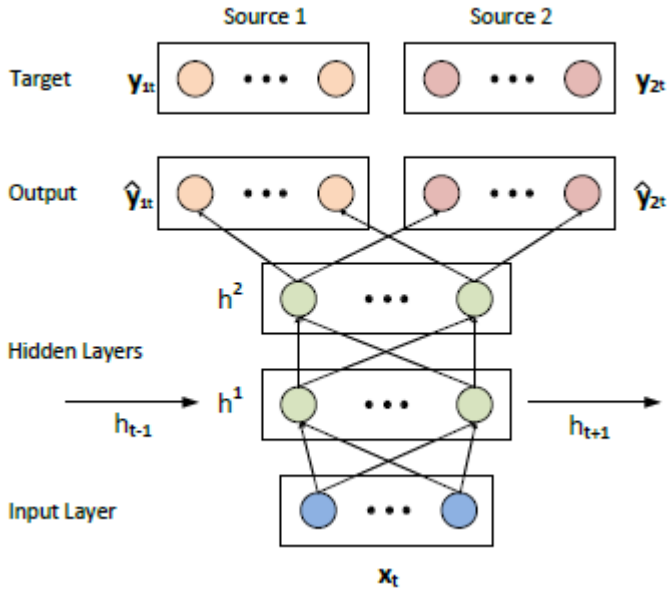
如果是 DNN，上式将简化为：

$$h^1(\mathbf{x}_t) = f(\mathbf{W}^1 \mathbf{x}_t + \mathbf{b}^1)$$

f()为 relu 函数，特别注意的是这里的输出层比较特殊，只是一个线性层：

$$\hat{\mathbf{y}}_t = \mathbf{W}^l h^{l-1}(\mathbf{x}_t) + \mathbf{c}$$

$\hat{\mathbf{y}}_t$是预测标签，$\hat{\mathbf{y}}_{1_t}$和$\hat{\mathbf{y}}_{2_t}$拼接。下图表示了上述算法的处理模型：



上面的模型有一个坏处就是缺少约束：预测结果的加和等于原始的输入。为了解决这一个问题，可以使用掩码的方法为原始混合音频添加上述约束，掩码一般有二值（硬）掩码和软掩码，它们分别是：

$$\mathbf{M_b}(f) = \begin{cases} 1 & |\hat{\mathbf{y}}_{\mathbf{1}_t}(f)| > |\hat{\mathbf{y}}_{\mathbf{2}_t}(f)| \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{M_s}(f) = \frac{|\hat{\mathbf{y}}_{\mathbf{1}_t}(f)|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}(f)| + |\hat{\mathbf{y}}_{\mathbf{2}_t}(f)|}$$

一旦掩码确定，复原出原来的源可以使用如下方法：

$$\hat{\mathbf{s}}_{\mathbf{1}_t}(f) = \mathbf{M}(f)\mathbf{X}_t(f)$$
$$\hat{\mathbf{s}}_{\mathbf{2}_t}(f) = (1 - \mathbf{M}(f))\,\mathbf{X}_t(f)$$

作为改进的手段之一，我们可以把掩码函数集成到网络中，作为额外的一层：

$$\tilde{\mathbf{y}}_{\mathbf{1}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{1}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{X}_t$$
$$\tilde{\mathbf{y}}_{\mathbf{2}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{2}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{X}_t$$

网络参数的优化过程将在$\tilde{\mathbf{y}}_{\mathbf{1}_t}, \tilde{\mathbf{y}}_{\mathbf{2}_t}$和$\mathbf{y}_{\mathbf{1}_t}, \mathbf{y}_{\mathbf{2}_t}$之间进行。如果想进一步平滑结果，也可以在调用一次掩码函数得到估计的分离出来的 spectra$\tilde{\mathbf{S}}_{\mathbf{1}_t}$和$\tilde{\mathbf{S}}_{\mathbf{2}_t}$，重构出时域信号可以用 ISTFT。

训练时常用的损失函数是平方损失：

$$||\hat{\mathbf{y}}_{\mathbf{1}_t} - \mathbf{y}_{\mathbf{1}_t}||_2^2 + ||\hat{\mathbf{y}}_{\mathbf{2}_t} - \mathbf{y}_{\mathbf{2}_t}||_2^2$$

不过由于我们得到高的 SIR（体现当前的源不受其它源干扰的能力），用如下的

损失函数会更加合适（discriminative training criterion）：

$$||\hat{\mathbf{y}}_{\mathbf{1}_t}-\mathbf{y}_{\mathbf{1}_t}||_2^2-\gamma||\hat{\mathbf{y}}_{\mathbf{1}_t}-\mathbf{y}_{\mathbf{2}_t}||_2^2+||\hat{\mathbf{y}}_{\mathbf{2}_t}-\mathbf{y}_{\mathbf{2}_t}||_2^2-\gamma||\hat{\mathbf{y}}_{\mathbf{2}_t}-\mathbf{y}_{\mathbf{1}_t}||_2^2$$

实验数据集

TIMIT corpus，一共有 8 段来自一男性和一女性的 TIMIT sentences 用于训练，剩余的 sentences 中，有一段来自男性和一段来自女性的 sentences 被用于验证，其它的全部用于测试。测试的 sentences 被加和到一起形成混合音频，SNR 为 0dB。为了增加训练样本的量，我们在时域上循环切换男性说话者的信号，并将这些信号与女性说话者的发言混合到一起。

特征提取

Spectral 和 log-mel filterbank features，前者提取时使用的帧大小为 n_fft=1024，帧重叠 50%，后者表现会更加良好，40 维的 log-mel 表示及其一二阶的派生特征会被用于实验。另外，我们发现 32ms 的窗大小和 16ms 的帧位移表现会更加良好。与 output spectra 相关的输入帧比率在提取时使用 n_fft=512。

度量

SIR，SDR，SAR，理论上说，它们都是越高越好，SIR 体现出对干扰的抑制能力，SAR 体现出分离过程体现出来的加工能力，SDR 则反映总体的性能表现。

实验结果分析

基线：n_fft=512 和 n_fft=1024 的标准 NMF，使用 KL 散度更新。基向量集合 Wm 和 Wf 分别来自男性和女性的训练数据，在计算出 Hm 和 Hf 两大激活矩阵后，硬掩码或者软掩码将被用于预测目标的 magnitude spectrogram，下图展现出不同的基向量数目（10，30，50），不同的 STFT 窗大小的结果（10 次不同随机初始化的平均）。

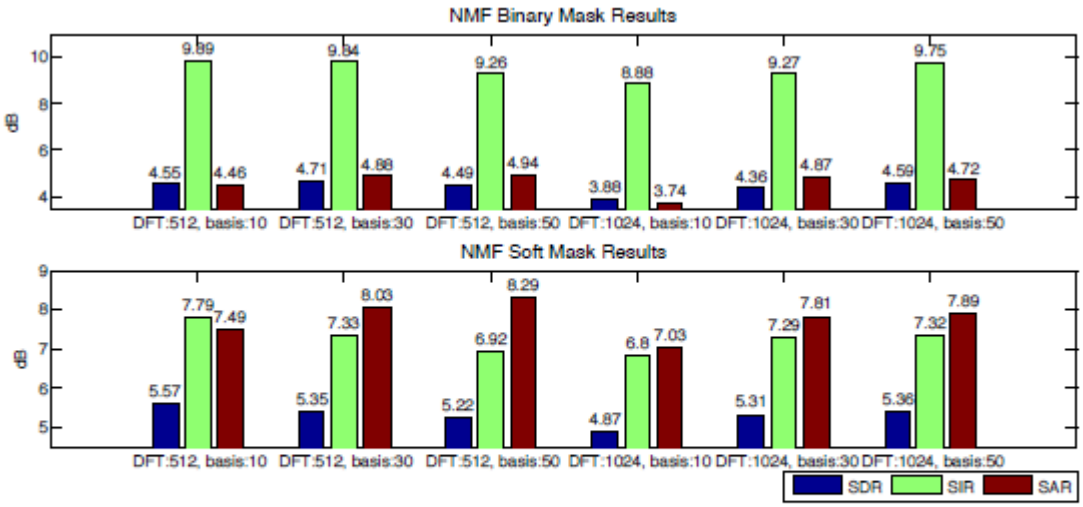


Fig. 3: NMF results with the 512-point and 1024-point STFT and basis vector sizes (10, 30, 50) using binary and soft time-frequency masking

神经网络模型使用两个隐藏层，分别 150 个神经元，从下图可以看出，使不适用相邻的帧差别并不大，类似于 NMF，二值掩码使用了硬决策来促进分离，，因此导致 SIR 会偏高，但也导致了低的 SAR。软掩码则相反，SDR 和 SAR 偏高而低 SIR。在图中的第 1，2 列，我们比较了使用 spectra 作为特征的 DNN 和 RNN，但区别不大，使用其它的语音特征也一样不明显。在列 2，3，6，7 和列 4，5，8，9 之间，我们对比了 spectra 和 log-mel filterbank 两种输入特征，可以看出，在 2，3，4，5 列中（没有使用 joint training），spectra 比 log-mel 要好，在 6，7，8，9 列中（使用了 joint training），log-mel 比 spectra 的结果更好，2 和 3 比较，4 和 5 比较，6 和 7 比较，8 和 9 比较，使用 discriminative training criterion 的效果更好。γ 取值在 0.05-0.2 之间可以在提高 SIR 的情况下保持 SAR 和 SDR，列 2，3，4，5 和列 6，7，8，9 的比较可以看出，joint training 可以改善表现性能对比 NMF 的结果，我们的最优模型大约提升了 3.8-4.8dB 和 3.9-4.9dB（分别使用二值掩码和软掩码）的 SIR，SDR 和 SAR 也有相应的提升。在未来的工作中，将会进一步探索更多时间信息用于神经网络，我们提出的模型可以被用于其它应用例如 robust ASR。

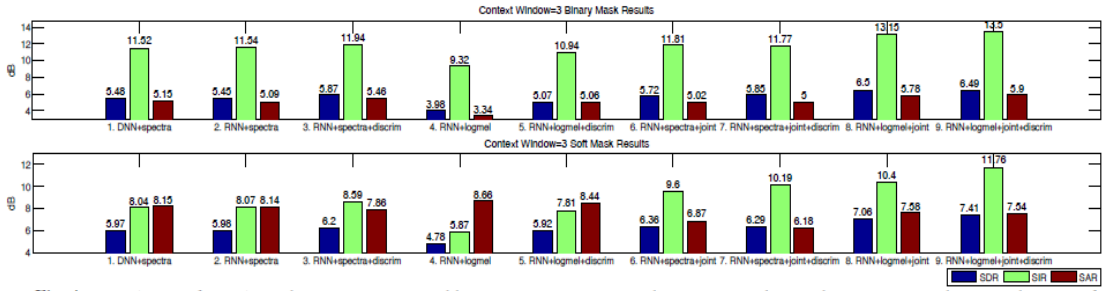


Fig. 4: Neural network results with concatenating neighboring 1 frame as input, where "joint" indicates the joint training between the network and the soft masking function, and "discrim" indicates the training with discriminative objectives

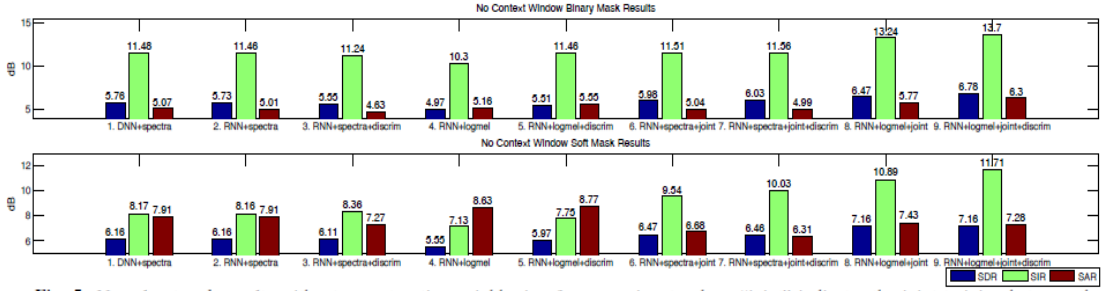


Fig. 5: Neural network results without concatenating neighboring frames as input, where "joint" indicates the joint training between the network and the soft masking function, and "discrim" indicates the training with discriminative objectives