论文标题

Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization

要点

NMF

给出非负的数据矩阵$\mathbf{V} \in \mathbb{R}_+^{m \times n}$，NMF 的目标是寻找到两个非负矩阵，$\mathbf{W} \in \mathbb{R}_+^{m \times r}$和$\mathbf{H} \in \mathbb{R}_+^{r \times n}$，从而最小化误差$D(\mathbf{V}, \mathbf{WH})$，D 是某种散度度量。在音频源分离的应用中，V 代表原始的 magnitude spectrogram. W 中的列代表了录音的特征化的 spectra。H 包含了这些基本的 spectra 的激活值。

度量方法一般用 KL 散度：

$$D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j}$$

更新的规则：

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}$$

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T}$$

注意 1 代表 ones 矩阵（全 1 的矩阵）

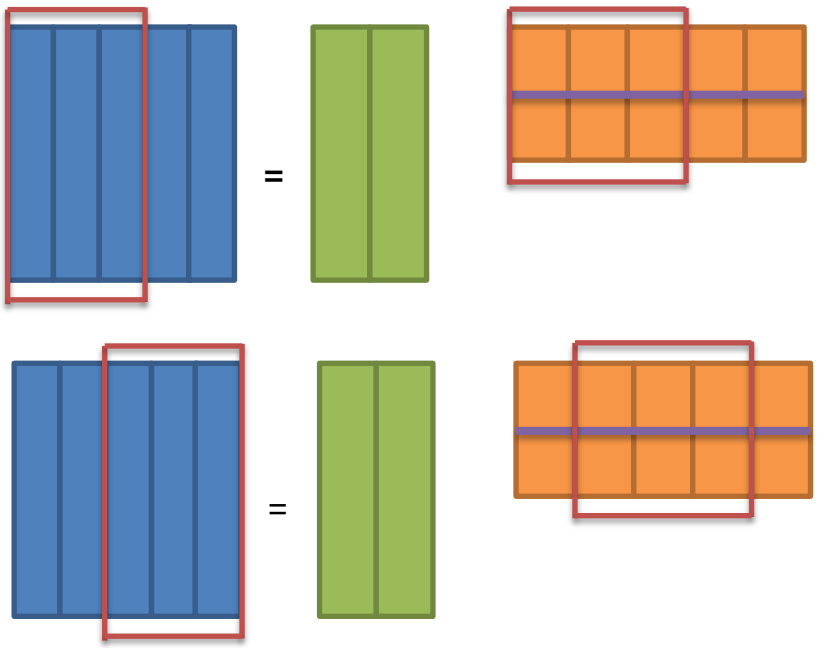
令 k=1,2...K 代表一共有 K 个源，那么：

$$\mathrm{V} \approx \mathbf{WH} = \sum\nolimits_{k} \mathbf{W_k H_k}$$

加粗部分是矩阵加法，$\mathbf{W_k}$，$\mathbf{H_k}$分别是**W**，**H**的子矩阵（加减法意义上），因此分离阶段复原出原来的各个音频源如下（比例的形式）：

$$\mathbf{V_k} = \mathbf{V} \cdot \frac{\mathbf{W_k H_k}}{\mathbf{WH}}$$

On-Line Superivised NMF



这种学习方法中，假设所有的音频源是已知的（从开始的 spectrum 到结束的 spectrum），因此很适合用在线的学习方式（增量的学习方式），换言之可以使用滑动窗口一帧帧处理，分离阶段 W 矩阵由隔离的源音频矩阵合成回来。用此方法时，W 矩阵会保持为一个常量矩阵，只有 H 矩阵增量更新。

$$\mathbf{h}_{:,t} \leftarrow \mathbf{h}_{:,t} \cdot \frac{\mathbf{W}^T \frac{\mathbf{v}_{:,t}}{\mathbf{W}\mathbf{h}_{:,t}}}{\mathbf{W}^T \mathbf{1}_{:,t}}.$$

小注：spectrum 为单数，spectra 和 spectrums 为其复数形式

On-Line Semi-supervised NMF

这种学习方法中，认为一种源是不可知的（可以是说话人的声音，也可以是噪音，很多时候是噪音），将未知源的 Wk 矩阵随机初始化，每一帧会在训练中更新，已知源的 Wk 矩阵则保持为常量矩阵。也正因为未知源的数据矩阵未知，混合音频的 H 矩阵和 W 矩阵会产生相互依赖，会因此不适合增量学习，而应该考虑全文信息（使用所有观测到的 spectra）。

Real-Time Implementation

Real-Time Implementation 是对 On-Line Semi-supervised NMF 的实时实现，这里有一个参数称为 delay 参数，它指示了滑动窗口要输出的帧的位置。如果该参数设置为 0，只考虑频谱中过去的上下文，否则会窗口外面（未来）的观测点。

激活值的精确性不仅依赖于对 W 矩阵的估计（可以由滑动窗口控制），还有迭代的次数（随着 delay 参数的增加而增加），总延迟 L=(d+1)s，d 为 delay 参数的大小，s 为帧的长度。

实验数据集

Buckeye database 采样频率为 16Hz，测试时的分析的帧的长度为 32ms，帧重叠 50%，实验中用 12 段随机选取的语音片段，长度在 3-20s 不等，每个训练序列用 20 个来自同一个说话人的语音片段拼接到一起，最后长度在 1.5min-5.5min 之间。

除了语音部分，还有噪音部分，大部分噪音来自CHiME噪声录音（拼接1024个0.5s的短片段），这部分噪音涵盖了大部分的训练样本，另外，为了加强泛化效果，17min的训练序列由SiSEC 2010 noisy speech database，SPIB noise database，street noise（来自the soundcities website）组成，两部分噪声分别代表matched和mismatched训练噪声，在上面提到的算法中，On-Line Superivised NMF的成分数目：语音和噪声分别为c_s=c_n=50；on-line semi-supervised NMF，语音的成分为$c_s = 50$，噪声的成分数目则可以改变。

Table 1. Tested values of the parameters for the on-line semi-supervised NMF system

Parameter		Tested Values
c_s	number of speech components	{50}
c_n	number of noise components	{1,2,4,8,12,16}
ℓ	sliding window length	{2,4,6,8,12,16,20,25,30}
d	delay	{0,1,2,3,4,5,6,7}
n	number of optimization iterations	{1,2,4,8,16,32,64}

几种评估标准：

the Source to Distortion Ratio (SDR)

the Source to Interference Ratio (SIR)

the Source to Artifact Ratio (SAR)

作为对比，提供一个离线版本的模型，$c_n \in \{1, 2, 4, 8, 12, 16, 20, 25, 30, 35, 40, 45, 50\}$，其中30最大化了测试数据库中的平均SDR，该模型的最优SDR为5.2dB，将作为一个基线。

实验结果分析

Supervised NMF的结果如下，SIR随着迭代次数的增加而增加，但是，对干扰的削弱也需要更多的加工代价，因此SAR同时也在下降。一种比较好的trade-off是用一次迭代实现该模型，得到4.2dB的结果，比起原始的混合音频，有0.6dB的提升，尽管最优的迭代次数与数据是独立的，但这也显示了一个比较小的迭代次数可以得到一个比较满意的分离。

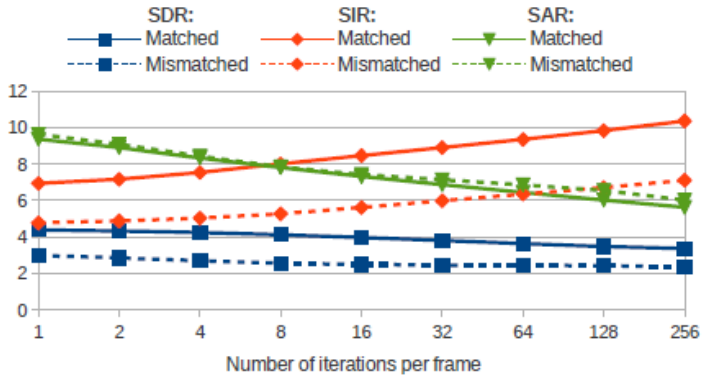


Fig. 1. Average source separation criteria (dB) for the supervised NMF systems, trained on the matched and mismatched noise

我们的模型显示了一个具有足够多的噪声的模型在分离上的重要性，而且实际上supervised NMF模型比off-line semi-supervised模型要好。另外，这两种模型的SAR粗略相同，但使用“matched” noise训练带来更高的SIRs，提升了2dB左右。下面的图2-4展示了on-line semi-supervised系统的一些表现，可以看出它比supervised NMF轻微好一些，最好的SDR大约为4.4dB。

最好的分数下的参数配置为：$c_n = 8,\ \ell = 20,\ d = 0\ \ i = 1$，图2显示出随着迭代次数的增加，SIR在下降，这很可能是由于过拟合引起的（更新中的component试图将语音和噪声一起学习进去）。而图3显示出随着窗大小的增大，SIR下降而SAR反而改善，这可以认为模型对环境的适应能力变差了些（抗干扰能力）但应对过拟合问题的鲁棒性更强，而图3噪声中噪声成分变化带来的效应却恰好相反，因此c_n =8，l =20和336ms的滑动窗口是一个合理的trade-off。图5中delay parameter的改变基本不会给模型带来任何变化，因此设置为0即可。由于迭代次数仅需1次，相应的好处是带来计算复杂度的降低。

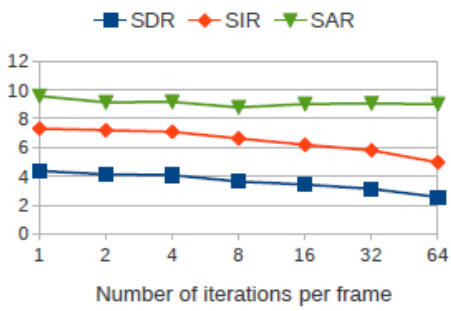


Fig. 2. Criteria (dB) as a function of i for constant $c_n = 8$, $\ell = 20$ and $d = 0$

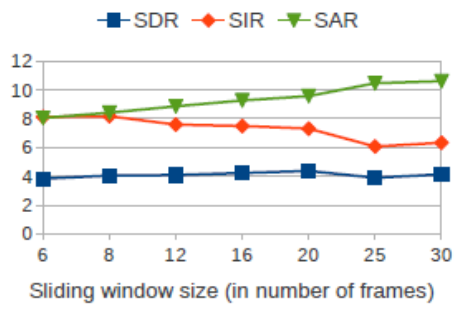


Fig. 3. Criteria (dB) as a function of ℓ for constant $c_n = 8$, $d = 0$ and $i = 1$

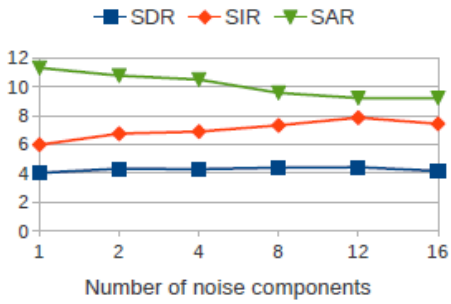


Fig. 4. Criteria (dB) as a function of c_n for constant $l = 20$, $d = 0$ and $i = 1$

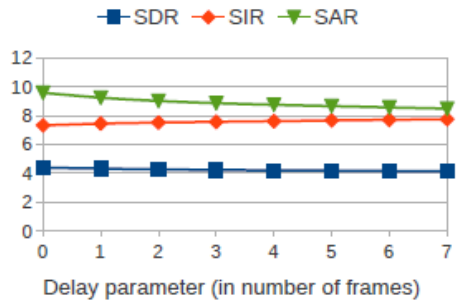


Fig. 5. Criteria (dB) as a function of d for constant $c_n = 8$, $\ell = 20$ and $i = 1$