第 38 卷 第 4 期
2019 年 7 月

应用声学
Journal of Applied Acoustics

Vol. 38, No. 4
July, 2019

◇ 李启虎院士八十华诞学术论文 ◇

基于字典学习和稀疏表示的单通道语音增强算法综述*

叶中付† 朱媛媛 贾翔宇

(中国科学技术大学信息科学技术学院 合肥 230027)

摘要 如何从带噪语音信号中恢复出干净的语音信号一直都是信号处理领域的热点问题。近年来研究者相继提出了一些基于字典学习和稀疏表示的单通道语音增强算法，这些算法利用语音信号在时频域上的稀疏特性，通过学习训练数据样本的结构特征和规律来构造相应的字典，再对带噪语音信号进行投影以估计出干净语音信号。针对训练样本与测试数据不匹配的情况，有监督类的非负矩阵分解方法与基于统计模型的传统语音增强方法相结合，在增强阶段对语音字典和噪声字典进行更新，从而估计出干净语音信号。该文首先介绍了单通道情况下语音增强的信号模型，然后对 4 种典型的增强方法进行了阐述，最后对未来可能的研究热点进行了展望。

关键词 单通道语音增强，稀疏表示，字典学习

中图法分类号: TN912.3 **文献标识码:** A **文章编号:** 1000-310X(2019)04-0645-08

DOI: 10.11684/j.issn.1000-310X.2019.04.022

Review for speech enhancement algorithms based on dictionary learning and sparse representation

YE Zhongfu ZHU Yuanyuan JIA Xiangyu

(*Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract How to recover the clean speech signal from the noisy signal has always been a hot issue in the field of signal processing. In recent years, single-channel speech enhancement algorithms based on dictionary learning and sparse representation have been proposed. These algorithms make full use of the sparsity of signals in time-frequency domain and construct the dictionary by learning the structure characteristics of signals. Finally, the clean speech is estimated by projecting the noisy signal in the dictionary. In terms of mismatched training data, a new approach combining the supervised non-negative matrix factorization method with conventional statistical model-based enhancement methods have been proposed, which can update the speech and noise dictionaries in the enhancement stage and estimate the clean speech. This paper first introduces the signal model of speech enhancement under single-channel condition, and then expounds four typical enhancement methods. Finally, the future research directions are prospected.

Key words Single-channel speech enhancement, Sparse representation, Dictionary learning

2019-01-29 收稿; 2019-04-09 定稿

*国家自然科学基金项目 (61671418)

作者简介: 叶中付 (1959-), 男, 安徽桐城人, 教授, 博士生导师, 研究方向: 信号与信息处理。

†通讯作者 E-mail: yezf@ustc.edu.cn

0 引言

在现实生产生活中，噪声的污染无处不在，语音信号的质量和可懂度也因此严重下降，影响人们的主观听觉感受，语音增强技术就是解决这类问题的主要方法之一。然而，传统语音增强算法[1−4]大多局限于抑制平稳噪声，在非平稳环境下增强语音的能力有限。近年来兴起的字典学习和稀疏表示理论由于能够学习到训练样本数据的分布特征和规律，在解决非平稳噪声抑制问题方面取得了丰硕的成果，其中发展比较迅速的两类算法主要包括基于生成性字典学习的算法以及非负矩阵分解(Non-negative matrix factorization, NMF)类算法。

文献[5–6]提出的生成性字典学习算法是第一类算法中的经典算法，首先利用样本数据训练出干净语音字典和噪声字典，然后将带噪语音信号投影到由语音字典和噪声字典组合而成的复合字典上计算相应的稀疏表示系数，从而估计出语音信号。为了更有效地计算稀疏表示系数，该文献将最小角度回归(Least angle regression, LARS)[7]算法拓展为基于相干性准则的最小角回归(Least angle regression with a coherence criterion, LARC)算法，在迭代过程中将信号残差向量和当前选择的字典原子之间的相关性大小作为算法终止的条件，避免了极不相关的信号成分在字典上的无意义投影，提升了算法的效率。然而，这种方法训练出来的语音字典和噪声字典之间的区分性不够好，会导致一定的源混淆，即部分与噪声相关的语音成分会被噪声字典表示，反之亦然。为了增强语音字典和噪声字典的区分性和差异性，学者们往往在字典学习训练阶段添加区分约束项来训练字典[8−9]。文献[10]则考虑了带噪语音信号与语音信号、噪声之间的映射关系，提出了基于互补联合字典学习和稀疏表示的有监督语音增强算法，联合训练混合-语音字典和混合-噪声字典，获得了较好的增强效果。在此基础上，文献[11]提出了两级联合稀疏表示和字典学习的语音增强算法，将由联合字典估计出的语音信号和噪声投影到由语音子字典和噪声子字典组成的复合字典上，进一步进行稀疏表示，实现语音增强。此外，文献[12]分析出不同声源的语音在分帧之后的相关性增加，在字典学习过程中会造成一定的混淆情况，因此在训练阶段通过构建合适的目标函数使各个训练样本的独特成分尽可能多地被相应的子字典进行稀疏表示，而相关性较大的成分则尽可能多地被缓冲子字典表示，然后逐层分离相似成分，降低混淆成分。文献[13]则是依据带噪语音信号在语音字典和噪声字典上的稀疏表示系数大小挑选出贡献较大的语音信号原子，移除贡献较大的噪声原子，提升降噪性能。文献[14]则定义了时频谱幅度的比率掩码(Ratio mask, RM)特征，联合训练信号的时频谱幅度字典和掩码字典，从而估计出带噪语音信号中的语音比率掩码和噪声比率掩码，然后构建不同的掩码滤波器以实现对语音信号的增强。

NMF 算法本质上是一种降维的工具，其基本思想是将一个矩阵近似分解为两个非负矩阵的乘积[15]，在图像表示、音乐标注、信源分离和语音增强等方面都有着广泛的应用。对于经典的有监督NMF类算法[16]，语音信号和噪声的字典(基向量)通常是由训练样本数据学习得到，然后在增强阶段固定语音信号和噪声的字典不变，迭代更新相应的稀疏表示系数(激活系数矩阵)。为了更好地依据数据的特性来进行字典学习，研究者们往往在NMF的代价函数中会引入一些先验知识作为正则化项来约束基矩阵和系数矩阵的更新，例如文献[17]中引入隐马尔可夫模型，文献[18]则挖掘了不同的说话人具有不同的调制率。在文献[19]中，语音信号和噪声的时频谱被认为是一个高斯混合模型，因此约束字典和系数矩阵的乘积为高斯混合模型并将其对数似然函数作为代价函数的正则化项。对于测试阶段出现了训练样本中不包含的噪声情况，文献[20]提出了具有环境自适应性的NMF 方法。首先用有监督NMF算法[16]从带噪语音信号中估计出语音信号和噪声的时频谱幅度值，接着用后验信噪比计算局部稀疏度以衡量每个频点处语音和噪声的分离程度，再将局部稀疏性与最小均方误差滤波器相结合进一步估计噪声。然后，通过带噪语音信号在语音信号字典和噪声字典上的表示系数大小来判断对应的噪声基向量是否需要更新，若需要更新则用估计的噪声对噪声字典进行更新，从而达到了更好的降噪效果。文献[21]则更加全面地考虑了对语音字典和噪声字典的在线更新，将基于统计模型的算法[22]与有监督的NMF算法相结合来挖

掘出在测试阶段带噪语音信号的先验知识，具有较好的灵活性和更高的实用价值。

1 信号模型

考虑单通道情况下的语音增强问题，带噪语音信号指的是被噪声污染的语音信号，基于加性噪声模型，则可以得到

$$x(m)=s(m)+n(m), \tag{1}$$

其中，$x(m)$、$s(m)$ 和 $n(m)$ 分别表示带噪语音信号、语音信号和噪声在 m 时刻的采样点。

对 $x(m)$ 进行短时傅里叶变换 (Short-time Fourier transform, STFT)，由于 STFT 变换的线性特性，将式 (1) 变化到如下的形式：

$$X(k,t)=S(k,t)+N(k,t), \tag{2}$$

其中，k 表示频率点，t 表示时间帧。$X(k,t)$、$S(k,t)$ 和 $N(k,t)$ 分别表示带噪语音信号、语音信号和噪声在时频点 (k,t) 的 STFT 复系数，忽略时频域中的相位信息，带噪语音信号的时频幅度谱近似为

$$|X(k,t)|=|S(k,t)|+|N(k,t)|. \tag{3}$$

将式 (3) 表示为矩阵形式为

$$\boldsymbol{X}=\boldsymbol{S}+\boldsymbol{N}, \tag{4}$$

其中，$\boldsymbol{X}\in\boldsymbol{R}^{K\times T}$、$\boldsymbol{S}\in\boldsymbol{R}^{K\times T}$ 和 $\boldsymbol{N}\in\boldsymbol{R}^{K\times T}$ 分别表示带噪语音信号、语音信号和噪声的时频幅度谱，K 表示频点数目，T 表示时间帧数目。

2 基于字典学习和稀疏表示的单通道语音增强算法

2.1 基于生成性字典学习的语音增强算法

针对非平稳噪声环境下语音增强问题，文献 [6] 提出了生成性字典学习算法，该算法包含训练和增强两个阶段，首先基于干净语音样本 $\boldsymbol{S}^{\mathrm{tr}}$ 和噪声样本 $\boldsymbol{N}^{\mathrm{tr}}$ 训练干净语音字典 $\boldsymbol{D}_s$ 和噪声字典 $\boldsymbol{D}_n$，其目标函数是最小化训练样本在对应字典上的稀疏表示误差：

$$\min_{\boldsymbol{D}_s,\boldsymbol{C}_s}\|\boldsymbol{S}^{\mathrm{tr}}-\boldsymbol{D}_s\boldsymbol{C}_s\|_{\mathrm{F}}^2\quad \text{s.t.}\|\boldsymbol{c}_{s,g}\|_1\leqslant q_s,\ \forall g, \tag{5}$$

$$\min_{\boldsymbol{D}_n,\boldsymbol{C}_n}\|\boldsymbol{N}^{\mathrm{tr}}-\boldsymbol{D}_n\boldsymbol{C}_n\|_{\mathrm{F}}^2\quad \text{s.t.}\|\boldsymbol{c}_{n,g}\|_1\leqslant q_n,\ \forall g, \tag{6}$$

其中，$\|\cdot\|_{\mathrm{F}}$ 表示弗罗贝尼乌斯范数 (Frobenius norm)，$\|\cdot\|_1$ 表示 l_1 范数。$\boldsymbol{c}_{s,g}$ 和 $\boldsymbol{c}_{n,g}$ 分别表示稀疏编码矩阵 $\boldsymbol{C}_s$ 和 $\boldsymbol{C}_n$ 的第 g 列，q_s 和 q_n 分别表示对 $\boldsymbol{c}_{s,g}$ 和 $\boldsymbol{c}_{n,g}$ 的稀疏度约束。需要注意的是，式 (5) 和式 (6) 并不是凸优化问题，文献 [6] 采用 LARC 算法以进行稀疏编码，采用近似 K-SVD(Singular value decomposition, SVD) 算法 [23] 以进行字典学习。

在增强阶段，将语音字典 $\boldsymbol{D}_s$ 和噪声字典 $\boldsymbol{D}_n$ 组合成一个复合字典 $\boldsymbol{D}=[\boldsymbol{D}_s,\boldsymbol{D}_n]$，将带噪语音信号 $\boldsymbol{X}^{\mathrm{te}}$ 投影到复合字典上采用 LARC 算法计算稀疏编码 $\boldsymbol{E}_s$ 和 $\boldsymbol{E}_n$，

$$\begin{bmatrix}\boldsymbol{E}_s\\ \boldsymbol{E}_n\end{bmatrix}\leftarrow \mathrm{LARC}\left(\boldsymbol{D},\boldsymbol{X}^{\mathrm{te}},\mu_{\mathrm{enh}}\right), \tag{7}$$

其中，μ_{enh} 表示 LARC 算法在增强阶段设定的相关性阈值。

最后，将语音字典 $\boldsymbol{D}_s$ 与相应的编码系数 $\boldsymbol{E}_s$ 相乘即可估计出干净语音信号的时频谱幅度 $\widehat{\boldsymbol{S}}$，再结合带噪语音信号的相位进行逆 STFT 变换，即可恢复出干净语音信号的时域信号。

该算法通过训练字典挖掘出信号的结构特征和时频域上的稀疏性，对非平稳噪声具有更好的抑制能力，但是当噪声的结构和语音信号存在相似之处时，如说话人 (babble) 噪声，就会出现部分噪声被语音字典所表示，反之亦然。这种源混淆的情况一旦出现，生成性字典学习算法就会在降噪的同时引入更多的失真，使得增强后的语音信号质量下降。文献 [6] 分析指出，LARC 算法中的相关性阈值 μ_{enh} 可以用于控制降噪性能和失真度之间的权衡：当 μ_{enh} 设置得过小时，则得到的稀疏编码系数会变得非常稀疏，降噪性能会相应地减弱；当 μ_{enh} 设得过大时，估计出的语音信号会有较多的失真成分。然而文献 [6] 并未给出明确的设置规则，μ_{enh} 的设置主要还是依赖于经验调整，这在一定程度上限制了这种算法的发展。

2.2 基于互补联合字典学习和稀疏表示的语音增强算法

文献 [10] 在生成性字典学习的基础上，在训练阶段将干净语音信号样本和噪声样本基于如式 (4) 所示的加性模型进行合成，得到带噪语音信号的训练样本 $\boldsymbol{X}^{\mathrm{tr}}$，利用带噪语音信号和干净语音信号、噪

声之间的映射关系训练互补联合字典，即有

$$\min_{\boldsymbol{D}_s,\boldsymbol{D}_{x1},\boldsymbol{C}_1}\left\|\begin{bmatrix}\boldsymbol{X}^{\mathrm{tr}}\\ \boldsymbol{S}^{\mathrm{tr}}\end{bmatrix}-\begin{bmatrix}\boldsymbol{D}_{x1}\\ \boldsymbol{D}_s\end{bmatrix}\boldsymbol{C}_1\right\|_{\mathrm{F}}^2$$
$$\text{s.t.}\ \|\boldsymbol{c}_{1,g}\|_1\leqslant q,\ \ \forall\, g\,, \tag{8}$$

$$\min_{\boldsymbol{D}_n,\boldsymbol{D}_{x2},\boldsymbol{C}_2}\left\|\begin{bmatrix}\boldsymbol{X}^{\mathrm{tr}}\\ \boldsymbol{N}^{\mathrm{tr}}\end{bmatrix}-\begin{bmatrix}\boldsymbol{D}_{x2}\\ \boldsymbol{D}_n\end{bmatrix}\boldsymbol{C}_2\right\|_{\mathrm{F}}^2$$
$$\text{s.t.}\ \|\boldsymbol{c}_{2,g}\|_1\leqslant q,\ \ \forall\, g\,, \tag{9}$$

其中，$\begin{bmatrix}\boldsymbol{D}_{x1}\ \boldsymbol{D}_s\end{bmatrix}^{\mathrm{T}}$ 被称为混合-语音联合稀疏表示，$\begin{bmatrix}\boldsymbol{D}_{x2}\ \boldsymbol{D}_n\end{bmatrix}^{\mathrm{T}}$ 被称为混合-噪声联合稀疏表示，$\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 为稀疏表示系数，$\boldsymbol{c}_{1,g}$ 和 $\boldsymbol{c}_{2,g}$ 表示 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 的第 g 列，q 是对应的稀疏约束阈值。由于对 $\boldsymbol{X}^{\mathrm{tr}}$ 和 $\boldsymbol{S}^{\mathrm{tr}}$、$\boldsymbol{X}^{\mathrm{tr}}$ 和 $\boldsymbol{N}^{\mathrm{tr}}$ 在相应字典上进行稀疏表示的时候约束的是相同的系数 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$，例如对同一帧的带噪语音信号和干净语音信号进行稀疏表示的时候，由于约束的稀疏表示系数 $\boldsymbol{c}_1$ 相同，则说明是采用 $\boldsymbol{D}_{x1}$ 和 $\boldsymbol{D}_s$ 中的同一列原子进行稀疏表示的，这样对于存在映射关系的 $\boldsymbol{X}^{\mathrm{tr}}$ 和 $\boldsymbol{S}^{\mathrm{tr}}$ 的每一帧，都反映在 $\boldsymbol{D}_{x1}$ 和 $\boldsymbol{D}_s$ 的每一列原子上，即 $\boldsymbol{D}_{x1}$ 和 $\boldsymbol{D}_s$ 在原子级上的映射关系就代表了带噪语音信号和干净语音信号之间的关系。

在增强阶段，对用于测试的带噪语音信号 $\boldsymbol{X}^{\mathrm{te}}$ 进行两路的稀疏表示：

$$\boldsymbol{E}_1^*=\arg\min_{\boldsymbol{E}_1}\left\|\boldsymbol{X}^{\mathrm{te}}-\boldsymbol{D}_{x1}\boldsymbol{E}_1\right\|_{\mathrm{F}}^2$$
$$\text{s.t.}\ \|\boldsymbol{e}_{1,g}\|_1\leqslant q,\forall g\,, \tag{10}$$

$$\boldsymbol{E}_2^*=\arg\min_{\boldsymbol{E}_2}\left\|\boldsymbol{X}^{\mathrm{te}}-\boldsymbol{D}_{x2}\boldsymbol{E}_2\right\|_{\mathrm{F}}^2$$
$$\text{s.t.}\ \|\boldsymbol{e}_{2,g}\|_1\leqslant q,\forall g\,. \tag{11}$$

利用得到的稀疏表示系数 $\boldsymbol{E}_1^*$ 和 $\boldsymbol{E}_2^*$，可以估计出语音信号和噪声：

$$\widehat{\boldsymbol{S}}_1^{\mathrm{est}}=\boldsymbol{D}_s\boldsymbol{E}_1^*, \tag{12}$$

$$\widehat{\boldsymbol{N}}_2^{\mathrm{est}}=\boldsymbol{D}_n\boldsymbol{E}_2^*. \tag{13}$$

基于式(4)所示的加性模型可得

$$\widehat{\boldsymbol{N}}_1^{\mathrm{est}}=\boldsymbol{X}^{\mathrm{te}}-\widehat{\boldsymbol{S}}_1^{\mathrm{est}}, \tag{14}$$

$$\widehat{\boldsymbol{S}}_2^{\mathrm{est}}=\boldsymbol{X}^{\mathrm{te}}-\widehat{\boldsymbol{N}}_1^{\mathrm{est}}. \tag{15}$$

然后将两路的估计信号进行加权融合：

$$\widehat{\boldsymbol{S}}^{\mathrm{est}}=(1-\alpha)\,\widehat{\boldsymbol{S}}_1^{\mathrm{est}}+\alpha\widehat{\boldsymbol{S}}_2^{\mathrm{est}}, \tag{16}$$

$$\widehat{\boldsymbol{N}}^{\mathrm{est}}=(1-\alpha)\,\widehat{\boldsymbol{N}}_1^{\mathrm{est}}+\alpha\widehat{\boldsymbol{N}}_2^{\mathrm{est}}. \tag{17}$$

这里设置权重 α 的目的在于衡量混合-语音联合稀疏表示和混合-噪声联合稀疏表示这两路对于稀疏表示的有效性。有效性高，表示该路估计的信号越准确，则相应的权重越大。文献[10]中分析指出影响这种有效性的重要因素就是信号的结构性，信号的结构性越好，相应其稀疏表示的误差就越小，相应的联合稀疏表示的权重就应该越大。考虑到说话人语音的结构性相对稳定，而不同类型的噪声在结构性上具有较大的差异，因此在设置权重 α 时仅考虑了不同噪声的结构差异性。为了衡量这种结构差异性，文献[10]提出了基于SVD 和基尼系数的信号结构特性度量方法，对于训练噪声样本的时频谱幅度矩阵 $\boldsymbol{N}^{\mathrm{tr}}$ 的每一列去均值得到 $\bar{\boldsymbol{N}}^{\mathrm{tr}}$，对其进行转置再进行SVD分解，得到奇异值并进行升序排列：$\sigma_1\leqslant\sigma_2\leqslant\cdots\leqslant\sigma_M$，然后基于洛伦兹曲线计算基尼系数来衡量奇异值的分布稀疏度，由于基尼系数能够有效地衡量信号的结构特性，因此可用基尼系数 G 作为权重系数 α。

最后，基于式(16)和式(17)估计出的语音信号和噪声，构造时频域的维纳滤波器：

$$\boldsymbol{M}=\frac{\left(\widehat{\boldsymbol{S}}^{\mathrm{est}}\right)^2}{\left(\widehat{\boldsymbol{S}}^{\mathrm{est}}\right)^2+\left(\widehat{\boldsymbol{N}}^{\mathrm{est}}\right)^2}, \tag{18}$$

其中，式(18)除法为元素级除。然后将 $\boldsymbol{M}$ 与带噪语音信号 $\boldsymbol{X}^{\mathrm{te}}$ 进行元素级相乘，对其进行滤波，得到最终估计的干净语音信号的时频谱 $\widehat{\boldsymbol{S}}$，同样结合混合信号 $\boldsymbol{X}^{\mathrm{te}}$ 的相位进行逆STFT变换，即可恢复出估计的语音信号的时域信号。

显然，该算法利用了带噪语音信号包含干净语音和噪声的关系，使得学习到的字典相比于生成性字典学习方法得到的单个信号字典更具有区分性。此外，用基尼系数融合两路稀疏表示结果，利用了联合稀疏表示的互补优势，实现对语音信号更精确的估计。文献[11]则在互补联合字典的基础上再加一级单独由干净语音和噪声训练出来的子字典，采用了两级稀疏表示结构，进一步降低信号混淆情况，提升语音增强的效果。然而值得一提的是，基尼系数仅仅取决于信号的结构性，并不能反映出在不同信噪比情况下两路稀疏表示的性能变化，且文献[10]中仅仅考虑了不同噪声的结构性，但实际上语音信

号的结构性对稀疏表示的有效性也有一定的影响，因此文献[10]中采用的权重系数并不是最优的权重系数，算法性能仍有进一步提升的空间。

2.3 基于联合特征字典学习的语音增强算法

文献[14]在生成性字典学习算法的基础上，引入RM特征来挖掘带噪语音信号中语音信号和噪声的时频谱幅度的比例信息。首先在训练阶段，联合学习信号的时频谱幅度字典和比率掩码字典：

$$\begin{aligned}&\min_{\boldsymbol{D}_s,\boldsymbol{W}_s,\boldsymbol{C}_s^{\mathrm{tr}}}\left\|\begin{bmatrix}\boldsymbol{S}^{\mathrm{tr}}\\ \lambda\boldsymbol{P}_s^{\mathrm{tr}}\end{bmatrix}-\begin{bmatrix}\boldsymbol{D}_s\\ \lambda\boldsymbol{W}_s\end{bmatrix}\boldsymbol{C}_s^{\mathrm{tr}}\right\|_{\mathrm{F}}^2\\&\quad\text{s.t.}\left\|\boldsymbol{c}_{s,g}^{\mathrm{tr}}\right\|_0\leqslant q,\forall g\,,\end{aligned}\tag{19}$$

其中，$\boldsymbol{P}_s^{\mathrm{tr}}=\boldsymbol{S}^{\mathrm{tr}}/\boldsymbol{N}^{\mathrm{tr}}$表示对于语音信号理想比率掩码，$\boldsymbol{D}_s$和$\boldsymbol{W}_s$分别表示语音信号的时频谱幅度字典和比率掩码字典，$\boldsymbol{C}_s^{\mathrm{tr}}$表示相应的稀疏系数矩阵，$\boldsymbol{c}_{s,g}^{\mathrm{tr}}$表示$\boldsymbol{C}_s^{\mathrm{tr}}$的第$g$列，$q$是对应的稀疏约束阈值，$\lambda^2$则是权衡时频谱幅度近似误差项和理想比率掩码近似误差项。

同理建立对噪声时频谱幅度字典和理想比率掩码字典的学习目标函数：

$$\begin{aligned}&\min_{\boldsymbol{D}_s,\boldsymbol{W}_s,\boldsymbol{C}_s^{\mathrm{tr}}}\left\|\begin{bmatrix}\boldsymbol{N}^{\mathrm{tr}}\\ \lambda\boldsymbol{P}_n^{\mathrm{tr}}\end{bmatrix}-\begin{bmatrix}\boldsymbol{D}_n\\ \lambda\boldsymbol{W}_n\end{bmatrix}\boldsymbol{C}_n^{\mathrm{tr}}\right\|_{\mathrm{F}}^2\\&\quad\text{s.t.}\left\|\boldsymbol{c}_{n,g}^{\mathrm{tr}}\right\|_0\leqslant q,\forall g\,,\end{aligned}\tag{20}$$

其中，噪声的联合比率掩码字典为$\boldsymbol{P}_n^{\mathrm{tr}}=\boldsymbol{1}-\boldsymbol{P}_s^{\mathrm{tr}}$，$\boldsymbol{1}$表示全1矩阵，其他变量与式(19)中的变量定义类似。

在增强测试阶段，基于训练得到的复合时频谱幅度字典$\boldsymbol{D}=[\boldsymbol{D}_s,\boldsymbol{D}_n]$和复合比率掩码字典$\boldsymbol{W}=[\boldsymbol{W}_s,\boldsymbol{W}_n]$，对带噪语音信号$\boldsymbol{X}^{\mathrm{te}}$和混合信号的比率掩码$\boldsymbol{P}^{\mathrm{te}}$进行联合稀疏投影：

$$\min_{\boldsymbol{C}^{\mathrm{te}}}\left\|\begin{bmatrix}\boldsymbol{X}^{\mathrm{te}}\\ \lambda\boldsymbol{1}\end{bmatrix}-\begin{bmatrix}\boldsymbol{D}_s & \boldsymbol{D}_n\\ \lambda\boldsymbol{W}_s & \lambda\boldsymbol{W}_n\end{bmatrix}\boldsymbol{C}^{\mathrm{te}}\right\|_{\mathrm{F}}^2,\tag{21}$$

其中，$\boldsymbol{C}^{\mathrm{te}}=\left[\left(\boldsymbol{C}_s^{\mathrm{te}}\right)^{\mathrm{T}}\left(\boldsymbol{C}_n^{\mathrm{te}}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$是稀疏系数复合矩阵，$c_g^{\mathrm{te}}$代表$\boldsymbol{C}^{\mathrm{te}}$的第$g$列，采用的是LARC算法对式(21)进行求解。

显然，基于得到的稀疏表示矩阵$\widehat{\boldsymbol{C}}^{\mathrm{te}}$可以计算出带噪语音信号$\boldsymbol{X}^{\mathrm{te}}$中对应的语音比率掩码$\widehat{\boldsymbol{P}}_s^{\mathrm{te}}$和噪声比率掩码$\widehat{\boldsymbol{P}}_n^{\mathrm{te}}$：

$$\widehat{\boldsymbol{P}}_s^{\mathrm{te}}=\boldsymbol{W}_s\widehat{\boldsymbol{C}}_s^{\mathrm{te}},\ \ \widehat{\boldsymbol{P}}_n^{\mathrm{te}}=\boldsymbol{W}_s\widehat{\boldsymbol{C}}_n^{\mathrm{te}}.\tag{22}$$

基于上述得到的比率掩码，文献[14]设计了两种掩码滤波器以实现更好的语音增强效果。第一种为软掩码滤波器，由理想二值掩码滤波器和维纳滤波器加权平均得到：

$$\begin{aligned}&SM_1(k,t)=\\&\beta J(k,t)+(1-\beta)\frac{\widehat{P}_s^{\mathrm{te}}(k,t)}{\widehat{P}_s^{\mathrm{te}}(k,t)+\widehat{P}_i^{\mathrm{te}}(k,t)},\end{aligned}\tag{23}$$

其中，$J(k,t)$为理想二值掩码滤波器，计算公式如下：

$$J(k,t)=\begin{cases}1, & \widehat{P}_s^{\mathrm{te}}(k,t)\geqslant\widehat{P}_n^{\mathrm{te}}(k,t),\\ 0, & \text{其他.}\end{cases}\tag{24}$$

而式(23)中的第二项即为维纳形式的滤波器，β为衡量这两个滤波器的权重值。显然，当$\beta=0$时，由式(24)得到的即为维纳形式的滤波器，反之当$\beta=1$时，得到的就是理想二值掩码滤波器。

考虑到在某个时频点往往会出现语音成分或噪声成分占主导作用的情况，文献[14]中提出了第二种滤波器：

$$SM_2(k,t)=\begin{cases}1, & \dfrac{\widehat{P}_s^{\mathrm{te}}(k,t)}{\widehat{P}_n^{\mathrm{te}}(k,t)}\geqslant\alpha,\\ 0, & \dfrac{\widehat{P}_s^{\mathrm{te}}(k,t)}{\widehat{P}_n^{\mathrm{te}}(k,t)}\leqslant\dfrac{1}{\alpha},\\ SM_1(k,t), & \text{otherwise.}\end{cases}\tag{25}$$

显然，α用于衡量语音成分是否占主导作用，当语音比率掩码值与噪声比率掩码值之比超过α时，则表明语音成分占主导作用，相应的掩码滤波值设为1，反之则认为是噪声，相应的滤波掩码值设为0。然而，当语音比率掩码值和噪声比率掩码值所占成分相差不大时，则保留为软掩码滤波值。

最后，将设计的掩码滤波器与混合带噪语音信号$\boldsymbol{X}^{\mathrm{te}}$相乘即可得到估计的语音信号的时频幅度谱$\widehat{\boldsymbol{S}}$，然后结合带噪语音信号的相位信息，经过逆STFT变换即可得到增强后语音信号的时域形式。

相比于上述介绍的生成性字典学习算法和互补联合字典学习算法，该算法不仅利用了信号时频

谱幅度的信息，也挖掘了带噪语音信号中语音信号和噪声的时频谱幅度的比例信息，基于多任务联合处理的思想联合学习了信号的时频谱幅度字典和比率掩码字典，提升了语音增强的性能，但同时也要求对λ、β和α等重要参数进行合理的设置，这就需要大量的实验调整和经验，降低了该算法的适应性和灵活性。

2.4 基于非负矩阵分解的语音增强算法

NMF算法的基本思想是将一个非负矩阵$\boldsymbol{Y}\in\boldsymbol{R}^{K\times T}$分解成非负字典$\boldsymbol{W}$和激活系数矩阵$\boldsymbol{H}$的乘积，常用的目标代价函数形式有Itakura–Saito距离、广义Kullback-Leibler散度和欧式距离[16]。当采用广义Kullback-Leibler散度时，基于乘法更新规则[24]可以得到对$\boldsymbol{W}$和$\boldsymbol{H}$的迭代更新公式：

$$\boldsymbol{H}\leftarrow\boldsymbol{H}\otimes\frac{\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{Y}/\boldsymbol{W}\boldsymbol{H}\right)}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{1}_{K\times T}},\tag{26}$$

$$\boldsymbol{W}\leftarrow\boldsymbol{W}\otimes\frac{\left(\boldsymbol{Y}/\boldsymbol{W}\boldsymbol{H}\right)\boldsymbol{H}^{\mathrm{T}}}{\boldsymbol{1}_{K\times T}\boldsymbol{H}^{\mathrm{T}}},\tag{27}$$

其中，符号$\otimes$表示元素级相乘，$\cdot/\cdot$表示元素级除，$\boldsymbol{1}_{K\times T}\in\boldsymbol{R}^{K\times T}$表示全1矩阵。当目标代价函数值趋于收敛或小于设定的阈值时，对$\boldsymbol{W}$和$\boldsymbol{H}$的更新停止。

在语音增强领域中，经典的有监督类NMF算法通常作用于信号的时频谱幅度，首先在训练阶段利用训练样本学习到语音字典$\boldsymbol{W}_s$和噪声字典$\boldsymbol{W}_n$，然后在测试阶段计算带噪语音信号在字典上的表示系数$\boldsymbol{H}_s$和$\boldsymbol{H}_n$，估计出语音信号和噪声，然后构造维纳滤波器与带噪语音信号进行元素级相乘，恢复出干净语音信号。

针对训练数据和测试数据出现不匹配的情况，文献[21]在增强阶段首先采用基于统计模型的传统增强方法[22]对带噪语音信号进行预增强，利用预增强信号和原始带噪语音信号学习新的语音字典和噪声字典：

$$\min_{\tilde{\boldsymbol{V}}_t,\widetilde{\boldsymbol{W}}_t,\tilde{\boldsymbol{H}}_t}D_{KL}\left(\tilde{\boldsymbol{V}}_t,\widetilde{\boldsymbol{W}}_t\tilde{\boldsymbol{H}}_t\right)+\delta\left\|\boldsymbol{W}_t-\widetilde{\boldsymbol{W}}_t\right\|^2,\tag{28}$$

其中，δ为设置的权重，$D_{KL}(\cdot)$表示采用广义Kullback-Leibler散度形式。$\tilde{\boldsymbol{V}}_t=\left[\boldsymbol{x}_t^{\mathrm{te}}\ \tilde{\boldsymbol{x}}_t^{\mathrm{te}}\right]$，$\boldsymbol{x}_t^{\mathrm{te}}\in\boldsymbol{R}^{K\times1}$代表原始带噪语音信号在时频域上的第$t$帧，$\tilde{\boldsymbol{x}}_t^{\mathrm{te}}\in\boldsymbol{R}^{K\times1}$代表预增强信号在时频域上的第$t$帧，$\widetilde{\boldsymbol{W}}_t=\left[\widetilde{\boldsymbol{W}}_t^s\ \widetilde{\boldsymbol{W}}_t^n\right]\in\boldsymbol{R}^{K\times(r_s+r_n)}$表示需要学习的语音字典和噪声字典，可基于乘法更新规则得到其迭代更新公式，r_s表示语音字典的原子数，r_n表示噪声字典的原子数。$\boldsymbol{W}_t=\left[\boldsymbol{W}_t^s\ \boldsymbol{W}_t^n\right]\in\boldsymbol{R}^{K\times(r_s+r_n)}$表示在第$t$帧基于训练字典更新后得到的语音字典和噪声字典，由训练学习得到的字典$\boldsymbol{W}_s$和$\boldsymbol{W}_n$作为初始化矩阵，具体更新公式如下：

$$\begin{aligned}\boldsymbol{W}_{t+1}^s&=\boldsymbol{\lambda}_t^s\otimes\widetilde{\boldsymbol{W}}_t^s+\left(\boldsymbol{1}_{K\times r_s}-\boldsymbol{\lambda}_t^s\right)\otimes\boldsymbol{W}_t^s,\\\boldsymbol{\lambda}_t^s&=\alpha_s(t)\,\boldsymbol{p}(t)\,\boldsymbol{1}_{r_s},\end{aligned}\tag{29}$$

$$\begin{aligned}\boldsymbol{W}_{t+1}^n&=\boldsymbol{\lambda}_t^n\otimes\widetilde{\boldsymbol{W}}_t^n+\left(\boldsymbol{1}_{K\times r_n}-\boldsymbol{\lambda}_t^n\right)\otimes\boldsymbol{W}_t^n,\\\boldsymbol{\lambda}_t^n&=\alpha_n(t)\,\boldsymbol{p}(t)\,\boldsymbol{1}_{r_n},\end{aligned}\tag{30}$$

其中，$\boldsymbol{p}(t)\in\boldsymbol{R}^{K\times1}$代表第$t$帧的语音存在概率，$\boldsymbol{1}_{K\times r_s}\in\boldsymbol{R}^{K\times r_s}$、$\boldsymbol{1}_{K\times r_n}\in\boldsymbol{R}^{K\times r_n}$、$\boldsymbol{1}_{r_s}\in\boldsymbol{R}^{1\times r_s}$和$\boldsymbol{1}_{r_n}\in\boldsymbol{R}^{1\times r_n}$表示全1矩阵，$\alpha_s(t)$和$\alpha_n(t)$表示最大更新比率，可以通过计算重构误差获得：

$$\alpha_s(t)=\max\left[\mathrm{sigm}\left(\tilde{e}(t)\right)\alpha_s^{\max},0.01\right],\tag{31}$$

$$\alpha_n(t)=\max\left[\mathrm{sigm}\left(\tilde{e}(t)\right)\alpha_n^{\max},0.01\right],\tag{32}$$

其中，$\mathrm{sigm}(\cdot)$代表sigmoid函数，$\alpha_s^{\max}$和$\alpha_n^{\max}$为设置的更新比率的最大上限，$\tilde{e}(t)$是由归一化的重构误差$e(t)$平滑得到，即

$$\tilde{e}(t)=\tau_e\tilde{e}(t-1)+(1-\tau_e)\,\tilde{e}(t),\tag{33}$$

显然，$0\leqslant\tau_e\leqslant1$为平滑因子，$e(t)$可由式(34)计算得到：

$$e(t)=\frac{\sum_{k=1}^{K}\left(\boldsymbol{x}_t^{\mathrm{te}}(k)-(\boldsymbol{W}\boldsymbol{H})_{k,t}\right)^2}{\sum_{k=1}^{K}\left(\boldsymbol{x}_t^{\mathrm{te}}(k)\right)^2},\tag{34}$$

其中，$\boldsymbol{W}=[\boldsymbol{W}_s\ \boldsymbol{W}_n]$表示训练阶段得到的语音字典和噪声字典，$\boldsymbol{H}$表示训练阶段对应语音信号和噪声的稀疏表示系数，$(\boldsymbol{W}\boldsymbol{H})_{k,t}$表示$\boldsymbol{W}$和$\boldsymbol{H}$乘积的第$(k,t)$个元素。

不难看出，该算法实现了对语音字典和噪声字典的在线更新，在非平稳环境下能够捕捉到更多的信号特征，且能够在训练数据与测试数据不匹配的情况下实现较好的语音增强，具有较好的灵活性和实用价值。但这也同时要求基于统计模型的传统增强方法在进行预处理时不能产生较多的失真成分，如果预增强后的$\tilde{\boldsymbol{x}}_t^{\mathrm{te}}$含有较多的失真，对语音字典的在线更新就可能会造成一定负面影响，甚至会降低语音增强的效果。

3 结论

本文主要介绍了基于字典学习和稀疏表示的单通道语音增强算法，首先介绍了基于生成性字典学习的语音增强算法，该算法利用信号的时频谱幅度学习语音字典和噪声字典，相比于传统算法在抑制非平稳噪声方面有一定的优越性。随后阐述的基于互补联合字典学习和稀疏表示的增强算法和基于联合特征字典学习的增强算法均是由生成性字典学习发展而来，通过挖掘带噪语音信号和语音信号、噪声之间的映射关系和引入比率掩码字典来进一步提升语音增强的性能。最后介绍的有监督类 NMF 算法则考虑了在训练数据和测试数据不匹配情况下进行语音增强，通过将 NMF 算法与基于统计模型的传统增强方法相结合来进一步挖掘测试数据的特性，对已有的语音字典和噪声字典进行更新，从而恢复出干净语音信号，更具有灵活性和应用前景。值得一提的是，文中介绍的前三种方法均采用 LARC 算法进行稀疏编码，由于 LARC 算法是基于相关性来判断迭代更新过程是否应该终止，因此当没有噪声训练样本用于初始化时，随机初始化得到的噪声字典并不一定与带噪语音信号中的噪声成分有较强的相关性，LARC 算法很可能放弃对噪声成分的稀疏表示，不能对噪声字典进行有效的更新，这也就限制了算法在无监督或者半监督情况下的应用。而基于 NMF 的算法由于对字典初始化的要求并不苛刻 (可由随机数生成)，因此具有较高的灵活性和较广的适用范围。然而，由于 NMF 类算法对字典和激活系数矩阵均有非负约束，在降噪方面则稍逊色于基于生成性字典学习一类的算法。

基于字典学习和稀疏表示的单通道语音增强算法在近十多年的时间里取得了飞速的发展并已获得了丰硕的成果，但依然存在一些问题值得进一步深入研究：(1) 建立不依赖于说话人特性的语音字典和稀疏表示方法。上述提到的有监督类学习算法均是针对特定说话人的语音学习相应的字典，在实际应用中会受到一定限制，研究如何强化各说话人语音数据的共性部分建立具有普适性的语音字典将会有更大的应用价值。(2) 寻找对语音信号和噪声更具有区分性的特征域。上述算法都是在信号的时频域进行操作的，虽然该特征域的变换是线性可逆的，但没有利用到人声在生理特性或听觉感知上对于语音和噪声的差异性，如何利用这些差异提取更有区分性的特征并建立该特征域与线性可逆特征域之间的映射关系是一个值得思考的问题。(3) 目前基于字典学习和稀疏表示的增强方法大多是对信号的时频域幅度谱进行处理，然后用带噪语音信号的相位作为估计语音信号的相位，如何很好地利用相位信息，实现对相位信息和幅度谱进行联合优化构造相应的字典是一个亟待突破的方向。

参 考 文 献

[1] Kamath S, Loizou P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise[C]. ICASSP, Citeseer, 2002.

[2] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1985, 33(2): 443–445.

[3] Hu Y, Loizou P C. A generalized subspace approach for enhancing speech corrupted by colored noise[J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(4): 334–341.

[4] Sun J, Zhang J, Small M. Extension of the local subspace method to enhancement of speech with colored noise[J]. Signal Processing, 2008, 88(7): 1881–1888.

[5] Sigg C D, Dikk T, Buhmann J M. Speech enhancement with sparse coding in learned dictionaries[C]. ICASSP, Dallas, TX, 2010.

[6] Sigg C D, Dikk T, Buhmann J M. Speech enhancement using generative dictionary learning[J]. IEEE Transactions on Audio, Speech and Language Processing, 2012, 20(6): 1698–1712.

[7] Efron B, Hastie T, Johnstone I, et al. Least angle regression[J]. The Annals of Statistics, 2004, 32: 407–499.

[8] Bao G Z, Xu Y F, Ye Z F. Learning a discriminative dictionary for single-channel speech separation[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2014, 22(7): 1130–1138.

[9] Nejati M, Samavi S, Soroushmehr S M R, et al. Coherence regularized dictionary learning[C]. ICASSP, Shanghai, 2016.

[10] Luo Y, Bao G Z, Xu Y F, et al. Supervised monaural speech enhancement using complementary joint sparse representations[J]. IEEE Signal Processing Letters, 2016, 23(2): 237–241.

[11] Fu J F, Zhang L, Ye Z F. Supervised monaural speech enhancement using two-level complementary joint sparse representations[J]. Applied Acoustics, 2018, 132: 1–7.

[12] Xu Y F, Bao G Z, Xu X, et al. Single-channel speech separation using sequential discriminative dictionary learning[J]. Signal Processing, 2015, 106: 34–140.

[13] He Y J, Sun G L, Han J Q. Spectrum enhancement with sparse coding for robust speech recognition[J]. Digital Signal Processing, 2015, 43: 59–70.

[14] Zhang L, Bao G Z, Zhang J, et al. Supervised single-channel speech enhancement using ratio mask with joint dictionary learning[J]. Speech Communication, 2016, 82: 38–52.

[15] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[J]. Advances in Neural Information Processing Systems, 2001: 556–562.

[16] Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the β-divergence[J]. Neural Computation, 2011, 23(9): 2421–2456.

[17] Grais E M, Emad H E. Hidden Markov models as priors for regularized nonnegative matrix factorization in single-channel source separation[C]. INTERSPEECH, 2012.

[18] Wilson K W, Raj B, Smaragdis P. Regularized non-negative matrix factorization with temporal dependencies for speech denoising[C]. INTERSPEECH, 2008.

[19] Chung H, Plourde E, Champagne B. Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement[J]. Speech Communication, 2017, 87: 18–30.

[20] Jeon K M, Kim H K. Local sparsity based online dictionary learning for environment-adaptive speech enhancement with nonnegative matrix factorization[C]. INTERSPEECH, 2016.

[21] Kwon K, Shin J W, Kim N S. NMF-based speech enhancement using bases update[J]. IEEE Signal Processing Letters, 2015, 22(4): 450–454.

[22] Rangachari S, Loizou P C. A noise-estimation algorithm for highly non-stationary environments[J]. Speech Communication, 2006, 48(2): 220–231.

[23] Aharon M, Elad M, Bruckstein A, et al. K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311–4322.

[24] Lee D D, Seung H S. Learning the parts of objects by nonnegative matrix factorization[J]. Nature, 1999, 401: 788–791.