第 40 卷第 4 期　　声　学　学　报　　Vol. 40, No. 4
2015 年 7 月　　ACTA　ACUSTICA　　Jul., 2015

# 稀疏低秩噪声模型下无监督实时单通道语音增强算法*

李轶南[1]　张雄伟[1]　贾　冲[1]　陈　亮[2]　曾　理[1]

(1　解放军理工大学指挥信息系统学院　南京　210017)

(2　解放军理工大学通信工程学院　南京　210017)

2013 年 12 月 26 日收到

2014 年 3 月 7 日定稿

**摘要**　针对现有基于字典学习的增强算法需要先验信息、不易实时处理的问题，提出一种便于实时处理的无监督的单通道语音增强算法。首先，该算法将无监督条件下背景噪声的建模问题转化为带噪语音幅度谱的稀疏低秩噪声分解；然后，采用增量非负子空间方法对背景噪声进行在线字典学习，获得能够体现背景噪声时变特性的自适应噪声字典；最后，利用所得的噪声字典，采用易于实时处理的逐帧迭代方式，对带噪语音进行处理。实验结果表明：相较于多带谱减法和基于低秩稀疏矩阵分解的增强算法，所提算法在噪声抑制方面的性能尤为显著，在多项性能评价指标上，均表现出更好的结果。

PACS 数：43.72

## Unsupervised real-time single channel speech enhancement with sparse low-rank and noise model

LI Yinan[1]　ZHANG Xiongwei[1]　JIA Chong[2]　CHEN Liang[1]　ZENG Li[1]

(1　*College of Command Information System, PLAUST*　Nanjing　210007)

(2　*College of Communications Engineering, PLAUST*　Nanjing　210007)

Received Dec. 26, 2013

Revised Mar. 7, 2014

**Abstract**　An unsupervised speech enhancement algorithm suitable for real-time processing in one channel record is proposed, aiming at resolving the prior-information-reliance and real-time processing difficulty in existing enhancement algorithms based on dictionary learning. With the magnitude of noisy speech, it recasts unsupervised background noise modeling problem into sparse, low-rank and noise decomposition. Subsequently, an adaptive noise dictionary which reflects the dynamic noise background is learned in an online fashion by employing incremental nonnegative subspace learning. Finally, frame-by-frame enhancement is conducted with the learnt dictionary, which makes the real-time processing much more convenient. Extensive experiments demonstrate that the presented algorithm outperforms state-of-the-art method such as multi-band spectral subtraction and method based on low-rank and sparse matrix decomposition, especially in terms of noise reduction.

## 引言

语音增强是指为消除噪声污染，提升语音信号质量所做的相应处理，其主要任务是从带噪语音中提取出尽可能纯净的语音信号或尽可能准确的语音参数，是语音处理领域的一个重要分支。语音增强根据语音信号在实际采集过程中所使用传声器数量的不同，又可分为单通道增强算法和多通道增强算法。相较于后者，单通道方式具有更易部署、成本更低、更易实现等优点，因此，长期以来，单通道语音增强算法一直是学者们研究的热点。

* 国家自然科学基金 (61072042, 61402519, 61471394) 和江苏省自然科学基金 (BK2012510, BK20140071, BK20140074) 资助

单通道语音增强算法根据是否具有语音和噪声的先验信息又可以进一步划分为有监督算法和无监督算法两大类。典型的无监督算法包括谱减法 [1]、Wiener 及 Kalman 滤波法 [2]、基于多元 Laplace 语音模型的短时谱估计算法 [3] 以及基于语音信号周期模型的增强算法 [4] 等。这些算法无需指定具体噪声类型或提供特定说话人语音特征等先验信息，即能够在带噪语音中估计出纯净语音信号。这类算法的主要技术难点在于对噪声功率谱密度的估计 [5]，尤其是当背景噪声具有非平稳特性时，对噪声功率谱的估计将会变得非常困难。经典的有监督算法分别对语音和噪声信号进行建模，并用语音和噪声样本分别对所建模型进行训练，估计出模型的具体参数，然后利用所得参数在带噪语音中估计出纯净语音。由于更合理地利用了语音和噪声的先验信息，这类算法往往能够获得比无监督算法更好的增强效果，因此受到了学者们的广泛关注，特别是近年来兴起的基于字典学习的语音增强算法更是成为语音处理领域的研究热点。

Kevin W. Wilson 等 [6] 基于 NMF (Nonnegative Matrix Factorization) 算法，对语音和噪声分别进行字典学习，获得二者的非负联合字典，通过将带噪语音在联合字典上进行投影，分离出纯净语音。Nasser Mohammadiha 等 [7] 深入研究了 Babble 噪声的特点，并将隐 Markov 模型引入到稀疏 NMF 算法中，提出一种能够从 Babble 噪声中分离出纯净语音的增强算法。这些算法虽然能够获得较好的增强效果，但由于对先验信息的大量需求及噪声模型自身的局限性，使其难以推广应用。为了获得具有更优增强性能和更具普适性的增强算法，学者们做出了很多卓有成效的努力。Mikkel Schmidt 等 [8] 基于 NMF 算法提出一种无监督批处理语音增强算法，该算法在语音间歇期获得噪声字典，并以 Wiener 滤波的方式实现了对带噪语音信号的增强。然而，这种将全部带噪语音采集下来进行集中处理的批处理方式大大降低了算法的实用性。黄建军等 [9] 将能够更好反映语音时频结构的卷积 NMF 模型运用到语音增强中，提出一种新的增强算法，该算法只需预先训练出噪声字典，即能够消除带噪语音中的特定噪声，但是，由于在进行增强前，仍需训练离线字典，使得该算法具有一定的局限性。Christian D. Sigg 等 [10] 在离线的条件下使用 K-SVD (K-Singular Value Decomposition) 算法训练得到语音字典，并通过 VAD (Voice Activity Detection) 在语音间歇期获取噪声字典，实现了语音增强，该算法在非平稳噪声环境下，也能保持良好的性能。然而，一方面，当训练样本与实际样本偏差较大时，离线字典学习方式的弊端将会非常明显；另一方面，算法性能严重依赖于 VAD 的性能，不准确的 VAD 会导致算法的性能急剧下降。

为了克服传统字典学习算法依赖语音或噪声的先验信息、需要 VAD 以及无法实现实时处理等问题，本文基于稀疏低秩噪声分解和在线字典学习，提出一种无监督的实时语音增强算法 (Unsupervised Real Time Enhancement, URTE) 。该算法首先对语音信号的幅度谱进行稀疏低秩噪声分解，获取所处环境的背景噪声。然后利用增量非负子空间方法对背景噪声进行在线学习，获得相应的自适应噪声字典。最后，利用所得噪声字典和乘性迭代公式，采用便于实时处理的逐帧处理模式，于带噪语音中分离出纯净语音。实验结果表明，本文算法消除噪声的性能显著优于多带谱减法和基于低秩稀疏矩阵分解的增强算法。

## 1 语音的稀疏低秩模型及存在的不足

在某种具体的噪声环境下，噪声将会在频域呈现出一些独特的特征，这些特征将会随着时间的推移反复出现，使噪声的幅度谱呈现出一定的低秩结构。常见的基于字典学习的增强算法正是通过对噪声幅度谱进行学习，来获取噪声的低秩结构。

假设被噪声污染的语音信号的幅度谱为 $\boldsymbol{Y}$。语音信号由于具有短时平稳性，其幅度谱会表现地十分稀疏 [11]，记 $\boldsymbol{Y}$ 中纯净语音为 $\boldsymbol{S}$；另一方面，由上文分析可知，$\boldsymbol{Y}$ 中的噪声具有潜在的低秩结构，如果将噪声的幅度谱记为 $\boldsymbol{L}$，那么，语噪分离问题即可以转化为在已知 $\boldsymbol{Y}$ 的情况下，同时求解低秩部分 $\boldsymbol{L}$ 和稀疏部分 $\boldsymbol{S}$ 之和，如式 (1) 所示。

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S}. \tag{1}$$

由于 $\boldsymbol{L}$ 具有低秩特性，$\boldsymbol{S}$ 具有稀疏特性，可以将式 (1) 改写为：

$$\min \operatorname{rank}(\boldsymbol{L}) + \lambda \|\boldsymbol{S}\|_0 \quad s.t. \ \boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S} \tag{2}$$

式中，$\|\cdot\|_0$ 为矩阵的零范数，即矩阵的势，所表示的是矩阵中非零元素的个数。$\lambda$ 是一个用以权衡矩阵 $\boldsymbol{L}$ 的秩和矩阵 $\boldsymbol{S}$ 的势的系数。此问题是 NP-hard 问题，无法直接求解。为了解决式 (2) 所描述的问题，Candès 等 [12] 于 2011 年提出鲁棒主成分分析 (Robust Principal Component Analysis, RPCA)，RPCA 通过使用矩阵的核范数 (矩阵奇异值之和) 和矩阵的 $\ell_1$ 范数分别替代矩阵 $L$ 的秩和矩阵 $S$ 的 $\ell_0$ 范数，将式 (2) 转化为形如式 (3) 的凸优化问题。

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad s.t. \ \ \boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S} \tag{3}$$

式中，$\|\cdot\|_*$ 和 $\|\cdot\|_1$ 分别表示核范数和 $\ell_1$ 范数，其中，可以将核范数看作是奇异值的 $\ell_1$ 范数。

通过增广拉格朗日乘子法 (Augmented Lagrange Multiplier, ALM)[13] 可以很方便地求取出式 (3) 的解，即实现语音和噪声的分离。但是，这种分离方法仍然存在三点不足。第一，幅度谱的稀疏低秩分解模型并不一定适用于所有的噪声环境，例如，高斯噪声既不稀疏，也难以使用低秩模型进行描述；第二，直接对幅度谱进行稀疏低秩分解本质上是一种批处理的增强方式，即需要将一段带噪语音全部采集下来进行集中处理，这不能够满足语音增强的实时性要求，因此具有较大的局限性；第三，直接对稀疏低秩分解所得的稀疏部分进行重构的增强方式会造成资源的浪费。因为该算法只对稀疏部分加以利用，低秩部分则被直接遗弃，而噪声主要集中于低秩部分，通过研究低秩部分的特点可以获取所在背景噪声环境的特征，充分利用这些特征能够帮助进一步提升算法的增强效果 (相应的实验与分析参见第 3 节)。

## 2　稀疏低秩噪声模型下的无监督实时单通道语音增强算法

本文增强算法针对第 1 节中指出的三点不足而提出，首先，将稀疏低秩模型进一步完善为稀疏低秩噪声模型，使分解所得的各部分都均具有可解释性；其次，以在线字典学习的方式对噪声特征进行学习，获取自适应噪声字典；最后，利用所得噪声字典，以帧为单位对语音信号进行处理，满足了增强算法的实时性要求。所提算法的总体框图如图 1 所示。

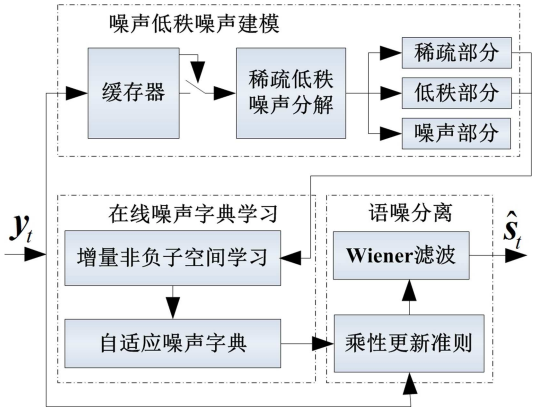


图 1　本文算法的总体框图

图中，所有处理均在频域完成，$\boldsymbol{y}_t$ 为 $t$ 时刻带噪语音帧的幅度谱；$\hat{\boldsymbol{s}}_t$ 为增强后的语音帧。由框图可知，本文算法主要包括三部分，这三部分将会分别在 2.2 — 2.4 节中进行详细介绍。由于增强算法所处理的是加性噪声模型下带噪语音信号的幅度谱，下面首先介绍加性噪声模型及带噪语音的幅度谱计算。

### 2.1　加性噪声模型及幅度谱计算

加性噪声模型是语音增强中最常见的噪声模型，假设 $\boldsymbol{s}(t)$ 和 $\boldsymbol{n}(t)$ 分别表示纯净语音信号和噪声信号，加性噪声模型认为带噪语音 $\boldsymbol{y}(t)$ 由 $\boldsymbol{s}(t)$ 与 $\boldsymbol{n}(t)$ 直接叠加得到，即 $\boldsymbol{y}(t) = \boldsymbol{s}(t) + \boldsymbol{n}(t)$。对带噪语音分帧加窗并变换到频域，即可得到 $\boldsymbol{y}(t)$ 的短时傅里叶变换 (Short-Time Fourier Transformation, STFT)，即：

$$\dot{\boldsymbol{Y}} = \mathrm{DFT}\left[h(0) \ \cdots \ h(L-1)\right] \begin{bmatrix} y(t_1) & y(t_2) & \cdots & y(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ y(t_1+L-1) & y(t_2+L-1) & \cdots & y(t_N+L-1) \end{bmatrix}, \tag{4}$$

式中 $L$ 为帧长，$h(n)$ 为归一化窗，$R = t_i - t_{i-1}$ 表示语音帧的帧移。由于人耳对相位不敏感，因此在增强时一般不考虑相位，对 $\dot{\boldsymbol{Y}}$ 取模值即可获得相应的幅度谱 $\boldsymbol{Y} = |\dot{\boldsymbol{Y}}|$。

### 2.2　带噪语音的稀疏低秩噪声模型

由第 1 节分析可知，RPCA 方法分解所得的稀疏和低秩结构可能并不适用于所有的噪声环境，本文将带噪语音模型进一步完善为稀疏低秩和噪声模型，通过增加噪声部分使得分解所得的各部分均具有物理上的可解释性。图 2 阐述了加性噪声模型下，带噪语音幅度谱的稀疏低秩噪声分解。从图中可以看出，分解所得的低秩部分主要反映实际噪声中结构性明显的成分，噪声部分则反映实际噪声中结构性不明显的成分，稀疏部分则主要由纯净语音成分构成。因此，可以认为，噪声的主要特征可以使用低秩部分和噪声部分之和来进行表示。

为了描述带噪语音幅度谱的稀疏低秩噪声模型，可以将式 (1) 改写为式 (5)：

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S} + \boldsymbol{N}, \ \ \mathrm{rank}(\boldsymbol{L}) \leqslant r, \ \ \mathrm{card}(\boldsymbol{S}) \leqslant k. \tag{5}$$

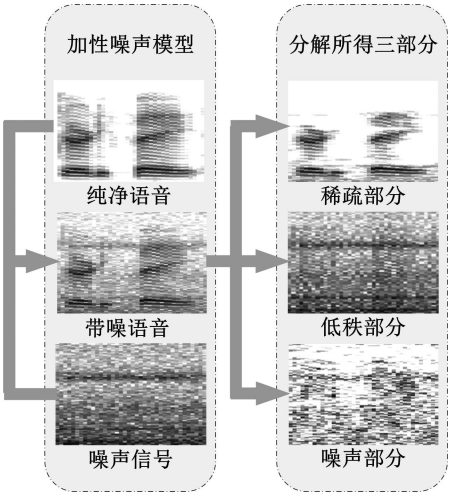


图 2 加性噪声模型下带噪语音的稀疏低秩噪声分解

文献 14 和文献 15 提出的 GoDec(Go Decomposition) 算法将式 (5) 的求解转化为两个最优化子问题, 其本质是在残差最小化的条件下, 对低秩和稀疏矩阵分别进行估计, 如式 (6) 所示:

$$\begin{cases} \boldsymbol{L}_t = \arg\min\limits_{\mathrm{rank}(\boldsymbol{L})\leqslant r} \|\boldsymbol{Y}-\boldsymbol{L}_{t-1}-\boldsymbol{S}_{t-1}\|_F^2, \\ \boldsymbol{S}_t = \arg\min\limits_{\mathrm{card}(\boldsymbol{S})\leqslant k} \|\boldsymbol{Y}-\boldsymbol{L}_{t}-\boldsymbol{S}_{t-1}\|_F^2. \end{cases} \tag{6}$$

式 (6) 在迭代的过程中, SVD 运算消耗了大量的计算量, GoDec 算法使用双边随机投影 (Bilateral Random Projection, BRP) 代替 SVD 运算, 从而大大提高了运算效率, 显著降低了算法的计算复杂度。在迭代出 $\boldsymbol{L}$ 和 $\boldsymbol{S}$ 之后, 利用已知的 $\boldsymbol{Y}$ 即可计算得到 $\boldsymbol{N}$ 。

需要指出的是, 稀疏低秩噪声分解的对象是图 1 缓存器中存储的若干帧语音信号的幅度谱, 因此目前还不能实现对带噪语音的实时增强。由上文分析, 低秩和噪声部分之和能够反映出实际噪声的主要信息, 通过在线学习的方式可以获取自适应噪声字典, 用于实时处理。

### 2.3 噪声字典的在线更新

假设 $\boldsymbol{M}$ 为稀疏低秩噪声分解以后所得的低秩部分和噪声部分之和。矩阵 $\boldsymbol{M}$ 可以表示为非负字典矩阵 $\boldsymbol{D}$ 与相应的非负增益矩阵 $\boldsymbol{C}$ 之积:

$$\boldsymbol{M} = \boldsymbol{L}+\boldsymbol{N} = \boldsymbol{D}\boldsymbol{C}. \tag{7}$$

由于缓存器中的存储的语音帧不断地发生改变, 因此 $\boldsymbol{M}$ 也是时变的。当 $\boldsymbol{M}$ 中有新的样本到来时, 传统的 NMF 算法为了同时适应新、旧样本需要将二者相结合, 重新计算相应的字典和增益系数, 使得计算开销非常大, 因此不适用于在线更新。增量非负子空间 [16] 学习方法在原来非负字典子空间和增益系数的基础上, 通过添加新样本的所产生的影响而实现在线学习。其过程可以使用式 (8) 来描述:

$$\boldsymbol{m}_t = \boldsymbol{D}\boldsymbol{c}_t, \tag{8}$$

式中, $t$ 时刻到来的新样本 $\boldsymbol{m}_t$ 在字典 $\boldsymbol{D}$ 上的时变增益为 $\boldsymbol{c}_t$, 文献 16 以迭代的方式对 $\boldsymbol{c}_t$ 和 $\boldsymbol{D}$ 进行更新, 相应的迭代公式为:

$$\boldsymbol{c}_t \leftarrow \boldsymbol{c}_t \otimes (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{m}_t \oslash \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{c}_t), \tag{9}$$

$$\boldsymbol{D}\leftarrow\boldsymbol{D}\otimes(\beta\boldsymbol{M}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{m}_t+\alpha\boldsymbol{m}_t\boldsymbol{c}_t^{\mathrm{T}})\oslash(\boldsymbol{D}(\beta\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}+\alpha\boldsymbol{c}_t\boldsymbol{c}_t^{\mathrm{T}})), \tag{10}$$

式中, $\otimes$ 表示的是矩阵中对应元素的乘, $\oslash$ 表示的则是矩阵中对应元素的除, $\alpha$ 和 $\beta$ 分别为原来样本和新到样本的权重系数。

### 2.4 语噪分离及后处理

在获得自适应噪声字典以后, 即可进行语噪分离。假设带噪语音幅度谱 $\boldsymbol{Y}$ 中的一帧为 $\boldsymbol{y}_t$, 语音和噪声字典分别用 $\boldsymbol{D}_s$ 和 $\boldsymbol{D}_n$ 进行表示, $\boldsymbol{y}_t$ 在 $\boldsymbol{D}_s$ 和 $\boldsymbol{D}_n$ 上的增益系数分别为 $\boldsymbol{c}_s$ 和 $\boldsymbol{c}_n$, 则有关系:

$$\boldsymbol{y}_t = \boldsymbol{D}\boldsymbol{c} = [\boldsymbol{D}_s\boldsymbol{D}_n]\begin{bmatrix}\boldsymbol{c}_s\\ \boldsymbol{c}_n\end{bmatrix}. \tag{11}$$

由于 $\boldsymbol{y}_t$ 和 $\boldsymbol{D}_n$ 已知, 可以通过迭代的方式估计出剩下的三个矩阵或向量 $\boldsymbol{D}_s$, $\boldsymbol{c}_s$ 和 $\boldsymbol{c}_n$ 。通过分别计算 $\boldsymbol{D}_s\boldsymbol{c}_s$ 和 $\boldsymbol{D}_n\boldsymbol{c}_n$ 能够分别估计出语音和噪声。此后, 通过式 (12) 所描述的 Wiener 滤波方式, 能够进一步提升增强语音信号的自然度, 获得更好的试听体验 [17]。整个语噪分离算法的原理如图 3 所示。

$$\widehat{\boldsymbol{y}}_t = \frac{\left(\boldsymbol{D}_s\boldsymbol{c}_s^{\mathrm{T}}\right)^2}{\left(\boldsymbol{D}_s\boldsymbol{c}_s^{\mathrm{T}}\right)^2+\left(\boldsymbol{D}_n\boldsymbol{c}_n^{\mathrm{T}}\right)^2}\otimes\boldsymbol{y}_t. \tag{12}$$

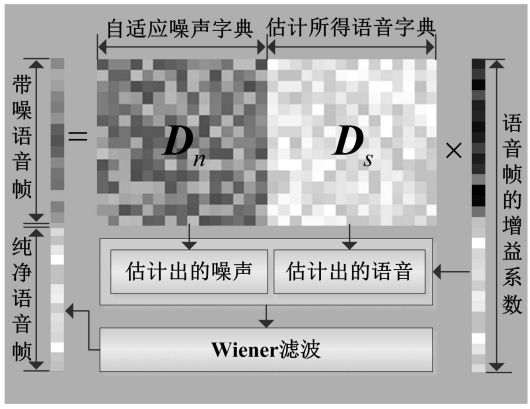


图 3 实时语噪分离原理图

文献 18 详细推导出了剩余三个矩阵的迭代更新公式, 本文将其改写为实时的帧处理形式:

$$\boldsymbol{c}_s \leftarrow \boldsymbol{c}_s \otimes \frac{\boldsymbol{D}_s^{\mathrm{T}}\boldsymbol{y}_t}{(\boldsymbol{D}_s^{\mathrm{T}}\boldsymbol{D})\boldsymbol{c}+\lambda_s}, \tag{13}$$

$$\boldsymbol{c}_n \leftarrow \boldsymbol{c}_n \otimes \frac{\boldsymbol{D}_n^{\mathrm{T}}\boldsymbol{y}_t}{(\boldsymbol{D}_n^{\mathrm{T}}\boldsymbol{D})\boldsymbol{c}+\lambda_n}, \tag{14}$$

$$\boldsymbol{D}_s \leftarrow \boldsymbol{D}_s \otimes \frac{\boldsymbol{y}_t\boldsymbol{c}_s^{\mathrm{T}}+\boldsymbol{D}_s\otimes(\boldsymbol{1}(\boldsymbol{D}_s\otimes\boldsymbol{D}(\boldsymbol{c}\boldsymbol{c}_s^{\mathrm{T}})))}{\boldsymbol{D}(\boldsymbol{c}\boldsymbol{c}_s^{\mathrm{T}})+\boldsymbol{D}_s\otimes(\boldsymbol{1}(\boldsymbol{D}_s\otimes(\boldsymbol{y}_t\boldsymbol{c}_s^{\mathrm{T}})))}, \tag{15}$$

式中，$\lambda_s$ 和 $\lambda_n$ 分别为控制 $\boldsymbol{c}_s$ 和 $\boldsymbol{c}_n$ 稀疏度的常数，$\boldsymbol{1}$ 为全 1 方阵；符号 $\otimes$ 与上文一致，表示的都是矩阵或向量中对应元素的乘。

最后，由于人耳对于相位信息不敏感，直接使用带噪语音的相位信息对 $\widehat{\boldsymbol{y}}$ 重构，再使用逆短时傅里叶变换 (Inverse Short Time Fourier Transformation, ISTFT) 输出相应的时域信号。

$$\widehat{\boldsymbol{y}}(t) = \mathrm{ISTFT}(\widehat{\boldsymbol{y}}_t \cdot \angle\boldsymbol{y}_t), \tag{16}$$

式中，$\angle\boldsymbol{y}_t$ 表示的是带噪语音帧 $\boldsymbol{y}_t$ 的相位，$\widehat{\boldsymbol{y}}(t)$ 表示的是增强后输出的语音时域信号。

## 3　增强算法性能评估

本节对所提算法进行实验仿真测试，并就其性能进行评估。实验中的纯净语音选自 TIMIT 标准语音库，噪声选自 Noisex-92 标准噪声库 [19]，将纯净语音和噪声进行下采样到 8 kHz，信噪比分别选取 −5 dB, 0 dB, 5 dB 和 10 dB 对算法进行评估。进行 STFT 时所使用的语音帧长为 256 点，选用 hamming 窗。缓存器能够存储 50 帧语音信号。自适应噪声字典中的原子个数设定为 40 。式 (10) 中的参数的选择参照文献 16 进行选取，分别令 $\alpha = 0.2$, $\beta = 0.8$ 。使用 GoDec 算法对缓存器中带噪语音帧的幅度谱进行稀疏低秩噪声分解时，设定低秩部分的秩为 2，并令迭代公式 (13)—(15) 的最大迭代次数为 200 。

在评价指标中，采用语音质量客观评估方法 (Perceptual Evaluation of Speech Quality, PESQ) 和 BSS-EVAL 评价体系 [20] 来分别评估增强语音的质量和增强算法的实际性能，并使用短时客观可懂度测量 (Short-Time Objective Intelligibility measure, STOI)[21] 来评估增强语音的可懂度。其中 PESQ 是 ITU-T 推荐的评估方法，是一种能够评价语音主观试听效果的客观计算方法，可以很好地近似平均意见得分 (Mean Opinion Score, MOS)，PESQ 的取值范围为 $-0.5 \sim 4.5$，得分越高说明算法增强效果越好； BSS-EVAL 是目前公认的性能较好的盲源分离算法评估体系，该评估体系通过计算信源引入噪声比 (Signal to Artifacts Ratio, SAR) 、信干比 (Signal to Interference Ratio, SIR) 和信源失真比 (Signal to Distortion Ratio, SDR)，从不同方面反映了增强算法的效果，该评估体系下的 SAR, SDR, SIR 指标经常被用于评估增强算法的性能； STOI 是一种较新的机器驱动的可懂度客观评估方法，其计算值与人对于语音的实际可懂度高度相关，测量结果的数值越高表明测试语音的可懂度越高。

测试实验中纯净语音为男女声各 5 句，主要选取了 Pink, F16, HF channel 和 Babble 四种典型同时又比较有挑战性的噪声进行测试，其中， Pink 是自然界中最常见的噪声，其频率分量功率主要分布在中低频段，与语音信号的能量分布类似； F16 为美军双座 F16 战斗机巡航过程中座舱内的噪声，能量集中在 0∼700 Hz 和 2750 Hz 并呈现出一定的非平稳特性； HF channel 为高频信道噪声，记录的是解调后的高频信道中的噪声，是通信领域经常需要处理的一类噪声； Babble 为一个容纳有大约 100 个同时在讲话的人的餐厅中的背景噪声，其能量分布与纯净语音类似，是典型的类语音噪声。

将本文所提的无监督实时增强算法 (URTE) 与多带谱减法 (MSS)[22] 和基于低秩稀疏矩阵分解的增强算法 [23] 两种无监督的语音增强算法进行比较和性能分析。多带谱减法已被证明具有比子空间方法或基于统计模型相当或更好的性能，该算法通过对于无语音期的噪声进行建模，得到平滑噪声幅度谱，并据此估计当前噪声。然而，实际的噪声幅度谱可能与所估计的平滑噪声谱存在较大偏差。基于低秩稀疏矩阵分解的增强算法的语音增强算法与本文算法一样，都使用稀疏低秩噪声模型 (其核心算法亦为 GoDec 算法，故性能测试时将其简记为 GoDec)，不同的是该算法直接遗弃低秩和噪声部分，只对稀疏部分进行重构获得增强语音，并没有对低秩噪声部分表现出的背景噪声特征加以利用来进一步提升增强效果；同时，该算法也是一种批处理的增强算法，无法对带噪语音进行实时增强。

表 1 和表 2 分别给出了 4 种噪声条件下各增强结果的 PESQ 和 STOI 测量值。从表 1 中可以看出，本文所提出的算法的性能显著优于另外两种算法，增强所得的语音失真更小，并且随着信噪比的下降，增强性能并不明显下降。这主要是因为本文算法是基于字典学习的增强算法，学习获得的自适应噪声字典反映的是噪声的结构特征，而对于噪声的能量或功率不敏感。表 2 中本文算法拥有最高的 STOI 测量值，说明使用本文算法进行增强后的语音具有最高的可懂度，这与通过 PESQ 测量值所得出的结论相类似。

然而， PESQ 和 STOI 都只能够从总体上评价算法性能，为了进一步评估算法各方面的性能，本文

使用 BSS-EVAL 评价体系来对几种算法进行进一步的评估。图 4 — 图 6 分别为 3 种算法在不同信噪比下的 SDR, SIR, SAR 性能曲线。

表 1 不同算法和噪声下的 PESQ 值

| 噪声 | SNR (dB) | 增强算法 | | |
|---|---|---|---|---|
| | | MSS | GoDec | URTE |
| Pink | −5 | 1.62 | 2.01 | 2.36 |
| | 0 | 2.28 | 2.37 | 2.61 |
| | 5 | 2.35 | 2.42 | 2.81 |
| | 10 | 2.59 | 2.61 | 2.94 |
| F16 | −5 | 1.63 | 1.89 | 2.38 |
| | 0 | 1.83 | 2.16 | 2.68 |
| | 5 | 2.31 | 2.33 | 2.83 |
| | 10 | 2.44 | 2.63 | 2.94 |
| HF channel | −5 | 1.44 | 1.66 | 2.25 |
| | 0 | 1.73 | 2.05 | 2.61 |
| | 5 | 2.17 | 2.43 | 2.78 |
| | 10 | 2.56 | 2.59 | 2.82 |
| Babble | −5 | 1.48 | 1.74 | 2.31 |
| | 0 | 1.94 | 1.98 | 2.55 |
| | 5 | 2.36 | 2.16 | 2.71 |
| | 10 | 2.57 | 2.40 | 2.81 |

表 2 不同算法和噪声下的 STOI 值

| 噪声 | SNR (dB) | 增强算法 | | |
|---|---|---|---|---|
| | | MSS | GoDec | URTE |
| Pink | −5 | 0.537 | 0.612 | 0.723 |
| | 0 | 0.582 | 0.676 | 0.780 |
| | 5 | 0.713 | 0.734 | 0.832 |
| | 10 | 0.785 | 0.787 | 0.883 |
| F16 | −5 | 0.543 | 0.590 | 0.739 |
| | 0 | 0.595 | 0.672 | 0.792 |
| | 5 | 0.697 | 0.735 | 0.849 |
| | 10 | 0.785 | 0.781 | 0.896 |
| HF channel | −5 | 0.521 | 0.625 | 0.740 |
| | 0 | 0.607 | 0.719 | 0.829 |
| | 5 | 0.695 | 0.798 | 0.885 |
| | 10 | 0.831 | 0.863 | 0.928 |
| Babble | −5 | 0.564 | 0.575 | 0.673 |
| | 0 | 0.640 | 0.671 | 0.735 |
| | 5 | 0.729 | 0.767 | 0.799 |
| | 10 | 0.779 | 0.808 | 0.828 |

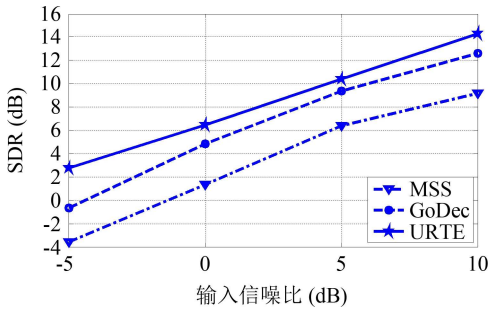


图 4 3 种算法在 SDR 测度下的性能曲线

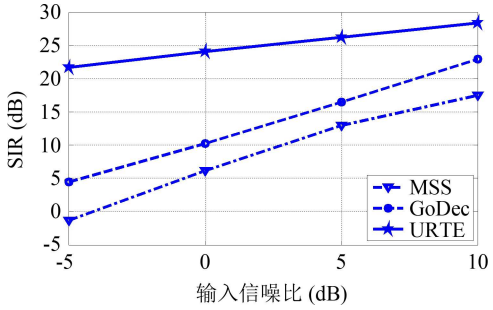


图 5 3 种算法在 SIR 测度下的性能曲线

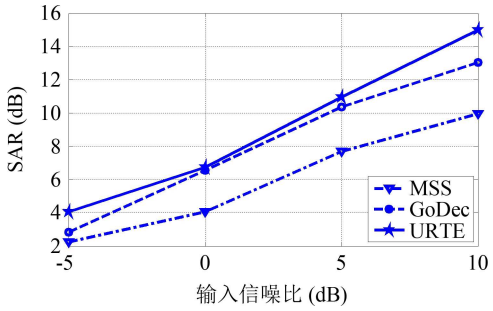


图 6 3 种算法在 SAR 测度下的性能曲线

图 4 为反映增强算法总体性能的 SDR 值随输入信噪比的性能变化曲线，由图中可以看出，所提的 URTE 算法比低秩稀疏矩阵分解的增强算法要高出大约 2 dB，比 MSS 算法要高大约 5 dB，据此可以得到与上文相一致的结论。

图5为反映增强算法对噪声抑制能力的 SIR 测度随输入信噪比的性能变化曲线。不难看出，所提算法对于噪声的抑制能力显著高于其他两种增强算法，比低秩稀疏矩阵分解的增强算法平均高出 11.5 dB，比 MSS 算法平均高出 16.2 dB 。同时，本文算法抑制噪声的性能随输入信噪比的降低下降缓慢，这主要得益于在线的字典学习方式能够更好地追踪噪声的统计学变化规律，使本文算法获得了相较于传统算法更强的噪声抑制能力，同时，基于特征的学习方式对于噪声的能量不敏感，因此在低信噪比条件下也能获得比较理想的抑噪性能。

图 6 为反映增强算法引入噪声大小的 SAR 测度随输入信噪比的性能变化曲线。可以看出，本文算法

与基于低秩稀疏矩阵分解的增强算法引入噪声的性能相类似，优于 MSS 算法。本文算法由于使用在线字典学习的学习方法，使得噪声字典能够反映噪声的实变特性，因此，对于纯净语音的估计相对比较准确，引入的噪声也比较少。基于低秩稀疏矩阵分解的增强算法在 SAR 测度上性能与本文算法类似，重构所得的语音中混杂的算法引入噪声相对较小。 MSS 算法的 SAR 性能相对较差是因为该算法使用的平滑噪声谱对于噪声的建模不一定准确，而不准确的模型是引入算法噪声的根源。

为了更好地描述出残留噪声和语音失真的细节信息，同时也为使实验结果更具说服力。本文给出了 3 种算法进行语音增强前后的语谱图。噪声选择富有挑战性的 Babble 噪声，输入信噪比为 5 dB 。

观察图 7 的语谱图可以看出，经过 3 种算法处理后的语谱图降低了噪声的含量，并且不难看出，本文算法的增强效果要好于前面两种方法， MSS 算法虽然抑制较多的噪声成分，但是由于估计不准确引入了很多额外的噪声；基于低秩稀疏矩阵分解的增强算法显著去除了高频部分的噪声，但是低频部分仍有较多残留，这主要是因为低频噪声具有一部分重复性不明显的成分，使得低秩和噪声部分无法将其完全分离。本文算法通过在线字典学习方式获取了背景噪声的特征，并将这些特征应用到实时增强的处理过程中，从而提升了增强算法的性能。

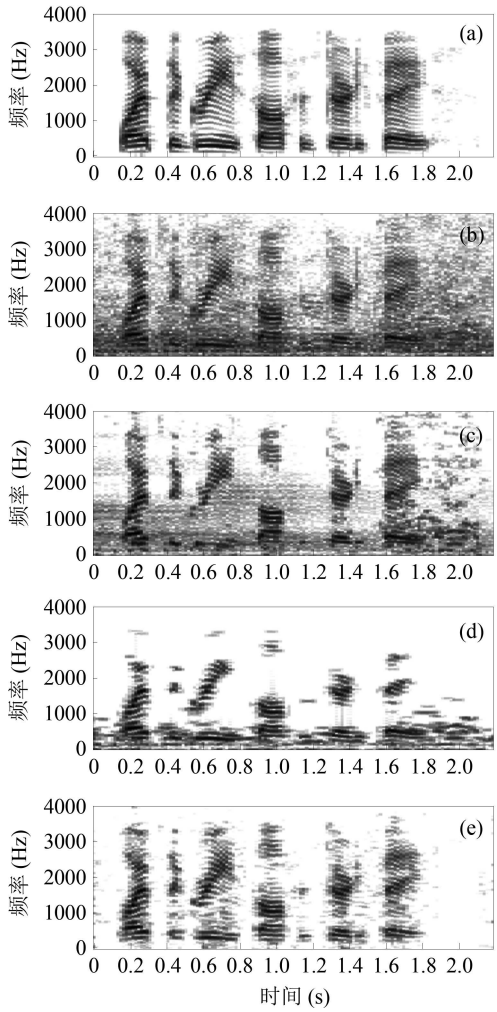


图 7　语音语谱图 ((a) 纯净语音； (b) 信噪比为 5 dB 的带噪语音； (c) 多带谱减法； (d) 基于低秩稀疏矩阵分解的增强算法； (e) 本文所提算法)

## 4　结束语

基于稀疏低秩噪声模型和增量非负子空间学习提出的一种无监督的实时增强算法，有效克服了传统字典学习算法依赖预先训练的离线字典、需要进行 VAD 以及难以实现实时增强等问题。算法继承了基于字典学习算法对噪声特征敏感而对能量不敏感的特性，在低信噪比条件下也能获得较理想的增强性能；同时，将增量非负子空间学习方法引入的噪声字典的训练过程中，使得增强算法能够动态追踪背景噪声的统计学变化规律，并获得了突出的噪声抑制能力。实验显示在具有挑战性的类语音噪声环境下，本文算法也具有较理想的增强效果。

## 参　考　文　献


1　Lu Yang, Loizou P C. A geometric approach to spectral subtraction. *Speech Communication*, 2008; **50**: 453—466

2　Lim J S, Alan V O. Enhancement and bandwidth compression of noisy speech. *Proceedings of IEEE*, 1979; **67**(12): 1586—1604

3　周彬，邹霞，张雄伟. 基于多元 Laplace 语音模型的语音增强算法. 电子与信息学报， 2012: 1562—1567

4　Jensen J R, Benesty J, Christensen M G, Jensen S H. Enhancement of single-channel periodic signals in the time-domain. *IEEE Trans. on Audio, Speech and Language Processing*, 2012; **20**(7): 1948—1963

5　Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Trans. on Audio, Speech and Language Processing*, 2013; **21**(10): 2140—2151

6　Wilson K, Raj B, Smaragdis P, Divakaran A. Speech denoising using nonnegative matrix factorization with priors. ICASSP, 2008: 4029—4032

7　Mohammadiha N, Leijon A. Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement. *IEEE Trans. on Audio, Speech and Language Processing*, 2013; **21**(5): 998—1011

8　Schmidt M, Larson J. Reduction of non-stationary noise using a non-negative latent variable decomposition. *IEEE Workshop on Machine Learning for Signal Process* (MLSP), 2008: 486—491

9　黄建军，张雄伟，张亚非，邹霞. 时频字典学习的单通道语音增强算法. 声学学报， 2012; **37**(5): 539—547

10 Sigg C D, Dikk T, Buhmann J M. Speech enhancement using generative dictionary learning. *IEEE Transactions on Audio, Speech and Language Processing*, 2012; **20**(6): 1698—1712
11 Po-Sen Huang, Chen S D, Smaragdis P *et al*. Singing-voice separation from monaural recordings using robust principal component analysis. ICASSP, 2012: 57—60
12 Candès E J, Li Xiaodong, Ma Yi *et al*. Robust principal component analysis? *Journal of the ACM*, 2011; **58**(3): 1—37
13 Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of a corrupted low-rank matrices. http://arxiv.org/abs/1009.5055, 2010
14 Zhou Tianyi, Tao Dacheng. GoDec: randomized low-rank & sparse matrix decomposition in noisy case. International Conference on Machine Learning (ICML), 2011: 33—40
15 Zhou Tianyi, Tao Dacheng. Shifted subspaces tracking on sparse outlier for motion segmentation. International Joint Conference on Artificial Intelligence, 2013: 1946—1952
16 Bucak S S, Gunsel B. Incremental subspace learning via non-negative matrix factorization. *Pattern Recognition*, 2009; **42**(5): 788—797
17 Smaragdis P. Convolutive speech bases and their application to supervised speech separation. *IEEE Trans. on Audio, Speech and Language Processing*, 2007; **15**(1): 1—12
18 Andersen K T. Wind noise reduction in single channel speech signals. Technical University of Demark, 2008
19 Rice University Digital Signal (Dsp) Group, Noisex92 Noise Database. http://spib.rice.edu/spib/select_noise.html, 1995
20 Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 2006; **14**(4): 1462—1469
21 Taal C H, Hendriks R C, Heusdens R. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. on Audio, Speech and Language Processing*, 2011; **19**(7): 2125—2136
22 Loizou P C. Speech Enhancement: Theory and Practice, Boca Raton, Taylor and Francis, 2007: 120—125
23 Huang J, Zhang X, Zhang Y, Zou X, Zeng L. Speech denoising via low-rank and sparse matrix decomposition. *ETRI Journal*, 2014; **36**(1): 167—170