doi:10.6043/j.issn.0438-0479. 201702026

**一种基于非参数贝叶斯理论的**

**语音增强算法**

吴佳雯1，刘沁婷1，曾德炉2，丁兴号1，李琳1\*

1. 厦门大学信息科学与技术学院，福建 厦门 361005; 2. 华南理工大学数学学院，

广东 广州 510641）

**摘要**：一种基于非参数贝叶斯理论的语音增强算法，在稀疏表示的框架下，把字典学习、稀疏系数表示和噪声方差估计融合成一个贝叶斯后验估计的过程，并利用Spike-Slab先验加强稀疏性（nonparametric Bayesian method with Spike-Slab priori，NBSP）。首先，将带噪语音分解为干净语音、高斯类型噪声和残余噪声三个子信号，分别对该三种子信号采用不同的先验概率模型表达，接着采用马尔科夫链-蒙特卡洛算法计算出三个模型中每个参数对应的后验概率，最后基于稀疏表示的框架重构出干净语音。实验数据使用NOIZEUS语音库，采用PESQ和SegSNR作为质量评价指标，分别在信噪比为0 ，5和10 dB的高斯白噪声、火车噪声和街道噪声上验证了其可行性，并与多种常用语音增强方法进行对比，发现其在低信噪比非平稳噪声情况下的增强效果更为理想。

**关键词**：稀疏表示；非参数贝叶斯理论；Spike-Slab先验；自适应字典；语音增强

**中图分类号：**TN 912 **文献标志码：**A

在日常生活中，由于周围环境嘈杂及录音设备的干扰，语音信号经常携带有冗余的噪声，在一定程度上影响了人们的信息交流。尤其，随着人工智能在日常应用的普及化，语音成为了一种重要的人机交互手段。因此，研究语音增强算法，以获取高自然度和清晰度的语音信号，是语音信号处理领域的热点问题之一。

从1960年起，研究者们提出了非常多的语音增强算法，常见有谱减法[1]、维纳滤波法[2]、小波阈值法[3]、子空间[4]、噪声估计[5-6]等。近几年来，基于稀疏表示理论的语音增强算法[7-11]受到越来越多的关注。稀疏表示理论以两个假设为前提，一个是噪声信号与语音信号是不相关的，另一个是语音信号在过完备字典上的表示是稀疏的，因此字典的构建和目标函数的优化是基于稀疏表示理论的语音增强算法的两个重要部分。目前主要的字典构建方法有两种：1）固定字典法，该方法直接采用小波变换，离散余弦变换等方法生成字典的基底，运算量小，但无法自适应地表达不同的语音信号；2）训练字典法，该方法采用如MOD[11]和K-SVD[12]这类的训练算法直接从样本数据中学得字典，这些字典原子更加符合信号本身的特性，但计算复杂，字典结构较为松散。不管是固定字典法还是传统的训练字典法，都需要使用者在训练字典事先提供稀疏度或者误差阈值来决定字典的结构，在实际应用中较难满足上述条件，而不准确的稀疏度和误差阈值将导致不佳的增强效果。

受到非参数贝叶斯理论在图像去噪上成功应用[12]的启发，本文采用基于非参数贝叶斯理论的字典学习方法进行语音信号的稀疏表示，利用Spike-Slab先验[13-15]以自适应保证稀疏表示结构中模型参数的稀疏性，以减少信号模型对参数的依赖程度。该模型假设带噪语音由干净语音、高斯噪声和残余噪声三个子信号相加而成，干净语音由字典和稀疏系数组成，接着对字典、稀疏系数、高斯噪声和残余噪声进行先验概率建模，假设其分别服从一定的先验分布，同时根据共轭先验分布对每个分布中的参数进行建模，其中稀疏系数和残余噪声服从Spike-Slab先验以保证稀疏性，最后采用马尔科夫链-蒙特卡洛算法[16-18]求得各个参数对应的后验概率，从而重构出干净噪声。基于非参数贝叶斯理论的语音增强方法具有如下优点：

1. 参数自适应。该算法的参数都是由模型通过数据学习到的，每个参数只需要一个合理的先验概率模型即可，不需要人为设定一个固定值，这使得模型更具有适应性和灵活性。同时，字典中被使用的原子数目也是根据数据学到的，因此只要设定一个稍大的初始字典规模，就能自适应收敛到合适的字典规模。
2. 不需要提前估计噪声方差。在实际应用中，要准确估计带噪语音的噪声方差是较为困难的，尤其对于那些被不同强度噪声污染的连续一段语音，而该方法不需要提前估计噪声方差，可以根据给定数据和后验概率估计出干净语音。
3. 提供了一种在线学习的方法。该方法不需要提前进行训练，对于单条短语音即可进行字典学习，具有较强的实用性和便捷性。
4. **Spike-Slab先验**

Spike-Slab先验[13-15]是一种稀疏先验，用来加强系数先验信息的稀疏性，一般由连续的分布（如高斯分布）和一个分布混合而成，其中，连续分布为Slab项，分布为Spike项。狄拉克函数（Dirac Delta function，函数），亦称为单位脉冲函数，表示在零点处有值，其他点处为零的函数。当均值为零的高斯分布方差无限小的时候就成为分布，该分布的结果具有严格的稀疏性，因此可以作为稀疏表示系数的先验信息来使用。 其一般形式为:



其中，Spike项为分布，Slab 项为高斯分布，为权重，服从超参数为和的Beta分布，越小，其对应的系数就越有可能落入Spike 项，即越稀疏，因此的取值可以控制系数的稀疏度。

1. **基于Spike-slab先验的非参数贝叶斯算法**

一条带噪语音可以表示为

 (1)

其中，、和分别表示时域上的带噪语音，干净语音和噪声。通常，对语音进行分帧处理，如把一条语音分成帧长为的帧，则每一帧语音可以表示为

  (2)

在常用的稀疏表示方法中，用字典和稀疏系数的乘积来近似表示干净语音，即

  (3)

其中，由个原子组成的字典，是对应的稀疏系数。

在实际应用中，噪声可以表示为高斯类型噪声和残余噪声，因此有

  (4)

假设字典中的每一个原子之间是独立同分布的，并且都服从均值为0的高斯分布，即

 (5)

其中i+εi, i=1,…,P其中是维度为的单位矩阵。

高斯噪声服从均值为0，方差为的高斯分布，其中，为噪声精度，即噪声方差的倒数，服从超参数为和伽马分布，如下所示，

 (6)

考虑到稀疏系数和残余噪声的稀疏特性，假设两者都服从Spike-Slab先验，其中稀疏系数的表达如下所示，

 (7)

残余噪声的表示为

 (8)

式中，参数表示对应的字典原子被用于表示重构语音的概率，其服从超参数为和的Beta分布；参数表示当前数据点为残余噪声的概率，同样地，参数服从超参数为和的Beta分布；和服从伽马分布，用于估计精度(方差的倒数)，超参数分别为和。需要注意的是，语音信号的所有帧共用一个字典，因此与字典原子相关的参数也是所有帧共用的，而与残余噪声相关的参数则不同，由于每个数据点为残余噪声的概率是相互独立的，因此一个数据点就对应着一个。

 式子（4）-（8）为所有参数的先验概率模型表达，接着根据训练数据和贝叶斯推理，可以得到参数的后验概率分布。采用马尔科夫链-蒙特卡洛算法[28–30]方法计算模型的近似结果，如下:

1. 字典原子：

 (9)

其中，方差为，均值为。

1. 稀疏系数：

 (10)

其中，方差为，均值为，且有。

1. 概率参数：

 (11)

其中，表示不等于0的个数，表示等于0的个数

1. 精度：

 (12)

1. 残余噪声：

 (13)

其中，方差为，均值为，且有。

另外，在计算残余噪声之前，需要移除掉理想信号，因此有。

1. 概率参数:

 (14)

其中，表示不等于0的个数，表示等于0的个数

1. 精度：

 (15)

1. 精度：

 (16)

根据式子（9）-（16），将所有的待求变量进行逐步迭代求解，在其他变量固定前提下求解当前变量，然后对下一个变量进行上述相同操作，直至收敛到最终的稳定状态（在后续实验中，根据实验经验设定迭代次数为100次），取稳定状态的和，根据式(3)重构出增强后的语音信号。

1. **实验结果及分析**

**3.1 实验设置**

用于实验的30条干净语音来自NOIZEUS数据库[19]，分别由男、女性录制而成。带噪语音由干净语音叠加噪声所得，共270条。信噪比分别为0dB、5dB、和10dB，噪声种类有高斯白噪声，火车噪声和街道噪声。所有语音的采样率均为8KHz，帧长为128，帧移为1。初始字典大小，考虑到实际使用的字典原子个数由数据本身决定，因此把初始字典大小设大一些，能更好地适应不同的语音数据。为了使先验信息对结果的影响尽可能的小，超参数取较小的值，如[20]。由于计算后验分布时，数据样本的影响远大于这些超参数，因此Spike-Slab先验中Beta分布的超参数和的选取对增强性能的影响较小，不同的取值影响着对应变量的初始稀疏度[15]，考虑到稀疏系数和残余噪声是稀疏的这一先验信息，因此取和的值分别为1，100P，0.1和0.9。

语音质量评价指标采用主观语音质量评价（PESQ）[21]和分段信噪比（SegSNR），实验结果均为30条语音的平均值，其中分段信噪比的公式如下所示，

 (17)

其中，表示语音帧数，表示语音帧长，**为增强语音，**为干净语音。

**3.2 与K-SVD算法的比较**

K-SVD算法作为稀疏表示算法中的典型算法，一直以来都受到学者们的广泛关注，其采用如下所示的目标优化函数

 (18)

其中是权重。在满足或的时候，迭代停止，因此需要事先给定一个稀疏度或者一个阈值。尽管目前话音激活算法[22]可以用于估计噪声频谱，但在实际应用中要做到准确估计仍旧较难实现。而当噪声方差估计不准确时，K-SVD的结果会有非常明显的下降，如图1所示。图中所使用的数据是用30条干净语音叠加不同程度的高斯白噪声，叠加后的信噪比分别为0dB、5dB和10dB。图1(1)和(2)分别为30条语音的平均PESQ值和平均SegSNR值，从图中可以看出，当噪声方差估计不准确（K-SVD Mismatch）的时候，K-SVD算法的结果会有明显的下降；当噪声方差估计准确（K-SVD Match）的时候，基于非参数贝叶斯理论的语音增强方法与其性能相当，但值得注意的是，基于非参数贝叶斯理论的语音增强方法不需要预先估计噪声方差，只需要对式（6）和（8）中的超参数设定一个初始值（具体见3.1实验设置），就可以通过数据本身和先验分布计算得到对应噪声精度（噪声方差的倒数）的后验分布，如式（15）和（16）所示，因此相比于K-SVD算法，该方法对参数的依赖性更低。

 

(1) (2)

图1 NBSP算法与K-SVD算法对高斯白噪声的增强效果对比，(1)PESQ的值；(2)SegSNR的值。

其中K-SVD Match为噪声方差估计准确的情况，K-SVD Mismatch为噪声方差估计不准确的情况。

Fig.1 Comparison with K-SVD. (1)PESQ; (2)SegSNR

**3.3 与常用语音增强算法的比较**

为了进一步探究NBSP算法的有效性，本节把NBSP算法同目前常用的几种语音增强算法进行了对比，主要有谱减法（SS）[1]、维纳滤波法（Wiener）[2]、子空间（Subspace）[4]、使用非因果先验的最小化均方误差算法(MMSE\_NPS)[5]、基于估计频谱幅值平方最大后验的增强算法( MSS\_MAP)[6]和基于低秩约束的稀疏矩阵分解法（CLSMD）[11]，对比数据使用的是参考文献[11]中的实验结果，如表1，表2所示。

由表1和表2观察可知，NBSP方法在处理低信噪比非平稳噪声的带噪信号时，具有较大的优势。在带噪语音信噪比为0dB和5dB两种情况下，从SegSNR的值来看，除了在0dB火车噪声情况下略差于CLSMD方法外，NBSP方法取得了所有的最好值，这说明NBSP方法在低信噪比情况下去除噪声的能力强，值得注意的是在5dB街道噪声时，NBSP方法比第二好的结果有大约125%的提

表1 采用不同方法增强后的SegSNR值比较

Tab.1 Performance Comparisons in Term of SegSNR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 噪声类型 | 去噪算法 | 0dB | 5dB | 10dB |
| 高斯白噪 | NBSP | **3.54** | **5.23** | **6.23** |
| CLSMD[11] | 2.15 | 3.69 | 4.86 |
| MSS\_MAP[8] | 0.36 | 2.63 | 5.14 |
| SS[1] | -1.00 | 0.60 | 3.14 |
| Subspace[6] | 0.47 | 3.00 | 5.53 |
| MMSE\_NPS[7] | -0.76 | 0.47 | 1.91 |
| Wiener[2] | -0.46 | 0.79 | 2.12 |
| Noisy Speech | -4.81 | -2.01 | 1.13 |
| 火车噪声 | NBSP | -0.69 | **2.15** | 3.54 |
| CLSMD[11] | **-0.18** | 1.65 | 3.34 |
| MSS\_MAP[8] | -1.58 | 0.93 | **3.59** |
| SS[1] | -1.45 | 0.86 | 3.12 |
| Subspace[6] | -2.45 | 0.16 | 2.73 |
| MMSE\_NPS[7] | -2.48 | -0.69 | 0.87 |
| Wiener[2] | -1.06 | 0.42 | 1.91 |
| Noisy Speech | -4.50 | 1.69 | 1.42 |
| 街道噪声 | NBSP | **-0.71** | **2.07** | 3.39 |
| CLSMD[11] | -1.39 | 0.92 | 2.55 |
| MSS\_MAP[8] | -1.71 | 0.54 | **3.51** |
| SS[1] | -0.99 | 0.71 | 3.26 |
| Subspace[6] | -2.62 | -0.22 | 2.37 |
| MMSE\_NPS[7] | -2.07 | -0.80 | 0.89 |
| Wiener[2] | -0.84 | 0.22 | 1.80 |
| Noisy Speech | -4.26 | -1.58 | 1.84 |

表2 采用不同方法增强后的PESQ值比较

Tab.2 Performance Comparisons in Term of PESQ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 噪声类型 | 去噪算法 | 0dB | 5dB | 10dB |
| 高斯白噪 | NBSP | **2.04** | **2.42** | 2.65 |
| CLSMD[11] | 2.01 | 2.38 | 2.59 |
| MSS\_MAP[8] | 1.99 | 2.32 | **2.66** |
| SS[1] | 1.63 | 1.85 | 2.33 |
| Subspace[6] | 1.72 | 2.19 | 2.60 |
| MMSE\_NPS[7] | 1.87 | 2.17 | 2.49 |
| Wiener[2] | 1.46 | 1.71 | 2.14 |
| Noisy Speech | 1.56 | 1.83 | 2.14 |
| 火车噪声 | NBSP | **1.69** | **2.09** | 2.37 |
| CLSMD[11] | 1.67 | **2.09** | **2.41** |
| MSS\_MAP[8] | 1.61 | 2.02 | 2.38 |
| SS[1] | 1.40 | 1.94 | 2.33 |
| Subspace[6] | 1.31 | 1.72 | 2.10 |
| MMSE\_NPS[7] | 1.56 | 1.99 | 2.36 |
| Wiener[2] | 1.43 | 1.82 | 2.22 |
| Noisy Speech | 1.60 | 1.86 | 2.16 |
| 街道噪声 | NBSP | **1.76** | **2.11** | **2.43** |
| CLSMD[11] | 1.63 | 2.04 | 2.35 |
| MSS\_MAP[8] | 1.67 | 2.04 | 2.41 |
| SS[1] | 1.66 | 2.03 | 2.41 |
| Subspace[6] | 1.38 | 1.76 | 2.11 |
| MMSE\_NPS[7] | **1.76** | 2.07 | 2.41 |
| Wiener[2] | 1.51 | 1.83 | 2.21 |
| Noisy Speech | 1.56 | 1.90 | 2.25 |

升；从PESQ的值来看，NBSP方法取得了全部的最高值，仅在5dB火车噪声情况和0dB街道噪声情况下效果分别与CLSMD方法和MMSE\_NPS方法持平。SegSNR用于衡量增强算法的去噪能力，PESQ是一种模拟人耳听觉感受的主观质量评价算法，NBSP方法在两者都取得好结果，说明该方法在低信噪比情况下不仅去噪能力强，而且增强后的语音拥有着良好的听觉效果，而目前大多算法都偏向某一个指标的优势，很难实现二者的均衡，尤其是在针对非平稳噪声情况下，如街道噪声。在10dB带噪信号情况下，NBSP方法具有比较平稳的增强性能。

另外，为了探究NBSP方法在不同数据库以及更多噪声上的适应性，选取TIMIT数据库[23]中的20条不同的语音（男女比例1：1，每人一条语音），分别叠加来自Noisex92 噪声库[24]的高斯白噪声（WGN）和babble噪声，噪声强度为-5dB，0dB，5dB和10dB，所有语音的采样率均为16KHz。用NBSP方法进行增强，结果如表3所示，从表中数据可以看出，NBSP方法在TIMIT数据库上的有效性。

表3 NBSP算法在TIMIT数据库上的增强结果

Tab.3 Performance of NBSP in TIMIT database

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 噪声类型 | -5dB | 0dB | 5dB | 10dB |
| SegSNR | PESQ | SegSNR | PESQ | SegSNR | PESQ | SegSNR | PESQ |
| Babble | NBSP | -2.02 | 1.49 | -1.58 | 1.90 | 0.67 | 2.21 | 3.23 | 2.51 |
| Noisy | -6.94 | 1.36 | -4.39 | 1.70 | -1.18 | 2.07 | 2.32 | 2.42 |
| WGN | NBSP | 1.87 | 1.72 | 3.47 | 2.13 | 5.04 | 2.47 | 6.20 | 2.69 |
| Noisy | -7.21 | 1.27 | -4.68 | 1.56 | -1.51 | 1.91 | 2.10 | 2.27 |

1. **总 结**

NBSP算法在稀疏表示框架下，引入了基于Spike-Slab先验的非参数贝叶斯理论。该算法不需要预先设定一个约束项或者稀疏度，可以直接从数据本身学得字典、干净语音、噪声方差等参数。除此之外，基于非参数贝叶斯理论的字典学习方法可以直接对单条语音进行处理，不需要利用其他数据提前训练字典。实验结果表明了该方法的可行性，尤其在低信噪比非平稳噪声情况下具有较大的优势。但不足之处在于，该方法与其他稀疏表示方法一样具有较高的计算复杂度，不适用于实时处理应用场景。

**参考文献：**

[1] Boll, S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1979, 27(2): 113–120.

[2] Scalart, P. Speech enhancement based on a priori signal to noise estimation[C] // IEEE. Acoustics, Speech, and Signal Processing(ICASSP), 1996 IEEE International Conference on. Atlanta, Georgia: IEEE, 1996: 629-632.

[3] Vetterli M. Wavelets, approximation, and compression[J]. IEEE Signal Processing Magazine, 2001, 18(5): 59-73.

[4] De M B. The singular value decomposition and long and short spaces of noisy matrices[J]. IEEE transactions on signal processing, 1993, 41(9): 2826-2838.

[5] Cohen I. Speech enhancement using a noncausal a priori SNR estimator[J]. IEEE signal processing letters, 2004, 11(9): 725-728.

[6] Lu Y, Loizou P C. Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty[J]. IEEE transactions on audio, speech, and language processing, 2011, 19(5): 1123-1137.

[7] Zhao N, Xu X, Yang Y. Sparse representations for speech enhancement[J]. Chinese Journal of Electronics, 2011, 19(2): 268-272.

[8] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(10): 2140-2151.

[9] 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.

[10] Baby D, Virtanen T, Barker T. Coupled dictionary training for exemplar-based speech enhancement[C]//IEEE. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. Florence, Italy: IEEE, 2014: 2883-2887.

[11] Sun C, Zhu Q, Wan M. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition[J]. Speech Communication, 2014, 60: 44-55.

[12] Zhuang P, Huang Y, Zeng D, et al. Mixed noise removal based on a novel non-parametric Bayesian sparse outlier model[J]. Neurocomputing, 2016, 174: 858-865.

[13] Mitchell T J, Beauchamp J J. Bayesian variable selection in linear regression[J]. Journal of the American Statistical Association, 1988, 83(404): 1023-1032.

[14] George E I, McCulloch R E. Variable selection via Gibbs sampling[J]. Journal of the American Statistical Association, 1993, 88(423): 881-889.

[15] Ding X, Mi Z, Huang Y, et al. Robust RVM based on spike-slab prior[J]. Journal of Electronics (China), 2012, 29(6): 593-597.

[16] Hammersley J. Monte carlo methods[M]. Springer Science & Business Media, 2013.

[17] Jackman S. Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo[J]. American Journal of Political Science, 2000: 375-404.

[18] Sun S. A review of deterministic approximate inference techniques for Bayesian machine learning[J]. Neural Computing and Applications, 2013, 23(7-8): 2039-2050.

[19] Hu Y, Loizou P. Subjective comparison of speech enhancement algorithms[C]//IEEE. Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on. Toulouse: IEEE, 2006: I-I.

[20] Tipping M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of machine learning research, 2001, 1(Jun): 211-244.

[21] Rix A W, Beerends J, Hollier M, et al. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs[J]. ITU-T Recommendation, 2001, 862.

[22] Tashev I, Lovitt A, Acero A. Unified framework for single channel speech enhancement[C]// IEEE. Communications, Computers and Signal Processing, 2009 IEEE Pacific Rim Conference on. Victoria, BC: IEEE, 2009: 883-888.

[23] Garofolo J S. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database[J]. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988, 107.

[24] Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech communication, 1993, 12(3): 247-251.

Speech Enhancement Based on Nonparametric Bayesian Method

WU Jiawen1, LIU Qinting1，ZENG Delu2，DING Xinghao1, LI Lin1\*

(1. School of Information Science and Engineer, Xiamen University, Xiamen 361005, China; 2. School of Mathematics, South China University of Technology, Guangzhou 510641, China)

**Abstract：**A new speech enhancement strategy is proposed by utilizing a Bayesian nonparametric method with Spike-Slab priori. As a sparse representation framework, the dictionary learning, sparse coefﬁcients representation and noise variance estimation are replaced by a single procedure of Bayesian posterior estimation. Firstly, the noisy speech is divided into clean speech, Gaussian noise and rest noise. Then, each part is modeled with a certain priori distribution. Finally, by using Markov Chain Monte Carlo sampling algorithm, the posterior distribution can be obtained, as well as the clean speech and all the other parameters. Without knowing the noise variance, the proposed method could be performed directly on the noisy speech to infer the sparsity of the speech. Experiments were executed on NOIZEUS database. The noisy speech has three types of noise (white, train and street) with SNR range from 0dB to 10dB. And the subjective and objective measures like PESQ score and the output SegSNR are implemented to evaluate the performance of the proposed method and the other state-of-the-art methods. The corresponding results show that this proposed speech enhancement method achieves better performance, especially in conditions of non-stationary noise with low input SNR.

**Key words:** sparse representation；nonparametric Bayesian Method；Spike-Slab priori; dictionary learning speech enhancement