doi:10.3969/j.issn.1000-436x.2014.02.012

基于多任务稀疏表达的二元麦克风小阵列话音增强算法

杨立春[1,2]，叶敏超[1]，钱沄涛[1]

(1. 浙江大学 计算机科学与技术学院，浙江 杭州 310027；2. 浙江万里学院 智能控制技术研究所，浙江 宁波 315101)

摘 要：针对常规二元麦克风小阵列话音增强算法通常需要话音活动检测技术支持，并且难以有效抑制第一帧含目标信号的噪声。提出了一种基于多任务稀疏表达的二元麦克风小阵列话音增强算法，首先利用字典学习方法分别获得目标信号和噪声信号的过完备字典，然后利用 ℓ_2/ℓ_1 混合范数对信号在其字典上的表示系数进行正则化稀疏约束，使得 2 个阵元接收到信号中的噪声信号被抑制，而话音信号尽量保持不变，从而达到话音增强的目标。仿真和实验数据表明，无论开始位置是否含有目标话音信号，所提出的非话音活动检测支持的二元麦克风小阵列话音增强算法均能有效实现话音增强的目标。

关键词：麦克风小阵列；话音增强；字典学习；多任务稀疏表达

中图分类号：TN912.35 **文献标识码**：A **文章编号**：1000-436X(2014)02-0087-08

Speech enhancement based on multi-task sparse representation for dual small microphone arrays

YANG Li-chun[1,2], YE Min-chao[1], QIAN Yun-tao[1]

(1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China;
2. Intelligent Control Research Institute, Zhejiang Wanli University, Ningbo 315101, China)

Abstract: Speech enhancement algorithms for dual small microphone arrays usually rely on the voice activity detection(VAD), and they may fail in some cases when target speech signal is included in the first frame. A multi-task sparse representation based speech enhancement algorithm was proposed. First, dictionaries for signal and noise were respectively formed via dictionary learning. Then the noise in signals obtain from two microphones was reduced by ℓ_2/ℓ_1 regularized sparse representation on the over-complete dictionary, while the target speech signals were mostly preserved, hence the speech signals were enhanced. Experimental results from synthetic and real-world data show that the proposed speech enhancement algorithm without VAD works well in all cases no matter speech signal is included in the first frame or not.

Key words: small microphone arrays; speech enhancement; dictionary learning; multi-task sparse representation

1 引言

二元麦克风小阵列被广泛应用在手机、助听器等受空间、运算能力和成本限制的设备中，用以实现话音增强。自适应波束形成是二元麦克风小阵列话音增强的常用算法[1~5]，其思想是通过期望目标方向信号获得最大增益，并通过权系数的更新估计非目标方向干扰信号实现话音增强。为了防止目标话音信号失真，权系数在话音段需停止更新，而这需要话音活动检测(VAD, voice activity detection)技术支持，同时要求处理信号的开始阶段为非话音段，因此 VAD 的准确性成为影响波束形成话音增强效果的重要因素。

另一种常见的二元麦克风小阵列话音增强算法是相干滤波器(coherence-based filter)方法[6]，通过假定阵元间目标话音信号相关而噪声信号不相关，

收稿日期：2012-12-07；**修回日期**：2013-05-06

基金项目：国家自然科学基金资助项目（61171151）；国家重点基础研究发展计划（“973”计划）基金资助项目（2012CB316400）；国家科技支撑计划基金资助项目（2011BAD24B03）

Foundation Items: The National Natural Science Foundation of China (61171151); The National Basic Research Program of China (973 Program) (2012CB316400); The National Key Technology R&D Program of China (2011BAD24B03)

使用基于互功率谱密度的相干函数进行降噪。实际环境中，尤其是在小阵列中，这种假设往往不成立，因而一般通过估计噪声谱的方法进行相干滤波[7]。与波束形成方法一样，噪声谱估计方法也要求目标话音信号不能出现在处理信号的第一帧位置，且通常需要 VAD 技术支持，以防止噪声谱估计错误造成话音信号失真。

近年来基于多任务稀疏表达的学习方法[8~10]在去噪领域得到研究，该方法通过构造固定的字典或动态学习得到的字典，在特定的约束下可以使信号在此字典上表示的系数稀疏化，当多个任务中某些信号在此字典上表达的系数近似相同时将会被保留，而那些系数不同的信号将被抑制。

稀疏编码(sparse coding)在单道话音增强[11~13]算法中的研究表明，话音信号可以使用合适字典中的少量基函数进行表达。而高斯白噪声等随机噪声不能被少量基函数完整表达，故单道稀疏编码算法对这些类型噪声的抑制较好。但无论是构造固定字典还是通过学习的字典均无法完全将目标信号和非平稳噪声分离，所以同其他单道算法一样，该算法也不能对非平稳噪声进行有效抑制。

本文提出了一种基于多任务稀疏表达的二元麦克风小阵列话音增强算法，当 2 个阵元接收到的目标信号通过时延补偿使得它们在同一时刻基本一致，而噪声信号不一致时，对目标信号和噪声信号在各自的字典上进行稀疏表达，2 个阵元中的目标信号对应其字典上的系数应基本相同，而噪声信号对应其字典上的系数不同，使用多任务稀疏表达即可将这些不一致的噪声信号系数进行抑制，从而实现降噪的目的。

话音信号的字典通过语料库离线学习获得。由于话音信号具有共性特征，用语料库离线学习得到的字典与说话人及环境都没有关系，具有通用性。而噪声字典使用通过阵列得到的参考噪声信号在线学习得到。因环境噪声的多变特性，在线学习方法能够保证噪声字典对环境噪声的适应性。由于噪声抑制主要通过其在 2 个阵元字典表达上的一致性实现，因而泄露到参考噪声中的少量话音信号对降噪效果影响可以忽略。因此本文算法无需使用 VAD 算法，也没有第一帧要求非话音的限制，保证了算法的稳定性和通用性。

2 信号模型

考虑一个由 2 个全指向性麦克风组成的二元麦克风小阵列，假定目标话音信号和噪声信号不相关，则到达 2 个阵元的信号可以表示为

$$y_i(t)=x_i(t)+n_i(t),i=1,2 \tag{1}$$

其中，$x_i(t)$ 表示阵元接收到的目标话音信号；$n_i(t)$ 表示阵元接收到的噪声信号。把式(1)两边同时进行短时傅里叶变换转换成频域形式为

$$Y_i(t,\mathrm{e}^{\mathrm{j}\omega})=X_i(t,\mathrm{e}^{\mathrm{j}\omega})+N_i(t,\mathrm{e}^{\mathrm{j}\omega}),i=1,2 \tag{2}$$

阵元间目标信号的时延差可通过时延估计算法实现[14,15]。由于在小阵列中阵元间距较小，在采样率不够高的情况下，2 个阵元的目标信号的时延一般小于一个采样点，此时 2 个阵元的目标信号仅在相位上有差别。当目标信号以与阵列的第1个阵元成 θ 角方向传播时，第 2 个阵元接收的目标信号与第1个阵元的目标信号的相位差为 $\mathrm{e}^{-\mathrm{j}\omega d\cos\theta/c}$。对第 2 个阵元的目标信号进行相位补偿后可得

$$\begin{aligned}Y_2(t,\mathrm{e}^{\mathrm{j}\omega})&=X_2(t,\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\mathrm{j}\omega d\cos\theta/c}+N_2(t,\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\mathrm{j}\omega d\cos\theta/c}\\&=X_1(t,\mathrm{e}^{\mathrm{j}\omega})+N_2(t,\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\mathrm{j}\omega d\cos\theta/c}\end{aligned} \tag{3}$$

其中，θ 为目标声源与阵列的第1个阵元方向的夹角；ω 为频率因子，d 为阵元间距；c 为声波在空气中传播的速度。式(3)和式(2)表示阵元1接收到的含噪信号(当 $i=1$ 时)，可以看出经过相位补偿后的 2 个阵元中的目标信号完全相同，而非目标信号方向上的噪声信号则不同。为了处理方便，把经过相位补偿后的 2 个信号经过傅里叶反变换转到时域形式，分别为 $y_{1p}(t)$ 和 $y_{2p}(t)$。

3 多任务稀疏性约束话音增强算法

假定阵列中每个阵元接收到的噪声信号是式(1)所示的加性噪声，经相位补偿后第 2 个阵元含噪信号如式(3)所示，与式(2)第 1 个阵元含噪信号相比，目标信号基本相同，而噪声信号不同，因而符合多任务稀疏学习降噪的条件。基于多任务稀疏性约束话音增强算法主要包括两部分：1) 通过字典学习找到目标话音信号和噪声信号合适的字典；2) 通过混合字典的多任务稀疏表达实现噪声抑制。

3.1 字典学习

通用基函数，例如小波基、离散余弦变换(DCT, discrete cosine transform)基，由于可以作为任意非随机信号的字典，因而很难使用它们分离目标信号和噪声信号。字典学习的目标是通过使用某种类型信

号的训练样本，获得符合其特征基向量组成的字典。学习得到的自适应字典可以较好地重构与训练样本信号较为类似的信号，而不能完全重构其他与训练样本信号差异较大的信号。因而相对通用的基函数字典，通过学习得到的字典可以更好地实现信号分离。在字典中，每列也被称为“原子”，非随机信号如果能使用少量字典中的“原子”线性表达，则称该信号能被该字典稀疏表达。一般在噪声抑制和信号分离的应用场景中，均使用过完备字典(或称冗余字典)，即字典中包含的原子数目大于信号帧长。

考虑信号序列 $\boldsymbol{s}_i \in R^m, i=1,2,\cdots,n$，其字典应该满足

$$\min_{\boldsymbol{D},\beta} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\| \boldsymbol{D}\beta_i - \boldsymbol{s}_i \|_2^2 + \lambda' \| \beta_i \|_1\right)$$
$$\text{s.t.} \quad \forall j=1,2,\cdots,k, \quad \boldsymbol{d}_j^{\mathrm{T}}\boldsymbol{d}_j \leqslant 1 \qquad (4)$$

其中，$\boldsymbol{D}\in R^{m\times k}$ 为字典，上式中的 λ' 为正则化约束系数，β 为系数矩阵。通过对 β_i 施以 ℓ_1 范数约束可以得到其稀疏解。为了能使用式(4)有效地对大训练样本集进行求解，文献[16,17]提出了一种基于随机梯度下降的字典学习算法。该算法采用交替优化系数与字典的方式进行求解，每一次迭代，首先固定字典 $\boldsymbol{D}$，求解系数 β 的优化问题，然后固定系数 β，进行字典 $\boldsymbol{D}$ 的更新[16,17]，详细步骤如算法 1 所示。

算法 1　基于随机梯度下降的字典学习算法

输入：

信号：$\boldsymbol{s}_i \in R^m$，

正则化参数：λ'，

初始字典：$\boldsymbol{D}_0 \in R^{m\times k}$，

迭代次数：T

输出：

字典矩阵：$\boldsymbol{D}_T \in R^{m\times k}$

步骤：

1) 初始化：$\boldsymbol{A}_0 \leftarrow 0, \boldsymbol{B}_0 \leftarrow 0$；

2) for $t=1$ to T do

3) 固定字典，求解稀疏编码的系数：

$$\beta_t = \underset{\beta\in R^k}{\operatorname{argmin}} \frac{1}{2}\| \boldsymbol{s}_t - \boldsymbol{D}_{t-1}\beta \|_2^2 + \lambda' \| \beta \|_1;$$

4) $\boldsymbol{A}_t \leftarrow \boldsymbol{A}_{t-1} + \frac{1}{2}\beta_t\beta_t^{\mathrm{T}}$；

5) $\boldsymbol{B}_t \leftarrow \boldsymbol{B}_{t-1} + \boldsymbol{s}_t\beta_t^{T}$；

6) 使用算法 2 的字典更新算法更新字典：

$$\boldsymbol{D}_t = \underset{\boldsymbol{D}}{\operatorname{argmin}} \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\| \boldsymbol{s}_i - \boldsymbol{D}\beta_i \|_2^2 + \lambda' \| \beta_i \|_1$$
$$= \underset{\boldsymbol{D}}{\operatorname{argmin}} \frac{1}{t}\left(\operatorname{Tr}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A}_t) - \operatorname{Tr}(\boldsymbol{D}^t\boldsymbol{B}_t)\right)$$
$$\text{s.t.} \quad \forall j=1,2,\cdots,k, \quad \boldsymbol{d}_j^{\mathrm{T}}\boldsymbol{d}_j \leqslant 1$$

7) end for

8) return $\boldsymbol{D}_T$；

算法 2　字典更新算法

输入：

字典：$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_k]\in R^{m\times k}$，

$\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_k]\in R^{k\times k}=\frac{1}{2}\sum_{i=1}^{t}\beta_i\beta_i^{\mathrm{T}}$，

$\boldsymbol{B}=[\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_k]\in R^{m\times k}=\frac{1}{2}\sum_{i=1}^{t}\boldsymbol{s}_i\beta_i^{\mathrm{T}}$

输出：

更新后的字典矩阵：$\boldsymbol{D}\in R^{m\times k}$；

步骤：

1) repeat

2) for $j=1$ to k do

3) 更新字典的第 j 列：

$$\boldsymbol{u}_j \leftarrow \frac{1}{\boldsymbol{A}_{jj}}(\boldsymbol{b}_j - \boldsymbol{D}\boldsymbol{a}_j) + \boldsymbol{d}_j,$$

$$\boldsymbol{d}_j \leftarrow \frac{1}{\max(\| \boldsymbol{u}_j \|_{2,1})}\boldsymbol{u}_j;$$

4) end for

5) until convergence

6) return $\boldsymbol{D}$；

由于目标话音信号和非随机噪声信号适合使用不同的字典进行稀疏表达，通过把目标话音信号字典和噪声字典连接的组合字典，实现每个阵元接收到的含噪信号可以通过混合字典的稀疏表示进行分离，将噪声字典对应的系数置 0 即可实现降噪的目的，因此在降噪前需分别得到话音信号字典和噪声字典。

为了得到目标话音信号和噪声信号的字典，需要分别使用这 2 种信号作为训练样本进行字典学习。对于二元麦克风阵列来说，其噪声相关信号可以表示为

$$y_n(t) = y_{1p}(t) - y_{2p}(t) \qquad (5)$$

式(5)表明 $y_n(t)$ 理论上不含目标信号且与原始噪声信号相关。由于信号数据字典是信号在其特征空间中基向量的集合，信号的衰减或增强不会影响信号字典本身。因而可以使用信号 $y_n(t)$作为原始噪

声信号字典学习的信号得到其字典 D_n。假定在话音增强过程中噪声环境不变，因此噪声字典的学习可以放在话音增强的开始阶段，利用式(5)获得开始一段的相关噪声，并作为训练样本学习得到该噪声字典。

对于目标话音信号，由于不能直接获得其不含噪声干扰的纯净信号，因而需要预先使用语料库进行学习获得其过完备字典 $\boldsymbol{D}_t$。本文使用 GRID[18] 语料库，该语料库提供了 18 个男性和 16 个女性每人 1 000 个句子的语料。训练中选取其中男女各 16 人，对其语料进行训练得到具有一定通用性的话音信号字典。

3.2　话音降噪

由于含噪信号 $y_{1p}(t)$ 和 $y_{2p}(t)$ 中含有共同的目标信号 $x_1(t)$，为了得到这些共同信号，可以通过 ℓ_2/ℓ_1 正则化稀疏回归得到稀疏的系数矩阵，然后将噪声字典对应的系数置 0，最后进行稀疏重构即可实现降噪，如图 1 所示。

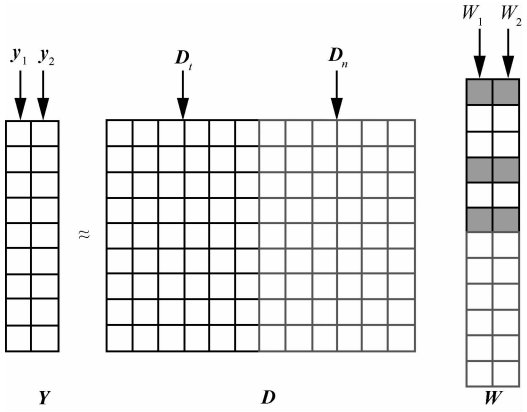


图 1　多任务稀疏表达降噪原理

图 1 中左侧 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 分别为稀疏重构后的 2 个阵元的信号，$\boldsymbol{D}$ 代表混合字典，它由上方左侧左子矩阵的话音字典 $\boldsymbol{D}_t$ 和右子矩阵的噪声字典 $\boldsymbol{D}_n$ 共同组成，右侧 $\boldsymbol{W}$ 是 2 个阵元信号在字典上表达对应的稀疏系数矩阵，其上部对应目标信号系数，下部为噪声信号系数，这些系数是通过 ℓ_2/ℓ_1 正则化约束得到。2 个麦克风所获得的噪声信号由于存在不一致性，故在多任务稀疏模型中噪声信号的系数得以抑制，图 1 中 $\boldsymbol{W}$ 对应框下半部分噪声对应系数较小或者为 0，去噪处理时统一做置 0 处理。

假定从 2 个阵元接收信号中每次取 m 个采样点并且经过相位补偿，然后定义 $\boldsymbol{Y}$ 为一个 m 行 2 列的矩阵，第 1 列和第 2 列分别为信号 $\boldsymbol{y}_{1p}(t)$ 和 $\boldsymbol{y}_{2p}(t)$ 的 m 个采样点，$\boldsymbol{Y}$ 在混合字典上表达对应的系数矩阵 $\boldsymbol{W}$ 满足下式约束

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{DW} - \boldsymbol{Y} \|_2^2 + \lambda \| \boldsymbol{W} \|_{2,1} \tag{6}$$

其中，$\boldsymbol{D} = [\boldsymbol{D}_t, \boldsymbol{D}_n]$，是目标话音字典和噪声字典组成的过完备基向量组成的数据字典矩阵；λ 为正则化系数，其值的大小决定了系数矩阵的稀疏性程度；式(6)中第二项为对系数进行稀疏性约束的 ℓ_2/ℓ_1 混合范数。求得式(6)中的 $\hat{\boldsymbol{W}}$ 最优解后，还需要对 $\hat{\boldsymbol{W}}$ 中对应噪声部分的系数置 0，得到 $\tilde{\boldsymbol{W}}$。此时阵列降噪后的输出信号 $\tilde{\boldsymbol{y}}_{\text{mts}}$ 为

$$\tilde{\boldsymbol{y}}_{\text{mts}} = \boldsymbol{D}\tilde{\boldsymbol{W}} \tag{7}$$

式(6)是以下多任务稀疏表示的一种特殊情况：

$$\underset{\boldsymbol{W}}{\operatorname{argmin}} f(\boldsymbol{W}) + \lambda \| \boldsymbol{W} \|_{2,1} \tag{8}$$

其中，$f(\cdot)$ 是一个光滑的凸代价函数。该问题可使用加速近似梯度算法[9,18,19]进行求解。该算法为迭代算法，每一次迭代中首先不考虑正则化项，使用加速梯度下降使得 $f(\cdot)$ 函数值减小，然后再将加速梯度下降得到的解通过近似算子“投影”到约束的可行域中。加速近似梯度算法的框架如算法 3 所示。

算法 3　加速近似梯度算法

输入：

代价函数：$f(\cdot)$

正则化参数：λ

初始化仿射组合参数：β^0

初始化系数矩阵：$\boldsymbol{W}^0$

收敛阈值：τ

输出：

系数矩阵：$\boldsymbol{W}^*$

步骤：

1) repeat

2) 通过仿射组合计算搜索点：
$\boldsymbol{S}^{(k)} = \boldsymbol{W}^{(k)} + \beta^{(k)}(\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(k-1)})$；

3) 使用自适应步长 $t^{(k)}$ 计算下一个梯度下降点 $\boldsymbol{U}^{(k+1)}$：
$\boldsymbol{U}^{(k+1)} = \boldsymbol{S}^{(k)} - t^{(k)} \nabla f(\boldsymbol{S}^{(k)})$；

4) 使用近似算子计算下一个系数矩阵：$\boldsymbol{W}^{(k+1)}$：
$\boldsymbol{W}^{(k+1)} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{W} - \boldsymbol{U}^{(k+1)} \|_2^2 + t^{(k)} \lambda \| \boldsymbol{W} \|_{2,1}$；

5) 更新 $t^{(k+1)}$ 和 $\beta^{(k+1)}$ 准备下次迭代

6) $k \leftarrow k+1$；

7) until $\|\boldsymbol{W}^{(k+1)}-\boldsymbol{W}^{(k)}\|_2 \leqslant \tau$

8) return $\boldsymbol{W}^{*}=\boldsymbol{W}^{(k+1)}$

对于算法 3 中的 4)，文献[19]给出了一种简便的按行分离的计算方法

$$\bar{\boldsymbol{W}}^{p}=\begin{cases}\left(1-\dfrac{\bar{\lambda}}{\|\bar{\boldsymbol{U}}^{p}\|_2}\right)\bar{\boldsymbol{U}}^{p}, & \|\bar{\boldsymbol{U}}^{p}\|_2>\bar{\lambda}\\ 0, & \|\bar{\boldsymbol{U}}^{p}\|_2\leqslant\bar{\lambda}\end{cases} \tag{9}$$

其中，$\bar{\boldsymbol{W}}=\boldsymbol{W}^{(k+1)}$，$\bar{\boldsymbol{U}}=\boldsymbol{U}^{(k+1)}$，$\bar{\lambda}=\boldsymbol{t}^{(k)}\lambda$，上标 p 表示该矩阵的第 p 行的向量。

3.3　算法复杂度分析

为了衡量本文算法的性能，选择广义旁瓣抵消器(GSC, generalized sidelobe canceller)和基于相干滤波器作为参考对象。由于在二元麦克风小阵列中 3 种算法均需要进行短时傅里叶变换和反变换，因而在比较中可以都不考虑傅里叶变换和反变换复杂度的影响。

本文方法是基于加速近似梯度算法，对于一个二元阵列，每次处理 m 个采样点，则基于 ℓ_2/ℓ_1 正则化约束多任务稀疏表达的算法复杂度是 $O(n(m+2)/\sqrt{\varepsilon})$[19]，其中，$\varepsilon$ 为信号重构误差，n 为字典中原子的数目。而相位补偿的算法复杂度为 $O(m)$，因而本文算法不考虑傅里叶变换与反变换的计算量时，对 m 个采样点的算法复杂度为 $O(m+n(m+2)/\sqrt{\varepsilon})$；GSC 算法复杂度[20]为 $O(4ml+m+3)$，其中，l 为 GSC 中自适应滤波器的长度，因此基于二元麦克风小阵列的 GSC 算法加上时延补偿后的总复杂度为 $O(4ml+2m+3)$；而相干滤波器[7]的算法复杂度为 $O(m)$。

由于本文算法中使用的是过完备字典，即 $n>m$，同时 ε 为很小的正数，因此本文的算法复杂度比另外 2 种算法复杂度高。不过由于在实际处理时，对信号是分帧进行的，只要帧不太长，对于当前主流的处理器运算速度，本文算法基本能满足实时性要求。

4　实验分析

本实验主要验证本文算法在目标话音信号阵列接收信号的开始位置和非开始位置对于非平稳噪声干扰的降噪效果，信号的帧长为 256 点，字典矩阵大小为 256×1 024。实验表明，正则化系数 λ 取值为 0.1 附近时，可以取得较好的去噪效果，故实验中使用 $\lambda=0.1$。每个实验采用开始 2 s 长度的噪声相关信号进行学习得到相应的噪声字典，该相关噪声信号通过式(5)计算得到。另外使用 GSC[1]和基于相干滤波器[7]2 种经典的二元麦克风小阵列话音增强算法作为比较，同时假设这 2 种方法在 VAD 估计时完全准确。

4.1　仿真实验

仿真实验使用阵元间距为 2 cm 的二元麦克风小阵列，干扰噪声信号来自 Noise92[21]噪声库，在实验中使用了 2 个相同的噪声干扰源，用来模拟真实环境多个噪声源情况，信号的采样频率降到 16 KHz；目标声源离阵列 15 cm，图 2 为噪声干扰源、目标信号源、阵列的位置关系图，为了与实际环境相似，图 2 中目标信号位置位于 2 个阵元中心线偏左一点，本实验中偏左 3 mm，处理时认为是在中心线上，以模拟实际目标信号位置估计略有偏差情景。阵列仿真信号使用 Kentucky 大学的 ArrayToolbox 工具箱产生。

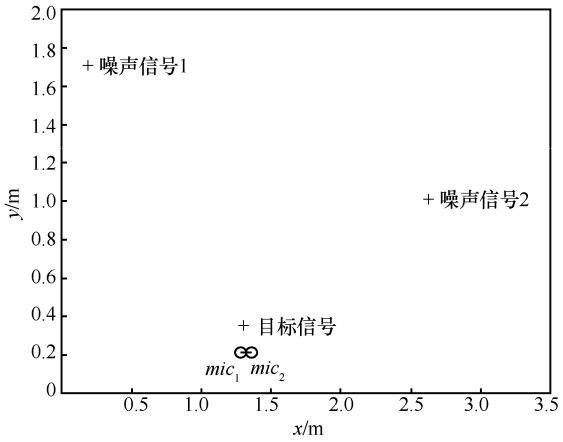


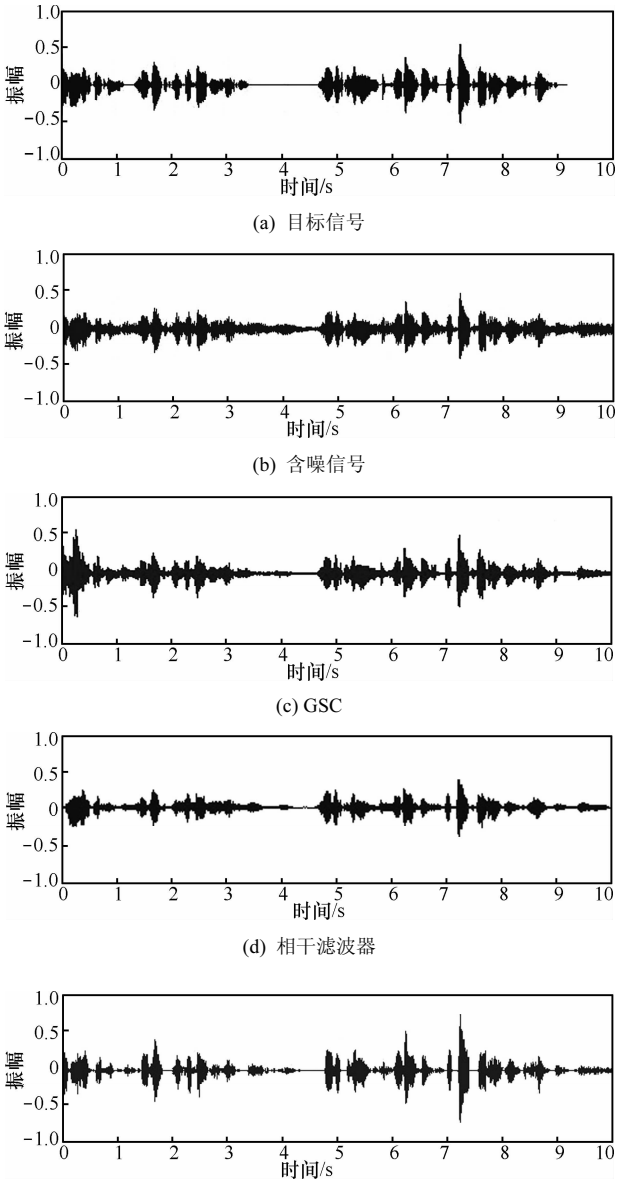
图 2　仿真环境阵列、目标信号、噪声信号位置关系

第 1 个实验的主要目的是验证本文算法对于目标信号处于开始帧位置时的降噪能力，干扰源为多人话音干扰（babble）噪声，图 3(a)为原始话音信号波形图，经过 babble 噪声干扰后第一个阵元接收到的含噪话音信号如图 3(b)所示。从图 3(c)和图 3(d)中可以看出，由于阵列接收的含噪信号开始含有目标信号，GSC 和互相干算法处理后开始阶段都发生了明显的失真，而从图 3(e)本文算法处理后的波形图可以看出，失真与另外 2 种算法相比明显较小，另外与图 3(b)的原始含噪信号相比，噪声已得到明显抑制。

为了对 3 种算法的话音失真情况进行定量分析，需对增强后的话音信号进行话音失真度(speech distortion)[22]计算

$$\upsilon_{\mathrm{sd}}(H)=\frac{\mathrm{E}[x(k)-H^{\mathrm{T}}x(k)]^2}{\sigma_x^2} \tag{10}$$

其中，H 表示话音增强函数，x 表示原始目标信号，k 为采样因子，E(·) 表示数学期望，σ_x 为原始目标信号的均方。υ_{sd} 的数值越小表明失真度也越小。分别对 GSC、相干滤波以及本文算法在第 1 个实验处理结果使用式(10)计算得到：0.031、0.103、0.012，这表明阵列接收到的含噪信号开始帧含有目标信号时，多任务稀疏性约束话音增强算法相对具有较小的失真度。



(a) 目标信号

(b) 含噪信号

(c) GSC

(d) 相干滤波器

(e) 本文算法

图 3　babble 干扰处理前后波形

第 2 个实验的主要目的是验证本文算法在不同信噪比环境下的降噪效果，干扰源为背景音乐，位置关系同第 1 个实验，唯一的区别是 2 个噪声源均为背景音乐。另外为了能与其他方法相比较，实验中阵列接收信号的开始位置不含目标话音信号。比较结果如图 4 所示，其中，信噪比通过式(11)计算。

$$SNR=10\lg(P(x)/P(n)) \tag{11}$$

其中，$P(x)$和 $P(n)$分别为目标信号和噪声信号功率谱密度。输出信噪比中的噪声谱密度采用最小统计[23~25]的方法进行估计，然后使用含噪信号谱密度减去噪声谱密度即为目标信号谱密度，进而利用式(11)计算出每个输出信号的信噪比。

从图 4 可以看出，在不同信噪比条件下，本文提出的话音增强算法信噪比大概能提升 12 dB 左右，与基于相干滤波器方法降噪效果大致差不多，但优于波束形成算法，另外本文算法的一个重要优点是无需使用话音活动检测支持，同时在上面 2 个实验中发现，在帧长为 256 时，本文算法基本能达到实时性要求。

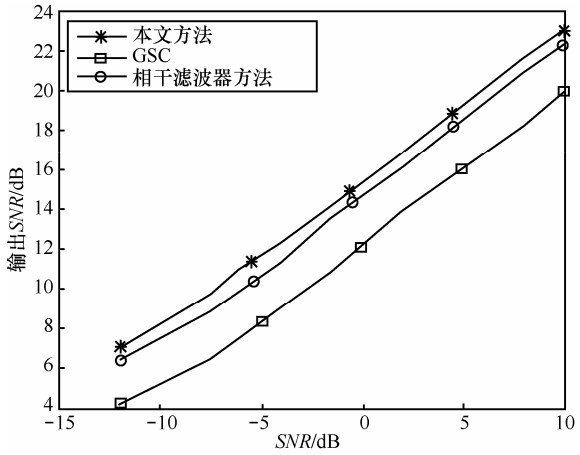


图 4　仿真环境阵列、目标信号、噪声信号位置关系

第 3 个实验主要验证相位补偿误差对本文算法的影响。目标信号与阵列中心距离为 15 cm，且偏左与阵列中心线成 45°角，含噪信号的初始信噪比为 0 dB。相位补偿阵列、信号源位置关系如图 5 所示。

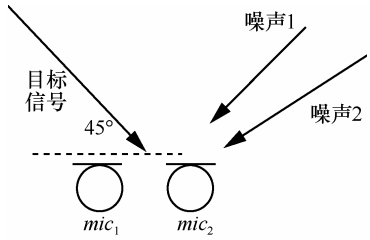


图 5　相位补偿阵列、信号源位置关系

图 5 中噪声 1 和噪声 2 均为音乐噪声，距离阵列中心分别约为 0.8 m 和 1.6 m。此时根据本文第 2 节分析，应该对阵元 2 信号乘以 $e^{-j\omega d\cos\theta/c}$，才能得到准确的相位补偿。

实验中设计估计的目标声源偏离实际声源误差以 2 mm 一个间隔增加，话音增强效果与目标声源偏离误差的关系如图 6 所示。从图 6 可以看出，在本实验条件下，当目标估计声源位置偏离越大时性能也越差，这主要是由于当估计误差越大时，训练用的参考噪声含有目标信号越多，导致学习得到的噪声字典无法较好地分离目标信号与噪声信号，从而对降噪产生一定的影响。

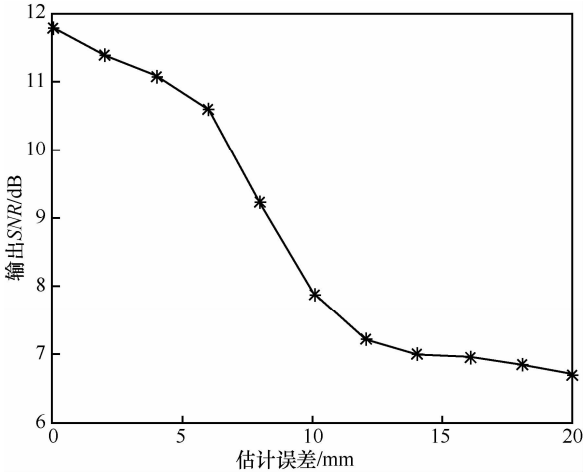


图 6　目标声源估计误差对处理效果的影响

但从图 6 中同时也可以看出，在估计误差较小的情况下，如小于 4 mm，与没有误差的估计性能相差较小。

考虑到本文应用场景是手机或助听器领域，目标信号距离阵列较近而噪声相对较远，此时相位估计相对较准确，因而相位补偿误差不会太大，对降噪不会产生明显影响。另外本实验也说明本文算法中相位补偿的步骤有益于噪声消除。

为了进一步验证本文算法的有效性，使用基于 ITU-T P.862.2 [25] 定义的话音质量感知评价(PESQ, perceptual evaluation of speech quality)标准作为重建话音质量的客观评价。基于 PESQ 标准的算法首先对原始话音信号和含噪话音经话音增强后的信号进行电平调整到标准听觉电平，再用 IRS(intermediate reference system)滤波器进行滤波。对通过电平调整和滤波后的 2 个信号在时间上对准，并进行听觉变换，该变换包括对系统中线性滤波和增益变化的补偿和均衡。2 个听觉变换后的信号之间的谱失真测度作为扰动，分析扰动曲面提取出的退化参数，并在频率和时间上累积起来，映射到对主观 MOS 的预测值。基于 PESQ 标准的算法可以比较待测试话音信号与指定参考信号之间的听觉距离，并提供类似主观平均意见分(MOS, mean opinion score)的 PESQ MOS 话音质量打分，其分值范围在−0.5~4.5 之间，分值越大表示增强后的话音与原始话音越接近。

实验环境同第 1 个实验，使用 GSC、相干滤波器以及本文方法分别对 babble、音乐、汽车、办公室、工厂 5 种背景噪声干扰的目标话音信号进行降噪处理，阵元接收到的含噪信号初始信噪比均为 1 dB 左右，使用基于 PESQ MOS 算法测试时需同时输入原始目标信号。表 1 为本实验环境下不同算法 PESQ MOS 得分情况，从表中可以看出，本文算法 PESQ MOS 评价结果也优于另外 2 种话音增强算法。

4.2　真实数据实验

实验中的二元麦克风阵列采用 2 个全指向性硅微麦克风组成，阵元间距为 1 cm，音频采集卡使用福建泉州恒通数码科技的 DAR-2000 进行信号采集，采样率为 32 KHz；实验环境为一个长、宽、高分别约为 6 m、5 m 和 3 m 的实验室内。目标信号源为真人朗诵且位于阵列的正前方约 15 cm；噪声信号源为位于阵列左前方的音箱，离阵列距离约为 1 m。实验中分别使用 babble、汽车、工厂、音乐以及办公室等作为背景噪声，不同二元麦克风小阵列降噪算法处理的结果如表 2 所示。

表 1　PESQ MOS 得分结果比较

噪声	不同方法的得分		
	GSC	相干滤波器	本文方法
babble	2.7	3.4	3.7
汽车	3.1	3.5	4.0
工厂	2.6	3.3	3.9
音乐	2.6	3.1	4.0
办公室	2.8	3.4	3.8

表 2　不同类型背景噪声处理信噪比比较

噪声	不同信号的信噪比/dB			
	输入信号	GSC	相干滤波器	本文方法
babble	0.87	8.01	11.23	11.80
汽车	7.92	18.38	20.06	21.0
工厂	−8.23	−1.98	4.96	5.51
音乐	−3.62	3.69	5.15	5.46
办公室	3.19	12.31	14.68	15.19

由表 2 可以看出，在实际环境中，无论输入信号的信噪比如何，本文算法明显比 GSC 算法要好，比

相干滤波器略有改善。考虑到实验中的 VAD 是理想状况，实际情况中很难满足，因而无需 VAD 支持的本文降噪算法相对来说更具可靠性。此外实际实验环境是在室内，因而具有一定的混响干扰，本文算法虽然是基于加性噪声干扰的噪声抑制，但由于混响噪声在多个麦克风获取的信号中无一致性，故多任务稀疏表示对于此类乘性噪声也有一定的抑制能力。

5 结束语

本文把基于多任务稀疏性约束的方法引入到二元麦克风小阵列中。首先利用相位补偿使其满足多任务稀疏性学习算法的条件。文中通过语料库的离线字典学习获得通用的话音信号字典，利用噪声参考信号进行实时在线字典学习获得适应于环境噪声的噪声信号字典，进而可以通过 ℓ_2/ℓ_1 范数约束噪声信号的系数，从而达到降噪的目的。与传统的二元麦克风小阵列话音增强算法相比，不但可以克服话音活动检测的限制，而且也不需要假定处理信号初始阶段为非话音段的条件，并具有明显的降噪效果。

参考文献：

[1] GRIFFITHS L, JIM C. An alternative approach to linearly constrained adaptive beamforming[J]. IEEE Transactions on Antennas and Propagation, 1982, 30(1):27-34.

[2] ELKO G W, PONG A N. A simple adaptive first-order differential microphone[A]. Proceedings of IEEE International Conference on Applications of Signal Processing to Audio and Acoustics[C]. New Paltz, NY, USA, 1995.169-172.

[3] BRANDSTEIN M, WARD D. Microphone Arrays: Signal Processing Techniques and Applications[M]. Berlin: Springer Verlag, 2001.

[4] CHENA J, PHUA K, SHUEA L, *et al*. Performance evaluation of adaptive dual microphone system[J]. Speech Communication, 2009, 51(12):1180-1193.

[5] HUANG Y, CHEN J, BENESTY J. Immersive audio schemes[J]. IEEE Signal Processing Magazine, 2011, 28(1):20-32.

[6] ALLEN J B, BERKLEY D A, BLAUERT J. Multimicrophone signal-processing technique to remove room reverberation from speech signals[J]. The Journal of the Acoustical Society of America, 1977, 62(4):912:915.

[7] KALLEL F, GHORBEL M, FRIKHA M, *et al*. A noise cross PSD estimator based on improved minimum statistics method for two-microphone speech enhancement dedicated to a bilateral cochlear implant[J]. Applied Acoustics, 2012, 73(3):256-264.

[8] ARGYRIOU A, EVGENIOU T, PONTIL M. Convex multi-task feature learning[J]. Machine Learning, 2008, 73(3):243-272.

[9] LIU J, JI S, YE J. Multi-task feature learning via efficient l2,1-norm minimization[A]. Proceedings of the Conference on Uncertainty in Artificial Intelligence[C]. Montreal, Canada, 2009. 339-348.

[10] ROMERA PAREDES B, ARGYRIOU A, BIANCHI-BERTHOUZE N, *et al*. Exploiting unrelated tasks in multi-task learning[A]. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics[C]. La Palma, Canary Islands, 2012.951-962.

[11] GEMMEKE J F, CRANEN B. Sparse imputation for noise robust speech recognition using soft masks[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. 2009.4645-4648.

[12] HE Y J, HAN J Q, DENG S W, *et al*. A solution to residual noise in speech denoising with sparse representation[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Kyoto, Japan, 2011.4653-4656.

[13] SIGG C D, DIKK T, BUHMANN J M. Jordan speech enhancement using generative dictionary learning[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(6):1698-1712.

[14] COBOS M, LOPEZ J J, SPORS S. Analysis of room reverberation effects in source localization using small microphone arrays[A]. International Symposium on Communications, Control and Signal Processing[C]. Limassol, Cyprus, 2010.1-4.

[15] BLANDIN C, VINCENT E, OZEROV A. Multi-source TDOA estimation using SNR-based angular spectra[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Prague, Czech Republic, 2011.2616-2619.

[16] BACH F, PONCE J, SAPIRO G. Online learning for matrix factorization and sparse coding[J]. Journal of Machine Learning Research, 2010,2010(11):19-60.

[17] MAIRAL J, BACH F, PONCE J, *et al*. Online dictionary learning for sparse coding[A]. International Conference on Machine Learning[C]. Montreal, Canada, 2009.689-696.

[18] http://www.dcs.shef.ac.uk/spandh/gridcorpus/.

[19] CHEN X. Accelerated gradient method for multi-task sparse learning problem[A]. IEEE International Conference Data Mining[C]. Miami, FL, 2009.746-751.

[20] HERBORDT W, KELLERMANN W. Efficient frequency-domain realization of robust generalized, sidelobe cancellers[A]. IEEE Fourth Workshop on Multimedia Signal Processing[C]. Cannes, France, 2001. 377-382.

[21] http://spib.rice.edu/spib/select_noise.html.

[22] BENESTY J, CHEN J, HUANG Y. Microphone Array Signal Processing[M]. Berlin: Spring-Verlag, 2008.10-11.

[23] MARTIN R. Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(5):504-512.

[24] MARTIN R. Bias compensation methods for minimum statistics noise power spectral density estimation[J]. Signal Processing, 2006, 86(6): 1215-1229.

[25] Wideband Extension to Rec P862 for the Assessment of Wideband Telephone Networks and Speech Codecs[R]. Intl Telecom Union, 2007.

作者简介：

杨立春（1975-），男，安徽舒城人，浙江大学博士生，浙江万里学院讲师，主要研究方向为话音信号处理。

叶敏超（1987-），男，浙江杭州人，浙江大学博士生，主要研究方向为人工智能、图像处理、信号处理。

钱沄涛（1968-），男，浙江杭州人，浙江大学教授、博士生导师，主要研究方向为智能信息处理、机器学习、模式识别。