基于卷积非负矩阵部分联合分解的强噪声单声道语音分离

董兴磊 [1] 胡 英 [1] 黄 浩 [1, 2] 吾守尔 · 斯拉木 [1, 2]

摘 要 非负矩阵部分联合分解 (Nonnegative matrix partial co-factorization, NMPCF) 将指定源频谱作为边信息参与混合信号频谱的联合分解, 以帮助确定指定源的基向量进而提高信号分离性能. 卷积非负矩阵分解 (Convolutive nonnegative matrix factorization, CNMF) 采用卷积基分解的方法进行矩阵分解, 在单声道语音分离方面取得较好的效果. 为了实现强噪声条件下的语音分离, 本文结合以上两种算法的优势, 提出一种基于卷积非负矩阵部分联合分解 (Convolutive nonnegative partial matrix co-factorization, CNMPCF) 的单声道语音分离算法. 本算法首先通过基音检测算法得到混合信号的语音起始点, 再据此确定混合信号中的纯噪声段, 最后将混合信号频谱和噪声频谱进行卷积非负矩阵部分联合分解, 得到语音基矩阵, 进而得到分离的语音频谱和时域信号. 实验中, 混合语音信噪比 (Signal noise ratio, SNR) 选择以 −3 dB 为间隔从 0 dB 至 −12 dB 共 5 种 SNR. 实验结果表明, 在不同噪声类型和噪声强度条件下, 本文提出的 CNMPCF 方法相比于以上两种方法均有不同程度的提高.

关键词 卷积非负矩阵分解, 非负矩阵部分联合分解, 语音分离, 强噪声, 单声道

引用格式 董兴磊, 胡英, 黄浩, 吾守尔 · 斯拉木. 基于卷积非负矩阵部分联合分解的强噪声单声道语音分离. 自动化学报, 2020, **46**(6): 1200−1209

DOI 10.16383/j.aas.c180065

Monaural Speech Separation by Means of Convolutive Nonnegative Matrix Partial Co-factorization in Low SNR Condition

DONG Xing-Lei[1] HU Ying[1] HUANG Hao[1, 2] SILAMU Wushour[1, 2]

Abstract Nonnegative matrix partial co-factorization (NMPCF) is a joint matrix decomposition algorithm integrating prior knowledge of specific source to help separate specific source signal from monaural mixtures. Convolutive nonnegative matrix factorization (CNMF), which introduces the concept of a convolutive non-negative basis set during NMF process, opens up an interesting avenue of research in the field of monaural sound separation. On the basis of the above two algorithms, we propose a speech separation algorithm named as convolutive nonnegative matrix partial co-factorization (CNMPCF) for low signal noise ratio (SNR) monaural speech. Firstly, through a voice detection process exploring fundamental frequency estimation algorithm, we divide a mixture signal into vocal and nonvocal parts, thus those vocal parts are used as test mixture signal while the nonvocal parts (pure noise) participat in the partial joint decomposition. After CNMPCF, we can obtain the separated speech spectrogram. Then, the separated speech signal can reconstructed through Inverse short time fourier transformation. In the experiments, we select 5 SNRs from 0 dB to −12 dB at −3 dB intervals to obtain low SNR mixture speeches. The results demonstrate that the proposed CNMPCF approach has superiority over sparse convolutive nonnegative matrix factorization (SCNMF) and NMPCF under different noise types and noise intensities.

Key words Convolutive nonnegative matrix factorization (CNMF), nonnegative matrix partial co-factorization (NMPCF), speech separation, strong noise, monaural speech

Citation Dong Xing-Lei, Hu Ying, Huang Hao, Silamu Wushour. Monaural speech separation by means of convolutive nonnegative matrix partial co-factorization in low SNR condition. *Acta Automatica Sinica*, 2020, **46**(6): 1200−1209

语音分离是语音信号处理的重要分支, 旨在从被干扰的混合信号中分离出纯净的语音信号, 以提升语音信号的可懂度. 近年来, 语音分离算法不断地

收稿日期 2018-01-26 录用日期 2018-07-15
Manuscript received January 26, 2018; accepted July 15, 2018
国家自然科学基金 (61761041, 61663044), 国家自然科学基金青年基金 (61603323), 新疆维吾尔自治区自然科学基金 (2016D01C061), 新疆大学自然科学基金 (BS160239), 新疆自治区高校科研计划项目 (XJEDU2017T002) 资助
Supported by National Natural Science Foundation of China (61761041, 61663044), National Natural Science Foundation of Youth Foundation of China (61603323), Natural Science Grant of Xinjiang Uygur Autonomous Region (2016D01C061), Natural Science Grant of Xinjiang University (BS160239), and University Scientific Research Project of Xinjiang Uygur Autonomous Region (XJEDU2017T002)
本文责任编委 党建武
Recommended by Associate Editor DANG Jian-Wu
1. 新疆大学信息科学与工程学院 乌鲁木齐 830046 2. 新疆大学多语种信息技术实验室 乌鲁木齐 830046
1. Department of Information Science and Engineering, Xinjiang University, Urmuqi 830046 2. Laboratory of Multi-lingual Information Technology, Xinjiang University, Urumqi 830046

成熟发展, 单声道语音分离作为一个难点, 成为学者们聚焦的一个研究热点[1−2].

语音分离过程能够很自然地表达为一个有监督学习问题. 语音分离系统通过有监督学习算法, 学习一个从混合语音的频谱到纯净语音频谱的映射函数[3] 以实现语音分离. 近年来有监督语音分离技术取得了重要的研究进展, 其中主流的有监督学习算法包括基于非负矩阵分解[4]、基于计算听觉场景分析[5] 和基于深度神经网络[2, 6−7] 的语音分离算法.

基于非负矩阵分解的算法在语音分离领域引起广泛的研究. Smaragdis[8] 提出一种卷积非负矩阵分解算法, 该算法采用一系列语音基矩阵集进行语音频谱的矩阵分解, 这些基矩阵集不仅能描述频谱沿着频率变化的情况, 还能描述频谱沿着时间变化的情况. 因此卷积非负矩阵分解 (Convolutive nonnegative matrix factorization, CNMF) 在指定说话人条件下的单声道语音分离得到了较好的效果. O′Grady 等[9] 在此基础上加入稀疏约束, 提出一种稀疏卷积非负矩阵分解 (Sparse convolutive nonnegative matrix factorization, SCNMF) 的算法, 对语音基矩阵集增加稀疏约束又进一步提高了语音分离算法的性能. 由于噪声信号在时域上不具备时变特性, 据此, Sun 等[10] 提出一种基于块卷积的稀疏低秩模型的单声道语音增强算法以去除混合语音中的噪声成份得到增强的语音信号. 总的来说, 非负矩阵分解 (Nonnegative matrix factorization, NMF)[4] 在语音分离应用上有了长足的发展, 当带噪语音信号中的信号源是统计独立时, 强加一些约束条件后, NMF 算法对于源分离便是有效且鲁棒的. 然而, 在没有任何关于指定信号源的先验知识情况下, 标准 NMF 算法缺少对指定源的分离能力. 为了解决这个问题, 非负矩阵部分联合分解 (Nonnegative matrix partial co-factorization, NMPCF)[11−13] 应运而出. Hu 等[14] 将该算法应用到唱声分离和歌手识别中, 提出一种基于 NMPCF 的唱声分离的歌手识别方法, 该方法将检测出的纯伴奏片段频谱作为边信息参与混合唱声频谱的联合分解, 在分解过程中, 混合频谱与纯伴奏频谱只共用伴奏基矩阵, 混合频谱与干净唱声频谱只共享唱声基矩阵, 因此称为部分联合分解。相比基于其他基于 NMF 的分离算法, 部分联合分解算法分离的唱声性能有了很大提高.

基于上述分析, 结合 CNMF 和 NMPCF 算法的优势, 本文提出一种卷积非负矩阵部分联合分解算法用于强噪声条件下的单声道语音分离. 本文的组织结构如下: 第 1 节介绍稀疏约束卷积非负矩阵分解算法; 第 2 节描述非负矩阵部分联合分解算法; 第 3 节针对提出的分离算法进行量化评估; 第 4 节给出总结与讨论.

1　稀疏约束卷积非负矩阵分解

1.1　卷积非负矩阵分解 CNMF

假设语音信号 $s(t)$ 和加性噪声信号 $n(t)$ 独立不相关, 带噪语音信号 $v(t)$ 经过短时傅里叶变换得到带噪语音频谱 V, 可以表示为 $V = S + N$. 其中 S, N 分别表示干净语音频谱和噪声频谱.

NMF 算法 [4] 将矩阵 $V \in \mathbf{R}_+^{M\times N}$ 分解成两个矩阵 $W \in \mathbf{R}_+^{M\times R}$ 和 $H \in \mathbf{R}_+^{R\times N}$, 使得 $V = WH$. M 是频率单元的个数, R 是基向量的个数, N 是时间帧数, W 是基矩阵, H 是系数矩阵 (也称激活矩阵), 矩阵中的每一个元素都被限定为非负. W 中基向量能描述出频谱沿频率方向的特性, 不能描述出频谱沿着时间方向的特性. Smaragdis[8] 提出卷积非负矩阵分解 (CNMF) 算法, 使用一组二维基向量集 $\{W(t), t \in [0, T-1]\}$ 及其系数矩阵 H 对应元素之间卷积求和运算的结果来表示待分解矩阵 V. CNMF 的数学模型表示为

$$V = \sum_{t=0}^{T-1} W(t) \cdot \overset{t\rightarrow}{H} \tag{1}$$

其中, $W(t) \in \mathbf{R}_+^{M\times R}$ 是基矩阵集, 包含 T 个基矩阵, 其中 $t \in [0, T-1]$, H 是系数矩阵, $\overset{t\rightarrow}{(\cdot)}$ ($\overset{\leftarrow t}{(\cdot)}$) 运算符表示右 (左) 移 t 列, 同时将左 (右) 边空出的 t 列补零. 采用 K-L 散度作为目标函数

$$D = \|V \odot \ln \frac{V}{\hat{V}} - V + \hat{V}\|_{\mathrm{F}} \tag{2}$$

其中, $\hat{V}$ 是 V 的估计. 通过优化目标函数得到模型参数的乘性更新公式

$$H = H \odot \frac{W(t)^{\mathrm{T}} \cdot \overset{\leftarrow t}{[\frac{V}{\hat{V}}]}}{W(t)^{\mathrm{T}} \cdot \mathbf{1}} \tag{3}$$

$$W(t) = W(t) \odot \frac{\frac{V}{\hat{V}} \cdot \overset{t\rightarrow}{H}^{\mathrm{T}}}{\mathbf{1} \cdot \overset{t\rightarrow}{H}^{\mathrm{T}}} \tag{4}$$

其中, $\odot$ 表示矩阵元素相乘, $\mathbf{1}$ 表示全 1 矩阵. 在每一次更新迭代步骤中, 首先, 在保持 H 不变的情况下更新所有的 $W(t)$, 然后, 再分别用 T 个 $W(t)$ 去更新 H, 最终的 H 是这 T 个更新结果的平均值.

将基矩阵集中不同时刻的同一基向量按时间顺序连接后可以看出语音在时频域的声学特征. 以基向量个数 $R = 40$, 时间跨度 $T = 8$ 为例, 图 1 集中显示了 40 个从干净语音频谱中提取出的基向量如上

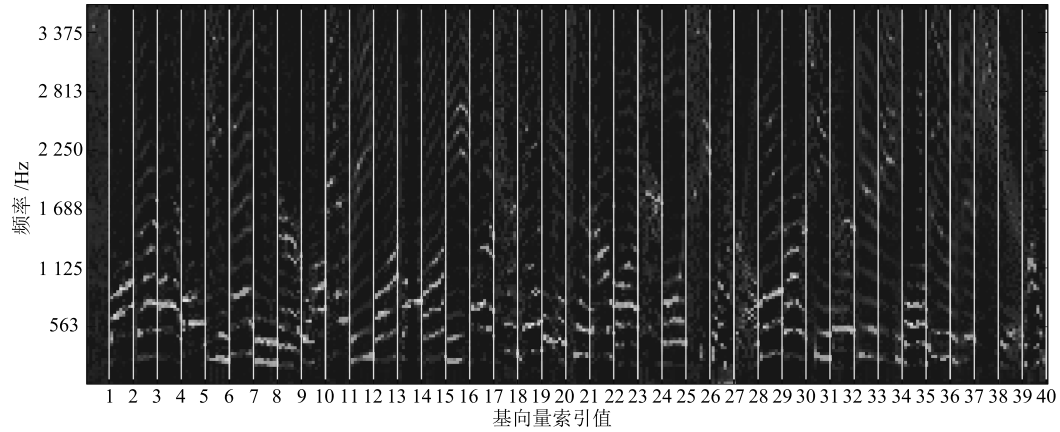


图 1 干净语音频谱经过 CNMF 分解后提取出的基向量

Fig. 1 The basis extracted from the clean speech spectrum after CNMF decomposition

所述连接后的结果, 可以看出每个基向量都能表示语音片段的一部分, 有的基向量与音素频谱一致, 有清晰的谐波成份且有音高的变化波形; 有的基向量与清音频谱一致, 在整个频域范围内都有能量.

1.2 稀疏卷积的 CNMF

采用 CNMF 分解语音信号幅度谱矩阵 V 得到的基是趋于稀疏的, 如果在系数矩阵 H 上增加稀疏性约束, 则可以得到稀疏的基向量, 而且重构误差更小, 这样可以使分离后的语音更加清晰可懂[15].

CNMF 对 H 增加 L_0 稀疏约束可得基于 L_0 的目标函数

$$D = \|V \odot \ln \frac{V}{\hat{V}} - V + \hat{V}\|_{\mathrm{F}} + \lambda \|H\|_0 \quad (5)$$

其中, $\lambda \in \mathbf{R}_+$ 为正则化参数, 用于控制稀疏程度. L_0 正则化会限制 H 中每一列非 0 元素的个数. L_0 是非凸的 NP-hard 问题[16]. Candés 等[17] 使用矩阵的 L_1 范数替代 L_0 范数, 将式 (5) 转化为如下的凸优化问题来近似求解.

$$D = \|V \odot \ln \frac{V}{\hat{V}} - V + \hat{V}\|_{\mathrm{F}} + \lambda \|H\|_1 \quad (6)$$

2 非负矩阵部分联合分解

NMPCF 算法[11−13] 利用额外的指定源信号频谱作为边信息参与矩阵联合分解, 能自动的区分出指定源信号和其他源的基向量. Hu 等[14] 在唱声分离中基于此联合分解算法做一扩展, 将混合信号中的两种源信号频谱作为边信息参与矩阵联合分解. 混合信号频谱矩阵 V 分解模型为

$$V = W_s H_s + W_n H_n \quad (7)$$

其中, W_s 和 H_s 分别是唱声的基矩阵和系数矩阵, 而 W_n 和 H_n 分别是纯伴奏的基矩阵和系数矩阵. 同样, 所有的矩阵都被约束为非负. 在此基础上, 再强加两个额外的先验频谱矩阵参与联合分解.

算法首先通过唱声检测过程, 将每段歌曲划分为两部分: 有唱声片断和无唱声片断 (纯伴奏片断). 唱声片断频谱 V 作为待分离混合信号频谱, 无唱声片段频谱 N 和干净唱声频谱 S 作为先验信息 (或称为边信息) 同时参与联合分解. 即给定三个输入矩阵 V, N, S, 在近似目标矩阵分解式 (7) 的同时做两个边信息矩阵分解: $S = W_s U_s$ 和 $N = W_n U_n$. 在联合分解中, V 和 S 的矩阵分解只共享基矩阵集 W_s, V 和 N 的矩阵分解只共享基矩阵 W_n. 采用 Euclidean 距离作为目标函数

$$D = \frac{1}{2}\|V - W_s H_s - W_n H_n\|_{\mathrm{F}}^2 + \frac{\lambda_s}{2}\|S - W_s U_s\|_{\mathrm{F}}^2 + \frac{\lambda_{\mathrm{n}}}{2}\|N - W_n U_n\|_{\mathrm{F}}^2 \quad (8)$$

其中, λ_s 是反映唱声频谱在联合分解中相对重要性的参数, λ_n 是反映伴奏频谱在联合分解中相对重要性的参数.

2.1 卷积非负矩阵部分联合分解

结合非负矩阵部分联合分解 (NMPCF) 和卷积非负矩阵分解 (CNMF) 的优势, 本文提出一种基于卷积非负矩阵部分联合分解 (Convolutive nonnegative partial matrix co-factorization, CNMPCF) 的算法, 用于强噪声条件下的单通道语音分离. 由于提取了混合信号中的纯噪声频谱参与到部分联合分解中, 因此, 本算法在一定程度上能减少噪声类型和噪声强度对语音分离性能的影响. 算法模型示意图如图 2 所示.

首先, 与 Kim 等[13] 的方法一致, 通过 Praat[18] 的基音检测算法检测混合语音信号的基音频率以确定语音片段的起始点, 据此将混合语音信号分割为

语音片断和无语音片断 (不同强度的纯噪声片断). 语音片断和噪声片断频谱分别用矩阵 V 和 N 表示. 混合信号频谱矩阵 V 分解模型为

$$V=\sum_{t=0}^{T-1}W_s(t)H_s+W_nH_n \tag{9}$$

考虑到混合信号中的语音成分具有时频域的相关性, 而噪声则只具有频域上的相关性, 因此, 在分解模型中, 语音片段采用 CNMF 分解, 即语音基向量包含在一系列基矩阵集中, 而噪声片段采用标准 NMF 分解, 噪声基向量包含于一个基矩阵中.

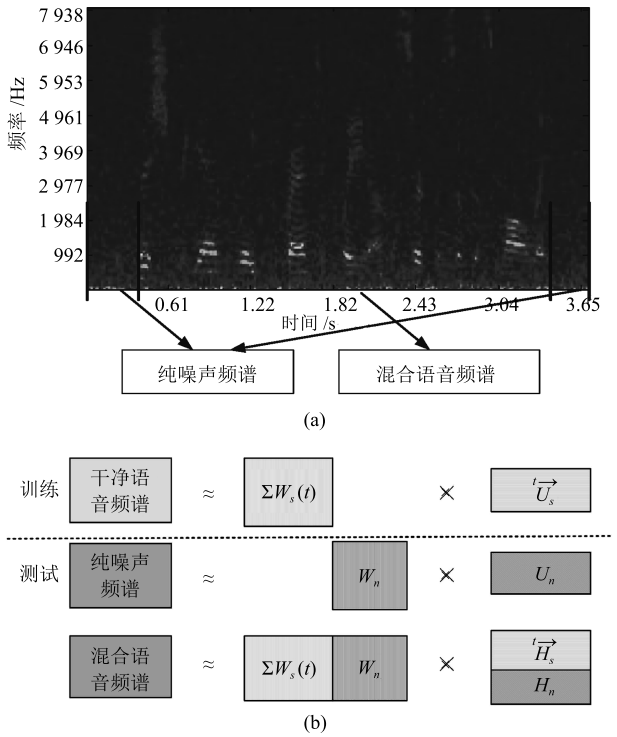


图 2　CNMPCF 算法的频谱分解示意图

Fig. 2　The illustration of magnitude spectrogram by CNMPCF

实验中采用的 TIMIT 数据通常是一条 10 s 左右的语音, 分割出的无语音片断通常位于带噪语音的首端与尾端. 带噪语音频谱 V、纯噪声段频谱 N 以及干净语音频谱 S 参与联合分解, 混合信号频谱 V 和干净语音频谱 S 在联合分解过程中共享语音基矩阵集 $\{W_s(t),t\in[0,T-1]\}$, 同时 V 和纯噪声频谱 N 共享噪声基矩阵 $W_n\in\mathbf{R}_+^{F\times NR}$.

强噪声条件下目标函数使用 K-L 散度相较于 Euclidean 距离更敏感[10], 本文要研究的是在强噪声条件 (低信噪比) 下的语音分离, 因此采用扩展 K-L 散度作为目标函数

$$\begin{aligned}D=&\lambda_v\mathrm{KLD}\left(V|\sum_{t=0}^{T-1}W_s\overrightarrow{H_s^t}+W_nH_n\right)+\\&\lambda_s\mathrm{KLD}\left(S|\sum_{t=0}^{T-1}W_s\overrightarrow{U_s^t}\right)+\\&\lambda_n\mathrm{KLD}(N|W_nU_n)+\lambda\|H_s\|_1\end{aligned} \tag{10}$$

其中, λ_v, λ_n 和 λ_s 分别是反映带噪语音频谱, 干净语音频谱和纯噪声频谱在联合分解中相对重要性的参数. 考虑到算法的泛化性, 样本应覆盖尽可能多的说话人. TIMIT 语音库包含 168 个说话人, 纯净语音样本数量过大, 这会大大降低联合分解的速度. 实验中, 选择事先采用 CNMF 算法训练干净语音样本, 得到语音基矩阵集 $\{W_s(t),t\in[0,T-1]\}$. 在测试阶段, 仅有 V 和 N 做联合分解. 在分解过程中, 语音基矩阵集 $\{W_s(t),t\in[0,T-1]\}$ 保持不变, 仅更新 H_s,W_n 和 H_n 即可. 相应的更新迭代公式为

$$W_n=W_n\odot\frac{\lambda_v\cdot\frac{V}{\hat{V}}\cdot H_n^{\mathrm{T}}+\lambda_n\cdot\frac{N}{\hat{N}}\cdot U_n^{\mathrm{T}}}{\lambda_n\cdot\mathbf{1}\cdot H_n^{\mathrm{T}}+\lambda_n\cdot\mathbf{1}\cdot U_n^{\mathrm{T}}} \tag{11}$$

$$H_s=H_s\odot\frac{W_s^{\mathrm{T}}(t)\cdot[\overleftarrow{\frac{V}{\hat{V}}}^t]}{W_s^{\mathrm{T}}(t)\cdot\mathbf{1}} \tag{12}$$

$$H_n=H_n\odot\frac{W_n^{\mathrm{T}}\cdot\frac{V}{\hat{V}}}{W_n^{\mathrm{T}}\cdot\mathbf{1}} \tag{13}$$

上述更新公式中, 式 (13) 中 H_s 是 T 个 $W_s(t)$ 基矩阵更新后的平均值.

利用事先训练得到的纯净语音的基矩阵集, 对带噪语音和纯噪声幅度谱矩阵进行联合分解, 分别得到语音的系数矩阵、噪声的基矩阵和系数矩阵. 最后, 采用式 (14) 维纳滤波方法得到分离语音的幅度谱, 进而利用带噪语音的频谱相位通过短时傅里叶变换的反变换 ISTFT 和重叠相加法重构出分离的时域语音信号.

$$\hat{S}=\frac{\sum_{t=0}^{T-1}W_s(t)\overrightarrow{H_s^t}}{\sum_{t=0}^{T-1}W_s(t)\overrightarrow{H_s^t}+W_nH_n}\odot V \tag{14}$$

3　性能评估和比较

本小节通过一系列实验评估语音分离算法的性能. 首先采用 3 种指标将本文所提出的 CNMPCF 算法同 NMPCF 算法和 SCNMF 算法分别作一比较, 然后再与本算法获得最佳性能的结果进行比较. 此外, 采用对比听音方式邀请 20 名大学生对上述 3 种算法分离出的语音进行打分作为主观评价指标.

3.1 实验数据及设置

实验中纯净语音选自 TIMIT 标准语音库, 该语音库共有 168 个说话人, 每个说话人有 10 条语音, 共 1 680 条语音. 随机选取每个说话人中的 1 条语音, 共 168 条语音作为训练样本. 噪声样本选取 Noisex-92 标准噪声库中 4 种典型的噪声: Pink 噪声、Babble 噪声、M109 噪声和 F16 噪声. 为了同噪声频率相匹配, 将纯净语音样本下采样到 8 kHz. 测试样本是由 1 680 条干净语音同 4 种噪声按 5 种不同信噪比进行混合, 信噪比以 −3 dB 为间隔从 0 dB 到 −12 dB 选取. 最终得到 33 600 条带噪语音作为测试样本. 采用应用广泛、技术成熟的 Praat[18] 基音检测算法检测语音信号的基音频率, 没有检测到语音基音频率的帧视为噪声. 由于语音的基音频率一般在 70 ∼ 500 Hz, 人为设定 Praat 方法检测频率上限为 500 Hz, 下限为 70 Hz. 采用 Hamming 窗计算幅度谱, 窗长为 32 ms, 帧移为 16 ms. 噪声基向量 NR 和语音基向量 R 个数分别设为 50 和 100, 时间跨度 $T = 5$. 反映频谱相对重要性参数的 λ_n, λ_v 分别设为 1, 稀疏因子 λ 为 0.01, 迭代次数为 200.

3.2 对比方法及评价指标

实验中, 将本文算法与 NMPCF[13] 和 SCNMF[9] 算法进行对比. 3 种方法均采用相同的实验样本. NMPCF 方法中参数与 CNMPCF 保持一致. SCNMF 方法中语音基矩阵集 $W(t)$ 的 R 设为 100, 时间跨度 $T = 5$, 采用 Hamming 窗, 窗长为 32 ms, 帧移为 16 ms.

同时, 为了对比本文算法的综合性能, 针对纯净语音进行语音边界检测, 获得的语音边界检测结果视为本算法能得到的最准确的检测结果, 据此语音边界检测结果确定的噪声段作为边信息参与 CNMPCF 分解过程, 可以认为这样得到分离语音的结果是本文所提分离算法能达到的最好结果, 在后续的论述中, 此过程的结果称为参考结果 (Reference). 采用语音质量客观估计方法 (Perceptual evaluation of speech quality, PESQ)[19]、BSS-EVAL 体系[20] 的信号失真比 (Source to distortion ratio, SDR) 以及信噪比 (Signal noise ratio, SNR) 增益 ΔSNR[21] 分别评估分离语音的质量和分离算法的实际性能. 其中 PESQ 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 ∼ 4.5, 得分越高说明算法分离效果越好. BSS-EVAL 是目前公认性能较好的源分离算法评估体系. 该评估体系中源失真比 SDR 是目前信号分离领域应用广泛且有效的评价指标, SDR 度量的是估计的源信号与失真的比, 失真包括分离误差、不同源之间的干扰和人工合成导致的误差. 信噪比增益 ΔSNR 是分离信号的信噪比与混合信号的信噪比差值, 是指定源信号的信噪比的增益, 定义为[21]

$$\mathrm{SNR}_{\mathrm{est}} = 10\lg\frac{\sum\limits_t s^2[t]}{\sum\limits_t (s[t]-\hat{s}[t])^2} \tag{15}$$

$$\mathrm{SNR}_{\mathrm{mix}} = 10\lg\frac{\sum\limits_t s^2[t]}{\sum\limits_t (s[t]-x[t])^2} \tag{16}$$

其中, $s[t]$ 是干净语音, $\hat{s}[t]$ 是估计语音, $x[t]$ 是混合信号. 信噪比增益则为

$$\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{est}} - \mathrm{SNR}_{\mathrm{mix}} \tag{17}$$

主观听音测试依据清晰度和失真度以及是否产生额外的干扰成份为标准进行打分, 得分取值范围为 1 ∼ 5, 得分越高说明分离语音的质量越好, 评分标准如下: 声音清晰, 音色饱满, 没有干扰判定为 5 分; 声音比较清晰, 有干扰, 但不易察觉判定为 4 分; 声音一般清晰, 干扰可察觉, 但影响不大判定为 3 分; 声音清晰度明显变差, 干扰增加较多, 影响收听判定为 2 分, 声音不清晰, 并有严重干扰, 无法收听判定为 1 分. 主观听音测试随机选取 25 条分离语音作为测试样本, 遵循双盲方式.

3.3 语音边界检测

本算法通过检测语音的起始点来确定噪声片段. 参与联合分解的噪声频谱矩阵越纯粹, 则后续联合分解的结果越理想, 若因为误检测使噪声频谱矩阵中含有语音片段, 则对后续联合分解性能会有一定的影响, 从一定程度上来讲噪声片段检测的准确性决定了语音分离算法的性能. 图 3 显示纯净语音与带噪语音的起始点和终止点的检测结果. 图 3 (a) 显示一段干净语音边界检测结果, 竖线指示语音段的起点 (上界) 和终点 (下界). 图 3 (b) 和图 3 (c) 显示了混合语音的两种 VAD 检测上、下界结果. 相邻两段语音中间是纯噪声片段, 为了保证参与联合分解的噪声片段的纯净, 检测的上界应位于语音起始点之前以及检测的下界应位于语音终止点之后, 即语音起点的负偏差和终点正偏差都不会影响噪声片段的纯粹度, 如图 3 (c) 所示. 因为这样确定的噪声段没有引入语音成分. 而图 3 (b) 检测的语音上界有正偏差, 下界有负偏差, 这样确定的噪声段引入语音成分, 则会影响后续联合分解的准确度, 从而影响整个算法的分离性能.

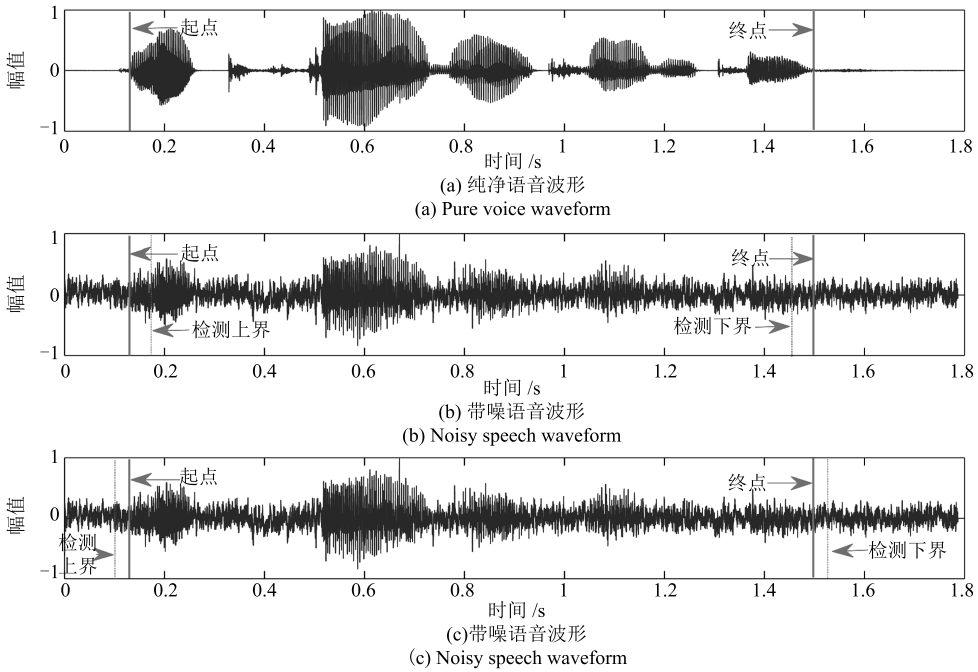


图 3　语音起点、终点 (边界) 检测示意图

Fig. 3　The illustration of start end points (boundary) detection of a speech

实验中选择两种语音边界检测方法来确定语音片段和纯噪声片段. 第 1 种方法是采用应用广泛的 Praat 基音检测算法 (后续用 VAD1 表示) 检测混合语音信号的基音频率, 语音片段为连续的非零基音频率段, 而噪声片段的基音频率为零. 基音频率连续为零的帧数大于 0.2 s 的片段视为噪声片段. 第 2 种方法是一种具有较强鲁棒性的声音激活检测算法[22] (后续用 VAD2 表示) 来确定语音的起点和终点.

选择 SNR 为 −12 dB 且被 Pink 噪声污染的混合语音为 VAD 测试样本, 图 4 显示 1 680 个样本语音边界上界和下界的检测结果的概率分布. 以人工标注语音上、下界作为标准, 显示两种 VAD 检测算法得到的上、下边界的偏差值的概率分布, 图中横坐标表示检测的上、下边界以帧为单位的绝对偏差, 选择显示绝对偏差在 $[-100, 100]$ 之间的概率分布. 虚竖线指示偏差值为 0.

从图 4 可以看出, 与 VAD2 算法相比, VAD1 算法检测的语音上界更多的分布在标准语音起始点之前, 且 VAD1 算法检测的语音下界更多的分布在标准语音起始点之后, 因此本文选择 Praat 基音检测算法作为语音边界检测算法以确定噪声段.

3.4　实验结果及分析

本小节将通过 PESQ、SDR、ΔSNR 和主观听音得分 4 种指标来度量分离性能. 每种指标都将展示对比的 3 种算法和本文算法在最优边界检测结果条件下, 分别在 4 种不同噪声、5 种不同信噪比下的平均结果.

从图 5 的 PESQ 结果来看, 本文提出的 CNMPCF 算法在 4 种不同噪声、5 种不同信噪比下, 都略优于 NMPCF 和 SCNMF. 在 4 种不同的噪声类型下, CNMPCF 算法的 PESQ 测量值相比于 NMPCF 算法平均高出约 0.1, 相比于 SCNMF 算法平均高出 0.26, 参考结果 (Reference) 的 PESQ 测量值相比于 CNMPCF 平均高 0.05. 随着信噪比的下降, CNMPCF 算法的分离性能反而愈发明显. 这是由于 CNMPCF 提取的纯噪声片段加入了联合分解, 因此本算法对噪声类型和强度不敏感, 即使强噪声条件下也能获得较好的分离结果.

图 6 显示 CNMPCF 算法相比于 NMPCF 和 SCNMF 均有较高的 SDR 值. CNMPCF 算法的 SDR 值相比于 SCNMF 算法平均提升了约 5 dB, 相比于 NMPCF 算法平均提升了约 3 dB. 参考结果 (Reference) 的 SDR 值相比于 CNMPCF 算法平均高出 1.4 dB 左右, 这表明 CNMPCF 算法在失真度较小的情况下, 仍能保证较好的语音质量. 强噪声下 CNMPCF 分离性能提升较为明显.

由图 7 显示的信噪比增益 ΔSNR 测量值看出, 随着混合语音信号信噪比的下降, NMPCF 和 SCNMF 两种分离算法的 ΔSNR 值也随之减小, 而 CNMPCF 算法在低信噪比时的 ΔSNR 值也能保持与信噪比较高的 ΔSNR 大略相当. 在 5 种信噪比条件下, CNMPCF 算法的 ΔSNR 测量值相比于

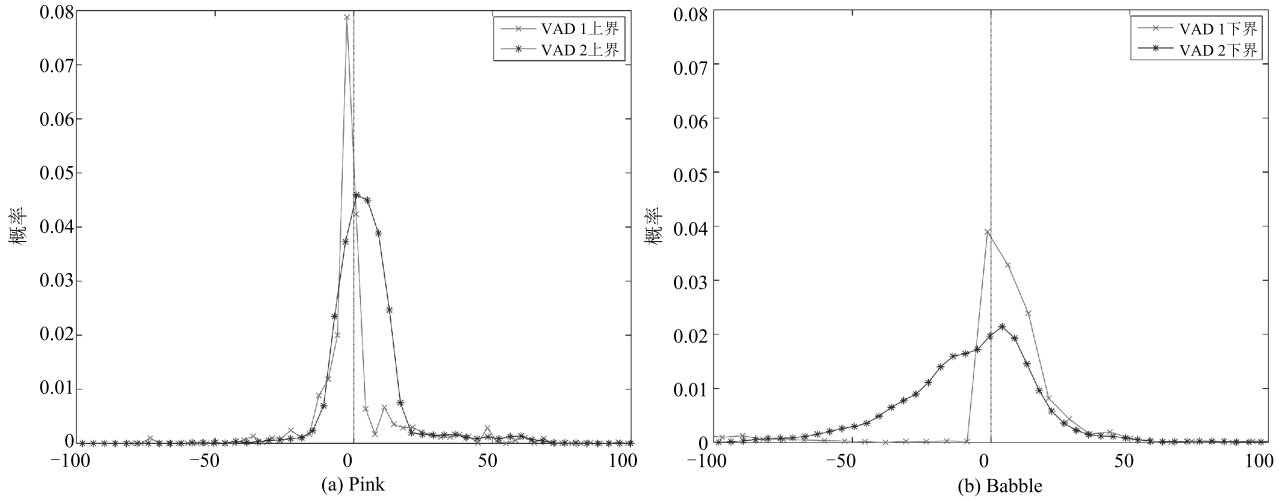


图 4 −12 dB 混合信号的语音上界、下界检测偏差概率分布

Fig. 4 The probability distribution of detection deviation of upper and lower bounds in −12 dB mixture speech

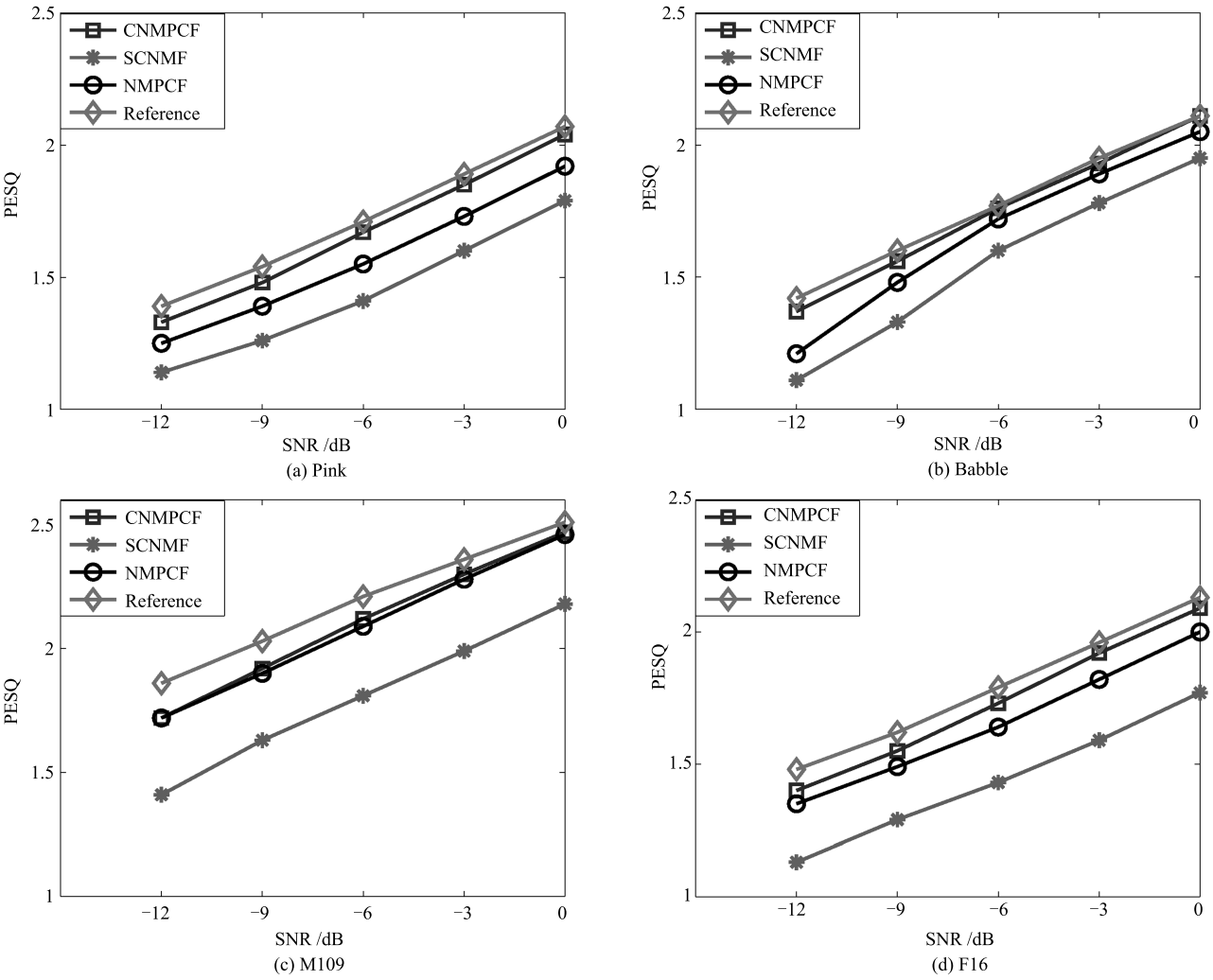


图 5 不同噪声下的 PESQ 性能对比

Fig. 5 Comparison of PESQ under different noises

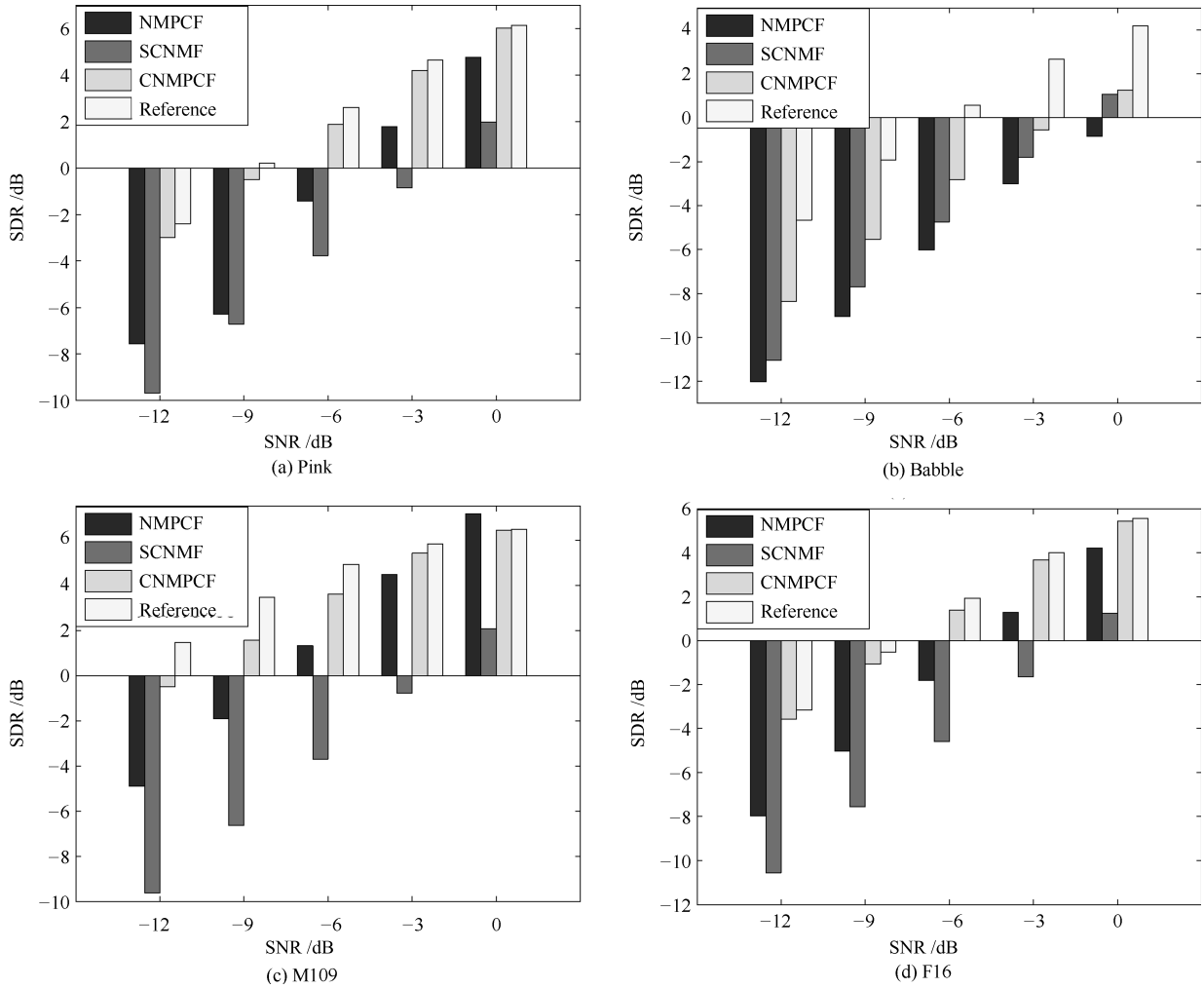


图 6　不同噪声下的 SDR 性能对比

Fig. 6　Comparison of SDR under different noises

SCNMF 算法平均提升了约 9 dB, 相比于 NMPCF 算法平均提升了约 3 dB. CNMPCF 算法获得的分离语音的信噪比增益 ΔSNR 明显优于另外两种分离算法的测量值. 参考结果 (Reference) 相比于 CNMPCF 算法平均高出约 0.01 dB, 但在 Pink 噪声和 F16 噪声条件下低信噪比时, 如 SNR = −9 dB, −12 dB 时, 参考结果 (Reference) 略低于 CNMPCF 结果. 而在 PESQ 和 SDR 指标中, 参考结果 (Reference) 均优于 CNMPCF 算法, 这也许说明不能单从 ΔSNR 一个指标去衡量语音分离的性能, 但 ΔSNR 还是可以作为分离性能的一个参考指标.

表 1 列举了在不同信噪比下, 3 种分离方法的主观听音得分平均值, 结果说明, 在 5 种信噪比情况下, CNMPCF 方法分离的语音的评价得分都略高于 NMPCF 和 SCNMF 方法分离的语音, 多人听音的统计结果相对可靠, 因此主观听音得分也作为语音分离性能的一个参考指标.

总的来说, 与 NMPCF 和 SCNMF 算法相比, 本文所提出的 CNMPCF 算法在不同噪声强度和噪声类型的情况下均能有不同程度的改进. 这主要得益于本算法结合了非负矩阵部分联合分解方法具备的指定源的分离能力, 以及卷积非负矩阵分解方法对语音信号时域相关性的描述能力的缘故.

表 1　5 种信噪比下, 不同方法的主观听音得分平均值

Table 1　The subjective listening score of different methods at five different input SNR levels

SNR (dB)	NMPCF	SCNMF	CNMPCF
−12	1.06	1.08	1.20
−9	1.37	1.46	1.62
−6	1.76	1.95	2.08
−3	2.20	2.29	2.42
0	2.74	2.59	3.05

4　结束语

本文提出了一种卷积非负矩阵部分联合分解的

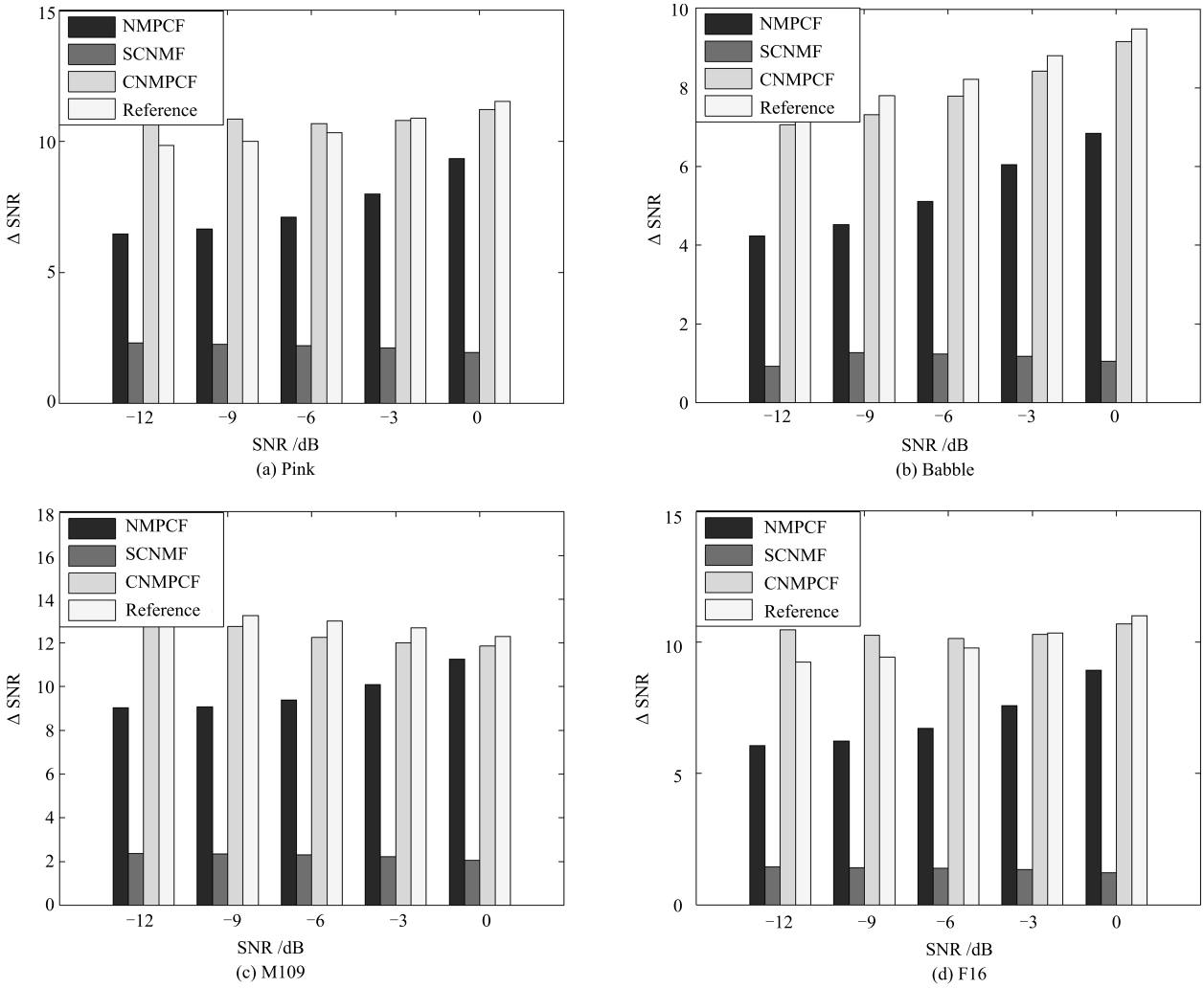


图 7 不同噪声下的 ΔSNR 性能对比

Fig. 7 Comparison of ΔSNR under different noises

语音分离算法, 有效的解决了传统的非负矩阵分解很难确定指定源基向量的困难, 同时考虑到语音信号的时频域相关性, 有效的表征原始语音信号的结构特征, 得到了较好的分离性能. 由于提取的混合语音中的纯噪声频谱段参与联合分解, 可以认为噪声类型和噪声强度都作为先验信息引入联合分解, 因此卷积非负矩阵部分联合分解算法对噪声类型和噪声强度而言有一定鲁棒性, 在低信噪比 (强噪声) 条件下也能获得较好的分离性能, 实验结果表明, 在非平稳噪声和低信噪比的条件下, 相比于以上两种方法均有不同程度的提高.

References

1 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Deep learning for monaural speech separation. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence: IEEE, 2014. 1562−1566

2 Huang P S, Kim M, Hasegawa-Johnson M, Smaragdis P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(12): 2136−2147

3 Liu Wen-Ju, Nie Shuai, Liang Shan, Zhang Xue-Liang. Deep learning based speech separation technology and its developments. *Acta Automatica Sinica*, 2016, **42**(6): 819−833 (刘文举, 聂帅, 梁山, 张学良. 基于深度学习语音分离技术的研究现状与进展. 自动化学报, 2016, **42**(6): 819−833)

4 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, **401**(6755): 788−791

5 Wang D L, Brown G J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Piscataway: IEEE Press, 2006.

6 Han Wei, Zhang Xiong-Wei, Min Gang, Zhang Qi-Ye. A single-channel speech enhancement approach based on perceptual masking deep neural network. *Acta Automatica Sinica*, 2017, **43**(2): 248−258

(韩伟, 张雄伟, 闵刚, 张启业. 基于感知掩蔽深度神经网络的单通道语音增强方法. 自动化学报, 2017, **43**(2): 248−258)

7 Yuan Wen-Hao, Sun Wen-Zhu, Xia Bin, Ou Shi-Feng. Improving speech enhancement in unseen noise using deep convolutional neural network. *Acta Automatica Sinica*, 2018, **44**(4): 751−759
(袁文浩, 孙文珠, 夏斌, 欧世峰. 利用深度卷积神经网络提高未知噪声下的语音增强性能. 自动化学报, 2018, **44**(4): 751−759)

8 Smaragdis P. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(1): 1−12

9 O′Grady P D, Pearlmutter B A. Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, 2008, **72**(1−3): 88−101

10 Sun M, Li Y N, Gemmeke J F, Zhang X W. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence. *IEEE Transactions on Audio, Speech, and Language Processing*, 2015, **23**(7): 1233−1242

11 Kim M, Yoo J, Kang K, Choi S. Blind rhythmic source separation: Nonnegativity and repeatability. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing. Dallas: IEEE, 2010. 2006−2009

12 Yoo J, Kim M, Kang K, Choi S. Nonnegative matrix partial co-factorization for drum source separation. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing. Dallas: IEEE, 2010. 1942−1945

13 Kim M, Yoo J, Kang K, Choi S. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE Journal of Selected Topics in Signal Processing*, 2011, **5**(6): 1192−1204

14 Hu Y, Liu G Z. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2015, **23**(4): 643−653

15 Lu Cheng, Tian Meng, Zhou Jian, Wang Hua-Bin, Tao Liang. A single-channel speech enhancement approach using convolutive non-negative matrix factorization with L1/2 sparse constraint. *Acta Acustica*, 2017, **42**(3): 377−384
(路成, 田猛, 周健, 王华彬, 陶亮. L1/2 稀疏约束卷积非负矩阵分解的单通道语音增强方法. 声学学报, 2017, **42**(3): 377−384)

16 Natarajan B K. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 1995, **24**(2): 227−234

17 Candés E J, Li X D, Ma Y, Wright J. Robust principal component analysis? *Journal of the ACM*, 2009, **58**(3): Article No. 11.

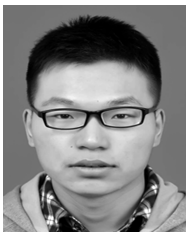
18 Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 1993, **17**: 97−110 2013. 704−708

19 Rix A W, Beerends J G, Hollier M P, Hekstra A P. Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City: IEEE, 2001. 749−752

20 Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(4): 1462−1469

21 Li Y P, Woodruff J, Wang D L. Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, **17**(7): 1361−1371

22 van Segbroeck M. A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice. In: Proceedings of the 2013 Interspeech. Lyon: Interspeech, 2013.



董兴磊 新疆大学信息科学与工程学院硕士研究生. 主要研究方向为语音信号处理, 语音分离.
E-mail: 15739578112@163.com
(**DONG Xing-Lei** Master student in the Department of Information Science and Engineering, Xinjiang University. His research interest covers speech signal processing and speech separation.)

胡 英 新疆大学信息科学与工程学院副教授. 研究方向为音频信息检索, 语音处理. 本文通信作者.
E-mail: huying_75@sina.com
(**HU Ying** Associate professor in the Department of Information Science and Engineering, Xinjiang University. Her research interest covers audio information retrieval and speech processing. Corresponding author of this paper.)

黄 浩 新疆大学信息科学与工程学院教授. 2008 年在上海交通大学电子工程系获博士学位. 主要研究方向语音识别, 多媒体人机交互技术.
E-mail: huanghao@xju.edu.cn
(**HUANG Hao** Professor in the Department of Information Science and Engineering, Xinjiang University. He received his Ph. D. degree from Shanghai Jiao Tong University in 2008. His research interest covers speech recognition and multi-media human-machine interaction.)

吾守尔 · 斯拉木 新疆大学信息科学与工程学院教授. 主要研究方向为语音识别, 语音合成, 多语种信息处理.
E-mail: wushour@xju.edu.cn
(**SILAMU Wushour** Professor in the Department of Information Science and Engineering, Xinjiang University. His research interest covers speech recognition, speech synthesis, and multi-lingual information processing.)