arXiv:2303.00829v2 [eess.AS] 7 Mar 2023

基于噪声空间协方差矩阵学习和最小方差无失真响应的移动机器人自噪声抑制

Pierre-Olivier Lagac´e、Franc¸ois Ferland 和 Franc¸ois Grondin

摘要——得益于深度学习方法，语音和事件识别系统的性能近期显著提升。然而，由于训练神经网络时未能涵盖机器人执行器产生的机械噪声和电气干扰，这些算法部署到机器人上时，部分任务仍颇具挑战。因此，在机器人上使用预训练的语音和事件识别算法时，将自噪声抑制作为预处理步骤有助于解决这一问题。本文提出一种仅使用麦克风阵列和不到两分钟噪声录音来抑制自噪声的新方法。我们利用主成分分析（PCA），从校准过程中在线创建的字典中选出最佳候选协方差矩阵，并将其用于最小方差无失真响应（MVDR）波束形成器。结果表明，所提方法可实时运行，将信号失真比（SDR）最多提高 10 dB，在某些情况下使词错误率（WER）降低 55%，并将事件检测的平均精度（AP）最多提高 0.2。

I. 引言

在人机交互中，音频模态在语音和声音事件识别中发挥着核心作用。过去几年中，人们提出了基于 Transformer 的鲁棒语音和事件识别方法（例如 Whisper [1], Google Speech-to-Text [2] 和 PANNS [3]）。尽管这些系统在环境噪声下表现良好，但遇到异常噪声和低信噪比（SNR）时，性能会迅速下降。当麦克风安装在机器人上时，电机和 WiFi/蓝牙天线产生的噪声会显著降低音频信号的质量。这使降噪变得困难，因为机器人产生的噪声（称为自噪声）通常是非平稳的。

人们提出了许多方法来降低无人地面车辆（UGV）和无人机（UAV）的自噪声。时频空间滤波（TFS） [4], [5] 利用目标声源在每个时频单元上的到达方向（DOA）来屏蔽干扰噪声。由于该方法依赖空间线索，当自噪声与目标声音接近时，性能会下降。另一方面，盲源分离（BSS） [6], [7]通过独立成分分析（ICA）和排列对齐，分别提取各个声源。BSS 在低信噪比下工作效率较高，但会受到排列歧义的影响。为解决排列歧义，研究者提出了 TFS-BSS 方法 [8]，但该方法依赖较长的音频片段，因此不适用于实时应用。

本研究由 NSERC-CREATE/CoRoM 和 NSERC探索基金资助。

P.-O. Lagac´e、F. Ferland 和 F. Grondin 任职于舍布鲁克大学电气工程与计算机工程系。

另一种策略是使用由靠近电机的麦克风 [9], [10], [11] 或压电传感器 [12]采集的参考信号。这种方法假设参考麦克风不会采集到目标声音。然而，这取决于机器人配置，因此难以推广至各种机器人。其他方法依赖于一个数据库，其中包含噪声模板和电机惯性数据，用于选择最符合观测到的执行器特征的噪声实例 [13], [14], [15], [16]。不过，这需要对音频信号和惯性信号进行同步。类似的方法是采用纯音频的基于字典的方法。训练字典以表示自噪声的多个成分，随后利用它从噪声信号中重建并抑制自噪声。非负矩阵分解（NMF） [17], [18], [19] 和相位优化奇异值分解（PO-KSVD）[20]可用于训练字典。利用电机状态从字典成分中重建自噪声。有些 NMF 方法在其框架中不使用电机状态 [21], [22], ，这可能引入非线性失真并降低语音识别性能。另一方面，Fang 等人 [23] 采用不会引入失真的线性运算，但在处理非平稳噪声时表现不佳。基于深度学习的方法也可用于从自噪声中分离目标语音 [24], [25], [26]¦[27], [28]。这些模型在包含自噪声和目标声音的大型数据集上训练时表现良好。然而，这些经过训练的网络在不同且未见过的音频场景中通常表现不佳，而这种情况在机器人技术中很常见。

本文提出了一种基于模板的方法，该方法利用自噪声空间协方差矩阵（SCM）与最小方差无失真响应（MVDR）算法。该方法在提高信号失真比的同时，能够稳健地应对非平稳噪声，并可在新的音频域中快速完成校准。这一方法颇具吸引力，因为它无需重新训练深度神经网络，而重新训练通常需要数小时的音频样本和机器人上无法提供的计算资源。通过校准步骤生成自噪声空间协方差矩阵字典。采用主成分分析（PCA）减少空间协方差矩阵的维度。利用这一紧凑表示，在数据库中的空间协方差矩阵与测得的空间协方差矩阵之间寻找最佳匹配。随后，将对应的空间协方差矩阵用于最小方差无失真响应算法，以分离目标声源与自噪声。结果表明，该方法能够提升语音识别和事件识别的性能。

本文的组织结构如下。第二节介绍整体自噪声抑制方法。第三节介绍实验设置，随后第四节给出实验结果

这些结果是使用真实录制的数据获得的，第五节则总结本文。

II. 基于空间协方差矩阵估计的最小方差无失真响应

麦克风 $m \in \{1, \ldots, M\}$ 的多通道目标声、自噪声和音频混合信号分别表示为 $\mathbf{s}[n] \in \mathbb{R}^M$、$\mathbf{b}[n] \in \mathbb{R}^M$ 和 $\mathbf{x}[n] \in \mathbb{R}^M$，其中 $M \in \mathbb{N}$ 表示麦克风数量，$n \in \mathbb{N}$ 对应采样索引。假设噪声为纯加性噪声，因此：

$$\mathbf{x}[n] = \mathbf{s}[n] + \mathbf{b}[n]. \tag{1}$$

这些信号可通过短时傅里叶变换（STFT）转换到频域，帧长为 $N \in \mathbb{N}$ 个采样点，帧移为 $\Delta N \in \mathbb{N}$ 个采样点，分别记为 $\mathbf{X}[k,l] \in \mathbb{C}$、$\mathbf{S}[k,l] \in \mathbb{C}$ 和 $\mathbf{B}[k,l] \in \mathbb{C}$，其中 $k \in \{0, 1, \ldots, K-1\}$ 和 $l \in \mathbb{N}$ 分别表示频率 bin 和帧索引。输入信号 $\Phi_{\mathbf{XX}}[k]$ 的空间协方差矩阵按以下方式在 L 帧上计算：

$$\Phi_{\mathbf{XX}}[k] = \frac{1}{L} \sum_{l=1}^{L} \mathbf{X}[k,l]\mathbf{X}[k,l]^H, \tag{2}$$

其中 $\{\ldots\}^H$ 表示厄米算子。

校准阶段首先录制机器人几分钟的自噪声（本文的方法只需 90 秒的自噪声数据）。自噪声录音被分割为 $J \in \mathbb{N}$ 个长度为 0.5 秒的片段，并使用式 (2) 计算每个片段的空间协方差矩阵 j。由于空间协方差矩阵是厄米矩阵，因此可将其上三角部分（包括对角线）展平，构成一个向量 $\mathbf{v}_j[k] \in \mathbb{C}^{M(M+1)/2}$，其中包含每个频率 bin k 的所有相关元素。随后，将所有频率 bin 对应的这些向量拼接成一个超向量 $\mathbf{V}_j \in \mathbb{C}^{KM(M+1)/2}$。使用主成分分析（PCA）可将每个超向量降维为一个稠密向量，记作 $\mathbf{D}_j \in \mathbb{C}^I$。

测试时，对由 0.5 秒噪声音频片段的空间协方差矩阵得到的超向量应用相同的主成分分析变换，结果记为 $\hat{\mathbf{D}} \in \mathbb{C}^I$。计算 $\hat{\mathbf{D}}$ 与字典中向量 $\mathbf{D}_j$ 之间的 l^2-范数，然后选出字典中距离最近的向量 $\mathbf{D}_{j^*}$：

$$j^* = \arg\min_j (\|\hat{\mathbf{D}} - \mathbf{D}_j\|_2^2). \tag{3}$$

使用主成分分析进行降维有两个目的：1）减少内存占用和计算量；2）将超向量观测投影到噪声子空间，从而简化带噪空间协方差矩阵（目标和噪声）与噪声空间协方差矩阵之间的比较。在低功耗嵌入式计算机上可以高效地进行主成分分析分解，因此，只要录制好校准音频信号，就能在机器人上快速生成字典。相应的噪声空间协方差矩阵估计 $\hat{\Phi}_{\mathbf{BB}}[k]$ 用于计算最小方差无失真响应权重：

$$\mathbf{w}[k] = \frac{\hat{\Phi}_{\mathbf{BB}}^{-1}[k]\Phi_{\mathbf{XX}}[k]}{\mathrm{Tr}\{\hat{\Phi}_{\mathbf{BB}}^{-1}[k]\Phi_{\mathbf{XX}}[k]\}}\mathbf{u}, \tag{4}$$

其中，每个带噪空间协方差矩阵（记为 $\hat{\Phi}^{-1}_{\mathbf{BB}}[k]$）的逆矩阵都可以在校准过程中预先计算一次，并直接存储在字典中，以加快计算速度。选择独热参考向量 $\mathbf{u} \in \{0,1\}^M$，以选取信噪比最高的麦克风；信噪比可根据所选噪声协方差矩阵估算。最后应用波束形成，生成增强信号：

$$Y[k,l] = \mathbf{w}[k]^H \mathbf{X}[k,l]. \tag{5}$$

随后可使用逆短时傅里叶变换（iSTFT）将信号变换回时域并获得 $y[n]$。

图1总结了所提出的框架。

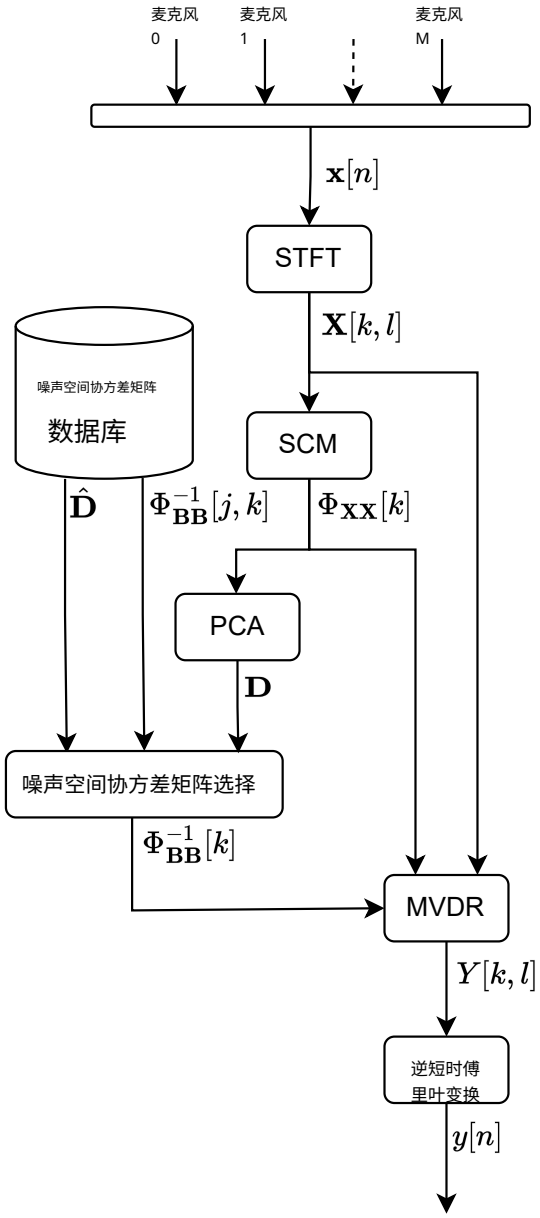

图 1. 所提框架的框图。

III. 实验设置

图2展示了本实验中使用的配备16SoundsUSB[2] MA的Clearpath Robotics Jackal无人地面车辆[1]。16个全向麦克风布置在机器人机架周围。其中一半比其余麦克风高3厘米，以确保在 z轴上的空间辨别能力。MA通过USB连接到一台笔记本电脑，该电脑运行以Python和ROS[3]实现的框架。音频以32000个采样点/秒的速率采样，并划分为0.5秒的片段。短时傅里叶变换使用长度为 $N = 2048$ 个采样点的汉宁窗，帧移为 $\Delta N = 256$ 个采样点。

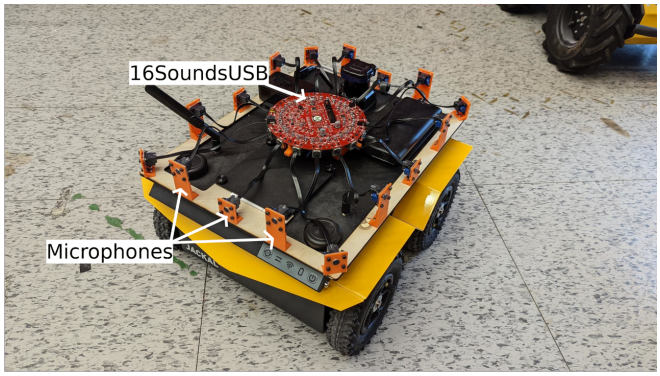

图 2. 配备麦克风阵列的 Clearpath Jackal 机器人

实验在三个不同的房间中进行：一间面积为 30 m^2，的小型办公室、一间面积为 150 m^2 的大房间，以及一条面积为 30 m^2 的走廊。各房间的混响时间 RT_{60} 分别为 200 毫秒、600 毫秒和 160 毫秒，这些数值是根据各房间录制的拍手声估算得出的。

为进行评估，自噪声和语音分别录制。机器人以不同速度（最高 0.4 米/秒）在每个房间内移动，并朝任意方向运动，同时录制自噪声。图 2 展示了机器人在大房间内移动的情形。录制自噪声时，除环境噪声外，机器人周围没有其他声源处于活动状态。随后，扬声器播放目标声源，音量调节至日常交谈的水平。扬声器由人工在机器人周围移动，距离介于 1 至 3 米之间，以再现机器人与目标声源之间的相对运动。目标声源包括语音、音乐、尖叫声、警报声和关门声。使用了 LibriSpeech [29] 数据集中的 36 个音频文件，分别来自 3 名男性说话者和 3 名女性说话者。这些语音片段的时长为 6 至 15 秒。扬声器还播放音乐（5 秒）、尖叫声（5 秒）和警报声（3 秒）。关门声来自房间内实际的门。为进行校准，共录制了 90 秒自噪声；另录制了 120 秒自噪声，并将其与目标声源录音随机混合，以供验证。每个房间由此获得 288 个含噪语音片段和 120 个含噪声音事件。

[1]https://clearpathrobotics.com/jackal-small-unmanned-ground-vehicle/ [2] https://github.com/introlab/16SoundsUSB [3] https://github.com/introlab/egonoise/

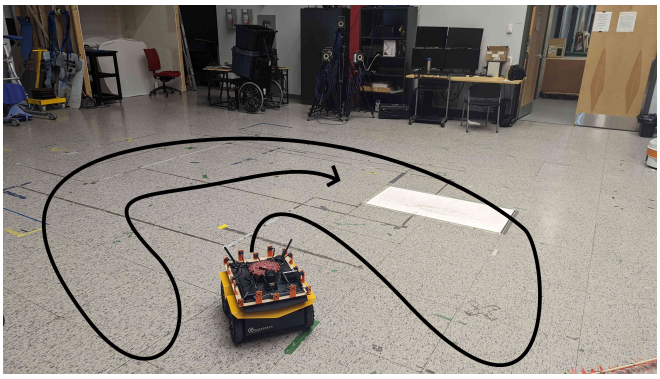

图 3. 机器人在大房间内移动

IV. 结果与讨论

用于分析所提出的框架性能的指标包括信噪比（SNR）、信号失真比（SDR）、语音识别的词错误率[4]（WER），以及用于事件检测分类的平均精度[5]（AP）。语音识别采用 Whisper 的小型英语模型。声音事件分类采用 PANNS 的 Wavegram Logmel Cnn14 模型。对于每项指标，均将框架输出的增强信号与输入信号和仅语音信号进行比较。输入信号和仅语音信号取自麦克风阵列（MA）中的一个麦克风；与其他一些更靠近执行器的麦克风相比，该麦克风的平均信噪比更高。

A. 语音识别

图4展示了大房间中输入、滤波和仅语音频谱图的一个示例，此时输入信噪比为 -6.09 dB。在输入频谱图中，语音几乎完全被自噪声掩盖。采用所提出方法得到的滤波信号信噪比为 6.31 dB，频谱图中现在可以看到许多语音特征。戴尔 XPS 笔记本电脑在 Python 环境中使用 NumPy [6], 可在 0.2 秒内处理一个 0.5秒片段，这证实了该框架能够实时执行自噪声抑制。

表I给出了每个房间中语音片段的平均信噪比和信号失真比。对于这些指标，仅使用包含目标声音的片段，并舍弃静音片段。

图5和图6展示了所有 0.5秒片段的输入和滤波信噪比及信号失真比的散点图。结果表明，与输入信号相比，增强信号具有更高的信噪比和信号失真比。结果显示出线性性能，但在低信噪比（约为 -5 dB 和 -10 dB）下除外；此时，音频信号中目标声音的功率较弱，可能导致最小方差无失真响应性能下降。

表II展示了输入信号、增强信号和仅语音信号的平均词错误率。正如预期，输入信号的词错误率

[4]https://torchmetrics.readthedocs.io [5] https://scikit-learn.org [6] https://numpy.org

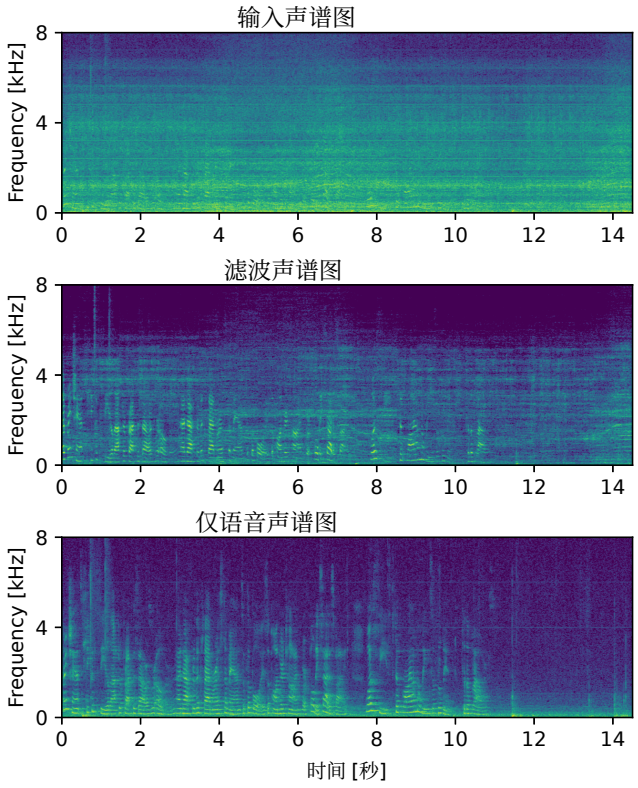

图4。输入、滤波和仅语音信号的频谱图。输入信号和滤波信号的信噪比分别为 -6.09 dB 和 6.31 dB，这意味着信噪比提高了 12.4 dB。

表I 各房间语音片段的平均输入信噪比、增强信噪比和信号失真比

Room	信噪比 (dB)		信号失真比 (dB)	
	输入	增强	输入	增强
小	1.5	11.29	2.14	11.87
大	-2.69	8.7	-1.75	9.57
走廊	1.6	12.06	2.28	12.72

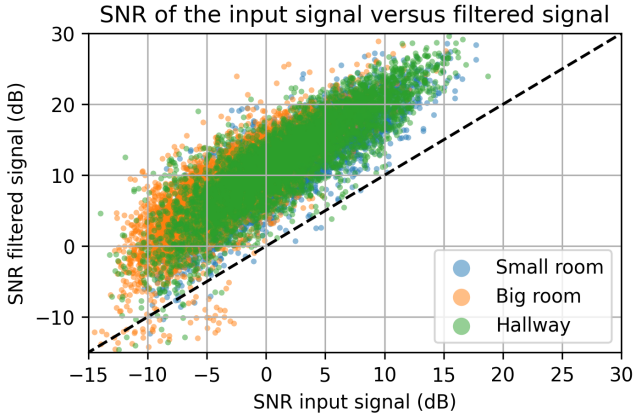

图5. 所有0.5秒片段滤波前后的信噪比

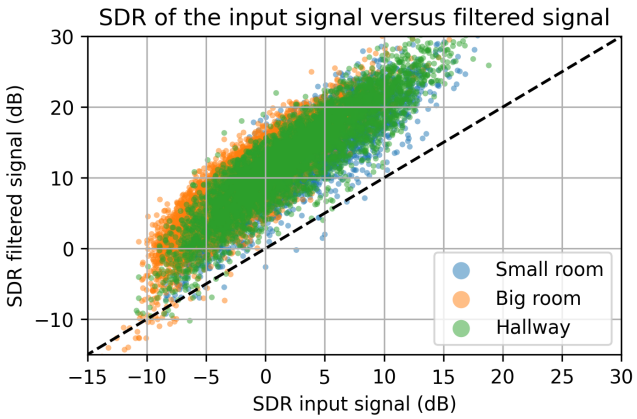

图6. 所有0.5秒片段滤波前后的信号失真比

表II 所提出的框架的词错误率（WER）

Room	输入	增强	仅语音
小	62.6%	30.9%	11.1%
大	90.3%	35.5%	14.9%
走廊	62.4%	25.3%	8.52%

在三个房间中均高于仅语音信号。所提出的自噪声抑制方法将词错误率显著降低了30%至55%。在小房间和走廊录制的输入信号信噪比较高，因此输入信号的词错误率表现更好。

B. 声音事件检测

图 7 展示了大房间中的输入、滤波和仅含音乐的频谱图示例。输入信号的信噪比为 -1.93 dB，滤波信号的信噪比为 13.13 dB。在输入信号频谱图中，音乐主要被自噪声掩盖，而在增强信号频谱图中，谐波则清晰可见。

图 8 和图 9 展示了大房间中所有事件的输入信噪比、滤波信噪比和信号失真比。由于三个房间中的结果相似，为便于说明，我们聚焦于大房间。尽管各事件之间的差异主要体现在输入信号的信噪比上，但该系统在大多数场景中都能提升声音质量。

使用 PANNS 模型对输入信号、增强信号和仅事件信号中的声音事件进行分类。表 III 所示的结果表明，增强信号提高了音乐信号的平均精度。一些结果显示，仅事件信号的平均精度低于噪声信号，而增强信号的平均精度甚至更高。结果总体上优于在[3]中观察到的结果。与 Audioset 不同，本实验中录制的数据清晰，且不受其他声音干扰。然而，对于关门撞击声事件，观察到的性能有所下降。我们认为，这是因为关门声短促且微弱，在计算输入信号的空间协方差矩阵时可能更难捕捉。尽管警报声事件仍然难以检测，

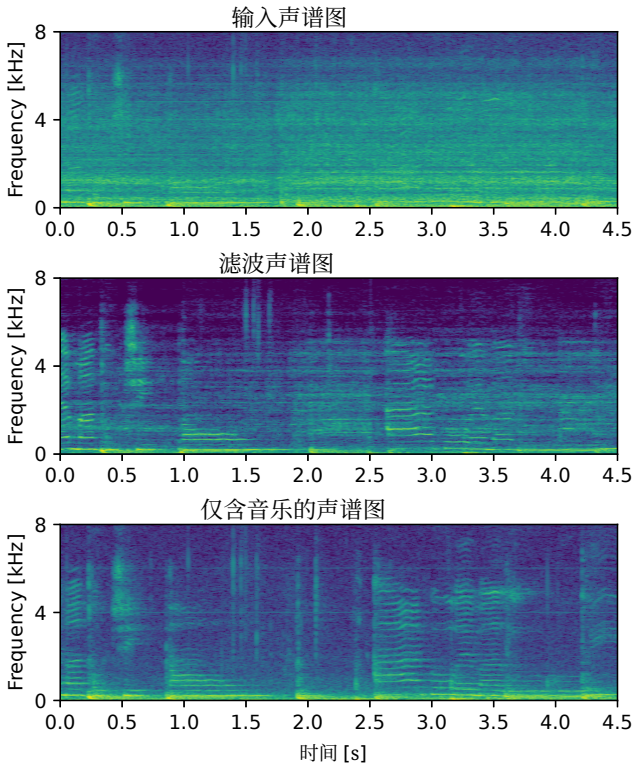

图 7. 输入信号、滤波后信号和仅含音乐信号的频谱图。输入信号和滤波后信号的信噪比分别为 -1.93 dB 和 13.13 dB。

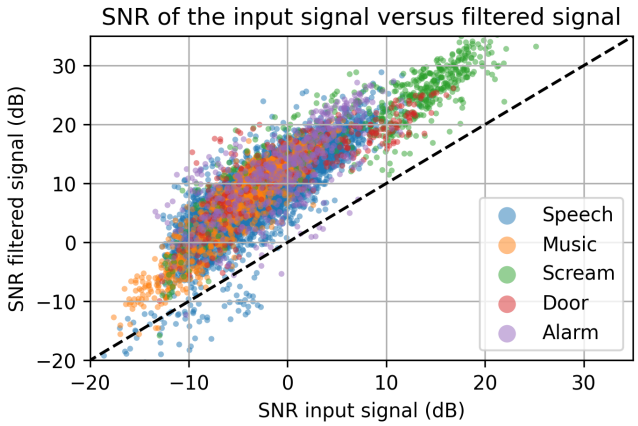

图 8. 各声音事件类型在滤波前后的信噪比

但使用 PANNS 时，该框架仍能略微提高平均精度。我们认为，该框架对信噪比较低的音频片段更有帮助。

V. 结论

本文介绍了一种高效框架，利用麦克风阵列降低机器人的自噪声。该框架依赖于在短暂校准期内训练得到的噪声空间协方差矩阵字典，以及用于增强目标声源的最小方差无失真响应算法。为获得更好的性能，空间协方差矩阵被串联成一个超向量，并使用主成分分析算法在离线阶段降低其维度。使用一辆无人地面车辆和一个 16 麦克风阵列获得的结果表明，

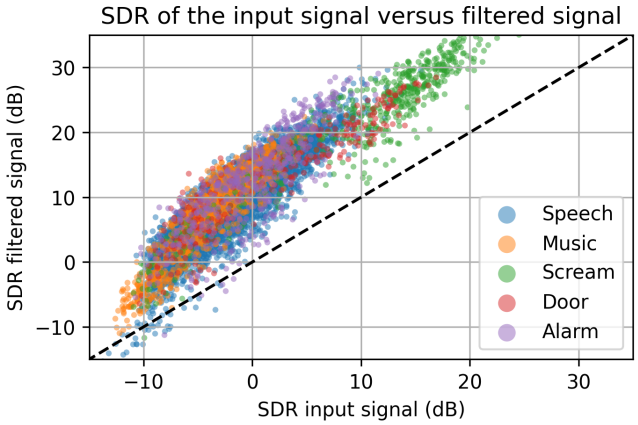

图 9. 各声音事件类型滤波前后的信号失真比

表 III 所提出的框架的事件检测平均精度

事件	Room	输入	增强	仅事件
语音	小	0.999	1.0	1.0
	大	0.994	1.0	1.0
	走廊	0.999	1.0	1.0
音乐	小	0.722	0.792	0.914
	大	0.487	0.685	0.784
	走廊	0.868	0.953	0.948
尖叫	小	0.993	0.994	0.992
	大	0.998	0.998	1.0
	走廊	0.994	1.0	1.0
Door	小	0.313	0.232	0.104
	大	0.178	0.109	0.091
	走廊	0.129	0.100	0.089
警报	小	0.091	0.105	0.108
	大	0.091	0.093	0.096
	走廊	0.090	0.098	0.105

仅需 90 秒校准时间，增强信号的信噪比和信号失真比平均提高约 10 dB。此外，将增强语音信号输入自动识别系统后，与使用含噪输入信号时的性能相比，词错误率降低了 55%。该方法还提升了声音事件检测的平均精度，在某些情况下最高提升 0.2。这个新框架为在机器人上集成语音识别和声源检测带来了新的机遇，尤其适用于机器人在深度神经网络模型训练阶段会产生未见过的自噪声的情况。

致谢

作者谨此感谢 Alexandre Girard 教授提供使用 Clearpath Robotics Jackal机器人的机会，并感谢 Samuel Faucherforthe 富有启发性的讨论。

参考文献

[1] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现稳健语音识别》，arXiv 预印本 arXiv:2212.04356，2022。[2] C.-C. Chiu、T. N. Sainath、Y. Wu、R. Prabhavalkar、P. Nguyen、Z. Chen、A. Kannan、R. J. Weiss、K. Rao、E. Gonina 等，《利用序列到序列模型实现最先进的语音识别》，载于 IEEE声学、语音与信号处理国际会议论文集，2018年，第4774–4778页。[3] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. D. Plumbley，《PANNS：用于音频模式识别的大规模预训练音频神经网络》，IEEE/ACM音频、语音与语言处理汇刊，第28卷，第2880–2894页，2020年。[4] L. Wang 和 A. Cavallaro，《微型飞行器声源定位的时频处理》，载于 IEEE声学、语音与信号处理国际会议论文集，2017年，第496–500页。[5] ——，《多旋翼无人机的声学感知》，IEEE传感器期刊，第18卷，第11期，第4570–4582页，2018年。[6] ——，《空中之耳：听觉微型飞行器的自噪声抑制》，载于 IEEE高级视频与基于信号的监控国际会议论文集，2016年，第152–158页。[7] ——，《听觉微型飞行器的麦克风阵列自噪声抑制算法》，IEEE传感器期刊，第17卷，第8期，第2447–2455页，2017年。[8] ——，《用于多旋翼无人机自噪声抑制的盲源分离框架》，IEEE/ACM音频、语音与语言处理汇刊，第28卷，第2523–2537页，2020年。[9] B. Kang、H. Ahn 和 H. Choo，《用于配备声音传感器的无人机降噪的软件平台》，IEEE传感器期刊，第19卷，第21期，第10 121–10 130页，2019年。[10] S. Yoon、S. Park、Y. Eom 和 S. Yoo，《用于广播多旋翼无人机的自适应降噪系统及先进声音采集方法》，载于 IEEE消费电子国际会议论文集，2015年，第26–29页。[11] S. Yoon、S. Park 和 S. Yoo，《用于广播多旋翼无人机的两阶段自适应降噪系统》，载于 IEEE消费电子国际会议论文集，2016年，第219–222页。[12] R. P. Fernandes、E. C. Santos、A. L. L. Ramos 和 J. A. Apolin´ario Jr.，《利用压电传感器增强四旋翼无人机信号的初步研究》，载于变革科学与工程、商业和社会创新国际会议论文集，2015年，第536–541页。[13] G. Ince、K. Nakadai 和 K. Nakamura，《基于模板的多通道自噪声估计在线学习》，载于 IEEE/RSJ智能机器人与系统国际会议论文集，2012年，第3282–3287页。[14] G. Ince、K. Nakadai、T. Rodemann、Y. Hasegawa、H. Tsujino 和 J. Imura，《利用模板减法抑制机器人的自噪声》，载于 IEEE/RSJ智能机器人与系统国际会议论文集，2009年，第199–204页。[15] G. Ince、K. Nakadai、T. Rodemann、Y. Hasegawa、H. Tsujino 和 J. Imura，《用于机器人自噪声消除的混合框架》，载于 IEEE机器人与自动化国际会议论文集，2010年，第3623–3628页。[16] G. Ince、K. Nakadai、T. Rodemann、J. Imura、K. Nakamura 和 H. Nakajima，《机器人自噪声估计的增量学习》，载于 IEEE/RSJ智能机器人与系统国际会议论文集，2011年，第131–136页。[17] A. Schmidt 和 W. Kellermann，《利用电机数据驱动字典进行知情式自噪声抑制》，载于 IEEE声学、语音与信号处理国际会议论文集，2019年，第116–120页。[18] A. Schmidt、A. Brendel、T. Haubner 和 W. Kellermann，《用于自噪声抑制的电机数据正则化非负矩阵分解》，音频、语音与音乐处理期刊，第11卷，第1–15页，2020年。[19] A. Schmidt 和 W. Kellermann，《采用电机数据正则化激活的多通道非负矩阵分解，实现稳健自噪声抑制》，载于 IEEE自主系统国际会议论文集，2021年，第1–5页。[20] A. Deleforge 和 W. Kellermann，《用于欠定多通道稀疏混合信号提取的相位优化 k-SVD》，载于 *IEEE*声学、语音与信号处理国际会议论文集，2015年，第355–359页。[21] T. Tezuka、T. Yoshida和K. Nakadai，“基于半盲无限非负矩阵分解的机器人自运动噪声抑制”，载于*IEEE*机器人与自动化国际会议论文集，2014年，第6293–6298页。[22] T. Haubner、A. Schmidt和W. Kellermann，“用于自噪声抑制的多通道非负矩阵分解”，载于语音通信论文集，2018年，第136–140页。[23] H. Fang、G. Carbajal、S. Wermter和T. Gerkmann，“利用部分自适应字典联合降低自噪声和环境噪声”，载于*ITG*语音通信会议论文集，2021年，第114–118页。[24] A. Ito、T. Kanayama、M. Suzuki和S. Makino，“小型机器人用于语音识别的内部噪声抑制”，载于欧洲语音通信与技术会议论文集，2005年，第2685–2688页。[25] A. Briegleb、A. Schmidt和W. Kellermann，“单通道自噪声抑制的深度聚类”，载于国际声学大会论文集，2019年，第2813–2820页。[26] J. Kim、J. Choi、J. Son、G. Kim、J. Park和J. Chang，“在移动机器人中保留声源定位空间线索的MIMO噪声抑制”，载于*IEEE*电路与系统国际研讨会论文集，2021年，第1–5页。[27] Z. Tan、A. H. T. Nguyen和A. W. H. Khong，“一种用于低输入信噪比下无人机噪声降低的高效空洞卷积神经网络”，载于亚太信号与信息处理协会年度峰会暨会议论文集，2019年，第1885–1892页。[28] L. Wang和A. Cavallaro，“用于无人机语音增强的深度学习辅助时频处理”，*IEEE*计算智能新兴主题汇刊，第5卷，第6期，第871–881页，2020年。[29] V. Panayotov、G. Chen、D. Povey和S. Khudanpur，“LibriSpeech：基于公有领域有声读物的自动语音识别语料库”，载于*IEEE*声学、语音与信号处理国际会议论文集，2015年，第5206–5210页。