arXiv:2302.04469v1 [cs.SD] 9 Feb 2023

使用卡尔曼滤波器的联合声学回声消除与语音去混响

Ziteng Wang、*Yueyue Na*、*Biao Tian*、*Qiang Fu*

阿里巴巴集团，中国

摘要

本文提出了一种在短时傅里叶变换域中联合进行声学回声消除（AEC）和语音去混响（DR）的算法。混响麦克风信号采用自回归（AR）模型描述。自回归系数和扬声器到麦克风的声学传递函数（ATF）被视为时变参数，并通过一阶马尔可夫过程同时建模。由此得到的方案可使用卡尔曼滤波器对这些参数进行最优估计。结果表明，与依次处理 AEC 和 DR 的常规方案，以及一种基于半盲源分离的先进联合 DRAEC 算法相比，所提出的算法在语音质量和回声抑制性能方面均表现更优。

索引词— 声学回声消除，混响抑制，卡尔曼滤波器

1. 引言

声学回声消除（AEC）和语音去混响（DR）在实时通信和自动语音识别等语音处理应用中发挥着重要作用。AEC旨在去除播放信号引入的回声。DR旨在抑制混响并恢复直达声，同时可选择保留早期反射。归一化最小均方（NLMS）滤波器 [1], 递归最小二乘（RLS）滤波器以及加权预测误差（WPE）算法 [2]等自适应滤波器广泛用于这两项任务。卡尔曼滤波器也分别应用于AEC [3, 4, 5, 6]和DR [7, 8, 9, 10]。

从自适应滤波的角度来看，声学回声消除和去混响在求解过程中有许多相似之处。这两项任务都涉及房间脉冲响应的估计，其中一项关联声源位置与麦克风位置，另一项关联扬声器位置与麦克风位置。能够同时处理声学回声消除和去混响的联合算法，已显示出优于分别处理部分任务的算法的性能提升 [11, 12, 13, 14]。Takeda 等人 [11] 通过应用频域独立成分分析（ICA）模型，实现了盲去混响和回声消除。该方法假设直达声帧与晚期混响及播放信号相互独立；在多输入/输出逆滤波定理（MINT）的条件下，这一假设近似成立。Togami 和 Kawaguchi [13]假设麦克风输入信号服从时变局部高斯模型，并据此结合声学回声抑制、语音去混响和噪声抑制。在统一似然函数下优化线性滤波器，该函数能够直接反映最终的语音增强性能。Cohen 等人 [15]提出了一种基于滑动平均模型的语音去混响卡尔曼-EM 方法，并在 E 步中采用卡尔曼滤波器，估计回声消除后的干净信号。

混响信号通常在短时傅里叶变换（STFT）域中用自回归（AR）模型描述。我们之前的工作 [16, 17]分别从半盲源分离的角度重新表述了AEC和DR。文献 [18]进一步推导出一种基于RLS的联合去混响与声学回声消除算法。鉴于RLS可视为卡尔曼滤波器的一种特殊情况 [19], , 本文进一步提出一种使用卡尔曼滤波器的联合去混响与声学回声消除算法。自回归模型系数以及扬声器到麦克风的声学传递函数（ATF）均被视为时变参数，并用一阶马尔可夫过程同时建模。通过最小化统一的均方误差损失函数，推导出一种新颖的联合去混响与声学回声消除算法。该联合算法不仅优于其RLS对应算法，也优于依次使用基于卡尔曼滤波器的AEC和基于卡尔曼滤波器的DR的级联方案。

2. 信号模型

我们考虑短时傅里叶变换（STFT）域中的多通道卷积混合。由 M 个麦克风组成的传感器阵列采集来自声源 $S(t,\ f)$ 的信号，以及由扬声器播放的信号$X(t,\ f)$，其中 t 和 f 分别表示时间索引和频率索引。第 f 个频带中第 m 个麦克风的信号为：

$$Y_m(t) = \sum_{l=0}^{\infty} A_{m,l}(t)S(t-l) + \sum_{l=0}^{\infty} B_{m,l}(t)X(t-l) + V_m(t) \quad (1)$$

其中，$A_{m,\ l}$ 表示声源到麦克风的传递函数，$B_{m,\ l}$ 表示扬声器到麦克风传递函数，V_m 表示背景噪声。该信号可用自回归模型 [13] 近似表示为：

$$Y_m(t) = S_m(t) + \sum_{n=1}^{M}\sum_{l=0}^{L_Y-1} C_{m,n,l}Y_n(t-\Delta-l) + \sum_{l=0}^{L_X-1} B_{m,l}(t)X(t-l) + \tilde{V}_m(t) \quad (2)$$

其中，S_m 表示直达路径声音和早期反射，Δ标记早期反射与晚期混响之间的边界，$C_{m,\ n,\ l}$ 表示多通道自回归系数，V包含建模误差。

式（2）右侧的第一项定义为待恢复的目标：

$$S_m(t) = Y_m(t) - \mathbf{w}_m^H(t)\mathbf{z}(t) \quad (3)$$

其中

$$\mathbf{w}_m = [\underbrace{B^*_{m,0},...,B^*_{m,L_X-1}}_{\mathbf{w}_{\text{AEC}}}, \underbrace{C^*_{m,1,0},...,C^*_{m,1,L_Y-1},...,C^*_{m,M,0},...,C^*_{m,M,L_Y-1}}_{\mathbf{w}_{\text{DR}}}]^T \quad (4)$$

是一个统一向量，由长度为 L_X 的声学回声消除相关滤波器和长度为 ML_Y的去混响相关滤波器组成，并且

$$\begin{aligned}\mathbf{z}(t) = [&X(t),...,X(t-L_X+1),\\ &Y_1(t-\Delta),...,Y_1(t-\Delta-L_Y+1),\\ &...,\\ &Y_M(t-\Delta),...,Y_M(t-\Delta-L_Y+1)]^T\end{aligned} \quad (5)$$

由播放信号和延迟后的麦克风观测值拼接而成。$(\cdot)^*$ 表示复共轭，$(\cdot)^H$ 表示共轭转置，$(\cdot)^T$ 表示转置。

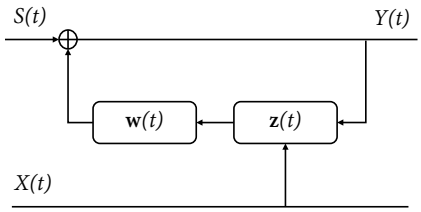
通过最小化平方误差损失函数，可以得到滤波器系数 (4) 的最优估计：

$$J_{\mathbf{w}} = \min \sum_{m=1}^{M} E\{|\mathbf{w}_m(t) - \hat{\mathbf{w}}_m(t)|^2\} \quad (6)$$

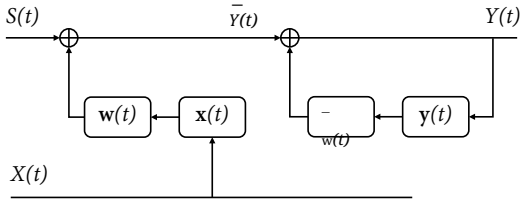
ˆ其中 $(\cdot)$ 表示变量估计值。我们假设每个麦克风独立执行以下算法，为简洁起见，省略下标 m。

3. 所提算法

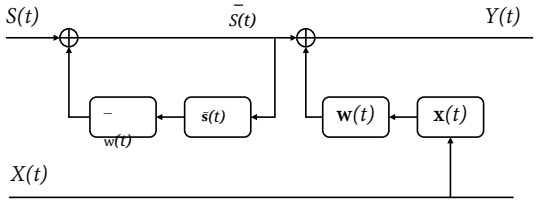
在实际声学场景中，声源到麦克风和扬声器到麦克风的传递函数都会随时间变化，



a. 联合 DRAEC



b. 级联 DR-AEC



c. 级联式 AEC-DR

图 1。联合 DRAEC 以及级联式去混响-声学回声消除和声学回声消除-去混响的声源模型。

因此，滤波器状态 $\{B_l, C_{n,l}\}$ 也被视为随时间变化。这可由一阶马尔可夫过程描述：

$$\mathbf{w}(t) = A(t) * \mathbf{w}(t-1) + \mathbf{u}(t) \quad (7)$$

其中 $A(t)$ 为状态转移参数，而 $\mathbf{u}(t) \sim \mathcal{N}(\mathbf{0},\ \mathbf{\Phi_u}(t))$ 描述服从零均值复高斯分布的过程噪声。

3.1. 基于卡尔曼滤波的双重回声消除与去混响

根据上述公式，应用著名的卡尔曼滤波器[20]来估计 $\mathbf{w}(t)$。我们将 $\mathbf{\Phi}$ 表示为状态向量误差协方差矩阵：

$$\mathbf{\Phi}(t) = E\{[\mathbf{w}(t) - \hat{\mathbf{w}}(t)][\mathbf{w}(t) - \hat{\mathbf{w}}(t)]^H\} \quad (8)$$

滤波器更新方程如下：

$$S(t|t-1) = Y(t) - \mathbf{w}^H(t|t-1)\mathbf{z}(t) \quad (9)$$

$$\hat{\mathbf{w}}(t) = \mathbf{w}(t|t-1) + \mathbf{k}(t)S^*(t|t-1) \quad (10)$$

$$\mathbf{\Phi}(t) = [\mathbf{I} - \mathbf{k}(t)\mathbf{z}^H(t)]\mathbf{\Phi}(t|t-1) \quad (11)$$

其中卡尔曼增益

$$\mathbf{k}(t) = \frac{\mathbf{\Phi}(t|t-1)\mathbf{z}(t)}{\phi_S(t) + \mathbf{z}^H(t)\mathbf{\Phi}(t|t-1)\mathbf{z}(t)}, \quad (12)$$

$\phi_S(t) = E\{S(t)S^*(t)\}$ 表示期望声源的功率谱密度，而$\mathbf{I}$ 是一个尺寸适当的单位矩阵。下一时间步的预测方程如下：

$$\mathbf{w}(t+1|t) = A(t)\hat{\mathbf{w}}(t) \tag{13}$$
$$\mathbf{\Phi}(t+1|t) = A^2(t)\mathbf{\Phi}(t) + \mathbf{\Phi}_\mathbf{u}(t) \tag{14}$$

最终通过以下方式恢复目标声源：

$$\hat{S}(t) = Y(t) - \hat{\mathbf{w}}^H(t)\mathbf{z}(t) \tag{15}$$

3.2. 级联方案

图 1 展示了联合 DRAEC 以及两种备选级联方案 DR-AEC 和级联式 AEC-DR 的声源模型，后两者根据其求解顺序命名。为便于比较，我们将 AEC-DR 的解表述为：

$$\bar{S}(t) = Y(t) - \mathbf{w}_{\text{AEC}}^H(t)\mathbf{x}(t)$$
$$S(t) = \bar{S}(t) - \bar{\mathbf{w}}_{\text{DR}}(t)\bar{\mathbf{s}}(t) \tag{16}$$

其中，$\mathbf{x}(t)$ 是播放信号向量，而 $\bar{}\mathbf{s}(t)$ 是经过声学回声消除后延迟$\bar{}$的信号向量 $\{S_m(t-\Delta-l)\}$。请注意，自回归系数 $\bar{}\mathbf{w}_{\text{DR}}$ 不同于式 (4) 中定义的$\mathbf{w}_{\text{DR}}$，因为所涉及的信号不同。因此，去混响滤波器的性能容易受到声学回声消除滤波器性能的影响。此时的损失函数为：

$$J_\mathbf{w} = \min E\{|\mathbf{w}_{\text{AEC}}(t) - \hat{\mathbf{w}}_{\text{AEC}}(t)|^2\} + \min E\{|\bar{\mathbf{w}}_{\text{DR}}(t) - \hat{\bar{\mathbf{w}}}_{\text{DR}}(t)|^2\} \tag{17}$$

通过依次执行基于卡尔曼滤波器的声学回声消除和基于卡尔曼滤波器的去混响，即可对其进行优化。

3.3. 参数估计

卡尔曼滤波器需要为 $\mathbf{\Phi}_\mathbf{u}(t)$ 和$\phi_S(t)$ 配备合适的估计器。类似地，如同在 [8], 中，我们使用 $\mathbf{\Phi}_\mathbf{u}(t) = \phi_\mathbf{u}(t)\mathbf{I}$，并假设$\mathbf{w}(t)$ 中的元素互不相关且同分布。方差参数

$$\phi_\mathbf{u}(t) = \frac{1}{L}E\{|\hat{\mathbf{w}}(t) - \hat{\mathbf{w}}(t-1)|^2\} + \eta \tag{18}$$

通过滤波器系数随时间的变化进行估计，其中 L 表示滤波器长度，η 是一个较小的正数，用于在声学环境发生变化时保持跟踪能力。

ϕ_S 的最大似然估计为：

$$\hat{\phi_S}(t) = \alpha\phi(t-1) + (1-\alpha)S(t|t-1)S^*(t|t-1) \tag{19}$$

and

$$\phi(t) = \alpha\phi(t-1) + (1-\alpha)\hat{S}(t)\hat{S}^*(t) \tag{20}$$

其中，α 为递归平滑因子。

对于初始化，我们使用$\mathbf{w(0)} = \mathbf{0}$ 和 $\mathbf{\Phi(0)} = \mathbf{I}$。

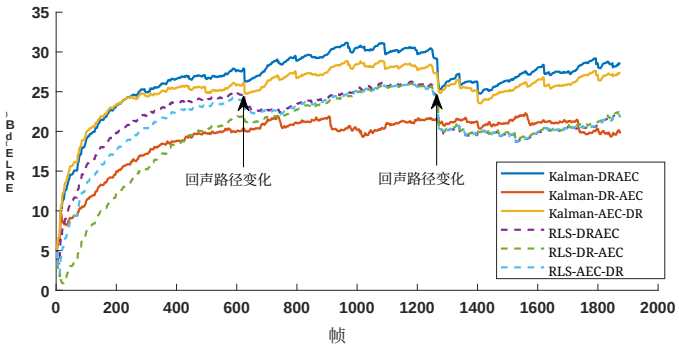


图 2。回声路径变化时的回声损耗增强。彩色显示效果最佳。

4. 实验

实验在有回声、有回声和混响，以及有回声、混响和噪声的环境中进行。不考虑仅有混响的情况，因为如图 1 所示，声源模型会退化为相同的模型。

采样频率为 16 kHz。帧长为 32 ms，帧间重叠率为 50%，短时傅里叶变换长度为 1024 点。声学回声消除滤波器的长度设为 $L_X = 5$，去混响滤波器的长度为 $L_Y = 5$，延迟为 $\Delta = 2$。我们在卡尔曼滤波器中使用 $A = 1$、$\eta = 1e^{-4}$ 和 $\alpha = 0.8$。实验设置主要遵循文献 [18]。我们将所提算法与基于递归最小二乘的变体进行了比较，这些变体的实现已开源[1]。

4.1. 回声

我们首先考虑单讲回声消除任务。使用配备$M = 2$ 个麦克风和一个扬声器的智能音箱录制回声信号。在图 2 中，我们研究了回声返回损耗增强（ERLE），其定义为输入信号功率与输出信号功率之比。通过串接两个不同的测试文件来模拟回声路径变化。Kalman-DRAEC 的稳态性能最高，ERLE 达到 31.15 分贝，比表现第二好的 Kalman-AEC-DR 高出 2.30 分贝，并优于所有递归最小二乘变体。联合DRAEC算法优于级联方案 DR-AEC 和 AEC-DR。其中一个原因是，麦克风信号包含播放声音的回声副本，而对麦克风信号进行滤波有助于抑制回声，尤其是在存在非线性回声时。

通过将纯净语音以 0 dB、-10 dB 和 -20 dB 的信号与回声比（SER）叠加到回声信号上，模拟双讲语音。感知语音质量评估（PESQ）分数经过评估，并报告于表1第一类中。当 SER=-10 dB 和 SER=-20 dB 时，Kalman-DRAEC 总体上取得最高分；在 SER=0 下，其得分与 Kalman-AEC-DR 相当（1.54 对 1.58）

[1]https://github.com/nay0648/unified2021

表1。双讲时段的感知语音质量评估分数。在无混响的情况下，使用纯净语音作为参考信号。

RT60	无混响			0.3s			0.6s		
SER	0dB	-10分贝	-20分贝	0dB	-10分贝	-20分贝	0dB	-10分贝	-20分贝
Orig	1.06	1.07	1.04	1.14	1.07	1.07	1.11	1.08	1.07
Kalman-DRAEC	1.54	**1.24**	**1.13**	1.92	**1.40**	**1.11**	1.57	**1.32**	**1.10**
Kalman-DR-AEC	1.23	1.06	1.09	1.55	1.14	1.08	1.35	1.12	1.09
Kalman-AEC-DR	**1.58**	1.23	1.11	**1.96**	1.39	1.10	**1.59**	1.32	1.10
RLS-DRAEC	1.40	1.19	1.04	1.74	1.34	1.08	1.48	1.28	1.09
RLS-DR-AEC	1.24	1.12	1.04	1.60	1.23	1.07	1.38	1.21	1.08
RLS-AEC-DR	1.39	1.18	1.04	1.74	1.33	1.08	1.48	1.27	1.09

dB.

4.2. 回声与混响

考虑混响时，先将纯净语音与房间脉冲响应进行卷积，然后再与回声信号叠加。脉冲响应是在大小随机的房间中使用镜像法 [21] 生成的。将包含早期混响（50ms）的直达声用作参考信号。双讲语音的 PESQ 分数列于表1右侧。其趋势与回声环境中的情况相同。总体而言，基于卡尔曼滤波器的算法优于基于 RLS 的变体。不建议先去混响再进行声学回声消除，因为如图 2(b) 所示，声源模型会混淆声源信号和回声信号。高混响（0.6s）和高回声级（SER=-20dB）对所有算法来说都具有挑战性。

表2. 相对于输入的信号失真比（分贝）提升。

SER	0dB		-10分贝	
RT60	0.3s	0.6s	0.3s	0.6s
Kalman-DRAEC	9.27	7.68	16.49	13.75
Kalman-DR-AEC	8.90	7.41	14.75	12.25
Kalman-AEC-DR	9.37	**7.78**	16.57	**13.77**
RLS-DRAEC	**9.69**	7.50	**16.82**	13.54
RLS-DR-AEC	9.51	7.32	16.05	12.74
RLS-AEC-DR	9.63	7.43	16.76	13.40

表 3。相对于输入的 SIER（分贝）改善值。

SER	0dB		-10分贝	
RT60	0.3s	0.6s	0.3s	0.6s
Kalman-DRAEC	**11.54**	**9.57**	**12.96**	**10.02**
Kalman-DR-AEC	9.48	7.85	8.48	6.72
Kalman-AEC-DR	11.36	9.38	12.46	9.22
RLS-DRAEC	11.12	9.25	12.09	8.93
RLS-DR-AEC	10.71	8.71	10.40	6.87
RLS-AEC-DR	10.95	9.04	11.90	8.35

4.3. 回声、混响与噪声

所提算法还在干扰与噪声共存的复杂场景中进行了评估。以 0 分贝的信干比（SIR）加入干扰信号。信失比（SDR） [22] 和一项非侵入式指标，即麦克风信号干扰回声比（SIER），分别列于表2和表3中。SDR 衡量处理后的语音的总体质量。SIER 衡量非目标抑制性能。

Kalman-DRAEC 在 SIER 方面具有优势，并且在所有测试案例中均取得最高分；然而，如 SDR 分数所示，这是以更大的语音失真为代价的。根据前述结果，DRAEC 的优势主要来自引入了更强的回声抑制。

5. 结论

本文提出了一种使用卡尔曼滤波器的联合去混响与声学回声消除算法，可同时执行回声消除和语音去混响。该联合算法基于统一的均方误差损失函数推导而来，其性能优于级联 DR-AEC 和 AEC-DR 方法。基于卡尔曼滤波器的算法也优于其基于递归最小二乘的变体，因为前者通过一阶马尔可夫模型对声学环境的变化进行了显式建模。

6. 参考文献

[1] J. J. Shynk 等，“频域与多速率自适应滤波”，*IEEE* 信号处理杂志，第9卷第1期，第14–37页，1992年。[2] T. Yoshioka 和 T. Nakatani，“盲 MIMO 脉冲响应缩短的多通道线性预测方法推广”，*IEEE/ACM* 音频、语音与语言处理汇刊，第20卷第10期，第2707–2720页，2012年。[3] G. Enzner 和 P. Vary，“用于免提电话声学回声控制的频域自适应卡尔曼滤波器”，《信号处理》，第86卷第6期，第1140–1156页，2006年。

[4] C. Paleologu、J. Benesty 和 S. Ciochin˘a，“用于回声消除的通用卡尔曼滤波器研究”，*IEEE/ACM* 音频、语音与语言处理汇刊，第21卷第8期，第1539–1549页，2013年。[5] F. Yang、G. Enzner 和 J. Yang，“突发回声路径变化后快速恢复的频域自适应卡尔曼滤波器”，*IEEE* 信号处理快报，第24卷第12期，第1778–1782页，2017年。[6]D. Yang、F. Jiang、W. Wu、X. Fang 和 M. Cao，“采用神经卡尔曼滤波的低复杂度声学回声消除”，*arXiv* 预印本 *arXiv:2207.11388*，2022年。[7] B. Schwartz、S. Gannot 和 E. A. Habets，“利用卡尔曼滤波器和 EM 算法进行在线语音去混响”，*IEEE/ACM* 音频、语音与语言处理汇刊，第23卷第2期，第394–406页，2014年。[8] S. Braun 和 E. A. Habets，“动态场景下采用自回归模型卡尔曼滤波器的在线去混响”，*IEEE* 信号处理快报，第23卷第12期，第1741–1745页，2016年。[9] ——，“基于线性预测的在线去混响与噪声抑制：采用交替卡尔曼滤波器”，*IEEE/ACM* 音频、语音与语言处理汇刊，第26卷第6期，第1119–1129页，2018年。

[10] J.-M. Lemercier、J. Thiemann、R. Koning 和 T.Gerk- mann，“用于噪声混响环境下稳健在线语音去混响的神经网络增强卡尔曼滤波”，*INTERSPEECH*，2022 年。

[11] R. Takeda、K. Nakadai、T. Takahashi、K. Komatani、T. Ogata 和 H. G. Okuno，“基于独立成分分析的可插话机器人听觉高效盲去混响与回声消除方法”，载于 *ICASSP*。IEEE，2009 年，第 3677–3680 页。

[12] ——，“基于独立成分分析的实际声学信号高效盲去混响与回声消除”，《神经计算》，第 24 卷，第 1 期，第 234–272 页，2012 年。

[13] M. Togami 和 Y. Kawaguchi，“针对相互干扰的声学回声抑制、语音去混响和噪声抑制联合优化”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 22 卷，第 11 期，第 1612–1623 页，2014 年。

[14] G. Carbajal、R. Serizel、E. Vincent 和 E. Humbert，“联合采用神经网络支持的多通道声学回声、混响和噪声抑制”，*IEEE/ACM*《音频、语音与语言处理汇刊》，第 28 卷，第 2158–2173 页，2020 年。

[15] N. Cohen、G. Hazan、B. Schwartz 和 S. Gannot，“一种基于卡尔曼-EM 方法的回声消除、去混响和噪声抑制在线算法”，*EURASIP*《音频、语音与音乐处理期刊》，2021 年，第 1 期，第 33 页，2021 年。[16] Z. Wang、Y. Na、Z. Liu、B. Tian 和 Q. Fu，“面向 AEC-Challenge 的加权递归最小二乘滤波器与神经网络残余回声抑制”，载于*ICASSP*。IEEE，2021 年，第 141–145 页。[17] Z. Wang 、 Y. Na 、 Z. Liu 、 Y. Li 、 B. Tian 和 Q. Fu，“一种用于语音去混响的半盲源分离方法”，*INTERSPEECH*，第 3925–3929 页，2021 年。[18] Y. Na、Z. Wang、Z. Liu、B. Tian 和 Q. Fu，“联合在线多通道声学回声消除、语音去混响与源分离”，载于*INTERSPEECH 2021* 论文集，2021 年，第 1144–1148 页。[19] A. H. Sayed 和 T. Kailath，“自适应递归最小二乘滤波的状态空间方法”，*IEEE* 信号处理杂志，第 11 卷，第 3 期，第 18–60 页，1994 年。[20] R. E. Kalman，“一种线性滤波与预测问题的新方法”，1960年。[21] J. B. Allen 和 D. A. Berkley，“用于高效模拟小房间声学的镜像法”，《美国声学学会杂志》，第 65 卷，第 4 期，第 943–950 页，1979年。[22] E. Vincent、R. Gribonval 和 C. F´evotte，“盲音频源分离中的性能测量”，*IEEE*《音频、语音与语言处理汇刊》，第 14 卷，第 4 期，第 1462–1469 页，2006年。