# 混合会议中联合抑制声学回声与声学啸叫的深度学习方法

张昊、于萌、于栋
腾讯人工智能实验室，贝尔维尤，华盛顿州，美国{aaronhzhang, raymondmyu, dyu}@global.tencent.com

arXiv:2305.01637v2 [eess.AS] 4 May 2023

## 摘要

后新冠时期，混合会议日益成为必要的会议形式，同时也带来了音频相关问题的新挑战。尤其是，混合会议中声学回声与声学啸叫的相互作用，使得对二者进行联合抑制颇具难度。本文提出一种深度学习方法来解决这一问题：采用教师强制训练策略，将循环反馈抑制过程建模为瞬时语音分离任务。具体而言，利用自注意力循环神经网络，从包含可获取及学习得到的参考信号的麦克风录音中提取目标语音，从而同时抑制声学回声和声学啸叫。本文还研究了不同输入信号与损失函数的组合，以提升性能。实验结果表明，所提出的方法能够有效地在混合会议中联合抑制回声与啸叫。

**索引词**：混合会议、声学回声消除、声学啸叫抑制、教师强制训练

## 1. 引言

混合会议结合了现场参与者和远程参与者，在后新冠时代变得日益重要 [1, 2]。截至 2022 年，许多工作场所（78 %）已采用混合办公策略，这表明混合办公正日益成为未来的工作模式[3]。然而，尽管混合会议具有诸多益处，声学回声和声学啸叫等音频问题仍会带来重大挑战，必须加以解决，以确保实现全双工通信。

声学回声是指通信系统一端扬声器发出的声音被另一端的麦克风采集，随后又回放到扬声器中，从而产生令人不快的回声现象 [4, 5]。声学啸叫则是指一端扬声器发出的声音被同一端的麦克风采集，形成反馈回路，使声音不断放大，直至令人难以忍受 [6, 7]。尽管声学回声和声学啸叫的底层机制相似，但它们是不同的问题；在混合会议中，两者可能同时出现，尤其难以应对。因此，必须采用稳健有效的算法，同时处理声学回声消除（AEC）和声学啸叫抑制（AHS），并考虑混合会议环境中的复杂声学环境。然而，一个问题的存在可能影响另一个问题的估计与抑制，使得传统算法难以有效地同时抑制回声和啸叫。

近年来，深度学习凭借建模复杂非线性关系的能力，已成为解决声学回声消除（AEC）和声学啸叫抑制（AHS）挑战的一种有前景的方法 [8, 9, 10, 11,12, 13, 14]。在声学回声消除中，该问题可以直接表述为监督式语音分离问题 [15, 8, 16]。然而，声学啸叫抑制面临着更复杂的挑战，因为它涉及播放信号的递归放大，因此将其表述为监督学习问题并非易事。为应对这一挑战，张等人 [14] 近期提出了一种基于深度学习的声学啸叫抑制方法（深度声学啸叫抑制），采用教师强制训练策略，与基线方法相比取得了更好的性能。我们认为，近年来基于深度学习的声学回声消除和声学啸叫抑制技术的进展，使开发有效的深度学习方法以联合解决这两个问题，并解决混合会议中的全双工通信问题成为可能。

本研究将联合声学回声消除与声学啸叫抑制视为一个整体的反馈抑制问题，以应对其带来的挑战，并提出一种深度学习方法来解决这一问题。通过教师强制训练策略 [17, 18], ，将递归反馈抑制过程转换为语音分离过程，从而简化问题建模并加快模型训练。为实现这一目标，采用自注意力循环神经网络（SARNN） [19]，利用多个参考信号作为额外输入，从麦克风信号中提取目标语音。我们探索了输入的多种组合，以充分利用可获取的参考信号。鉴于联合抑制这两种反馈形式存在困难，我们设计了专门的损失函数，以减轻因反馈抑制不当而引入的泄漏，结果证明该函数有效。实验结果表明，所提出的方法能够有效实现联合回声与啸叫抑制。

本文结构如下：第 2 节概述混合会议系统中的音频相关问题。第 3 节介绍所提出的方法。第 4 节说明实验设置，第 5 节报告相应结果。第 6 节总结全文。

## 2. 混合会议

### 2.1. 混合会议系统中的信号模型

对于一套同一端有 $J$ 台设备的混合会议系统，如果所有设备的扬声器和麦克风都已开启，那么系统中的声学路径总数为 $J^2$。以同一端有两台设备的简化系统为例，如图 1(a) 所示。在采集目标语音 $s_i$ 时，设备 $i$ 上的麦克风还会录到

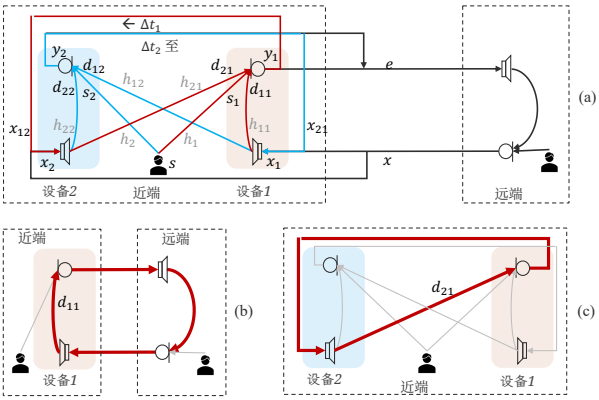



图1：*(a)* 简化的混合会议系统。*(b)* 和 *(c)* 展示了与设备*1*相关的两个闭合声学回路。

背景噪声 $n_i$，以及所有设备的回放信号：

$$y_i = s_i + n_i + \sum_{j=1}^{2} d_{ji} = s_i + n_i + \sum_{j=1}^{2} (x_j * h_{ji}) \quad (1)$$

其中，$x_j$ 是设备 $j$上的扬声器信号，$d_{ji}$ 是麦克风 $i$ 通过声学路径 $h_{ji}$从扬声器 $j$ 拾取的信号。在这些回放信号中，$d_{ii}$ 是从设备 $i$自身扬声器回放到其麦克风的信号，这被称为声学回声。与 $d_{ji}$ （$j \neq i$）相比，声学回声（$d_{ii}$）相对更容易抑制，因为每台设备通常只能获取其自身扬声器信号 $x_i$，该信号可在衰减 $d_{ii}$时用作参考信号。

远端和近端的发言者同时讲话时，会产生挑战。由于每台设备无法判断其他设备是否处于同一空间，因此会将所有其他设备视为远端，并向其发送处理后的信号。此时，扬声器信号 $x_i$ 将是远端信号 $x$ 与设备 $j$ 发送给设备 $i$ 的处理后信号（记为 $x_{ji}$，（$j \neq i$））的组合：

$$x_i = x + x_{ji}, j \neq i \quad (2)$$

如果每台设备上的反馈抑制模块都能正常工作，那么处理后的信号 $x_{ji}$ 应类似于近端语音 $s$ 的延迟、缩放和混响版本。从信号源的角度来看，式（1）中的麦克风信号可以改写为：

$$y_i = s_i + n_i + \sum_{j=1}^{2} d_{ji}^{x} + \sum_{j=1}^{2} d_{ji}^{s} \quad (3)$$

其中，$d_{ji}^{x}$ 和 $d_{ji}^{s}$ 分别表示由 $x$ 和 $s$ 产生的播放分量。抑制 $d_{ji}^{s}$ 更具挑战性，因为它与目标语音 $s_i$ 来自同一信号源，减少它可能会使目标信号失真。

### 2.2. 声学回声与声学啸叫联合抑制

让我们聚焦于设备1，分析混合会议系统中与音频相关的问题。系统中的每台设备都有两个闭合声学回路（CAL），如图1（b）和（c）所示，这些回路可能导致声学啸叫。第一个闭合声学回路由声学回声引起，较容易处理，因为声学回声在每次

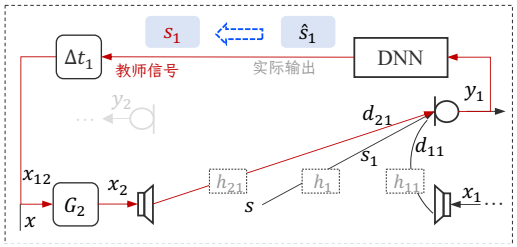



图2：采用教师强制学习策略的信号流。

传输中只出现一次，且在两端都会得到处理。第二个闭合声学回路更具挑战性，原因有二：1）设备1无法获取导致反馈的参考信号 $d_{21}$。2）参与该闭合声学回路的两台设备处于同一空间。

未经任何处理时，麦克风信号会通过扬声器播放出来，并反复进入拾音端。时刻索引为 $t$ 时的麦克风信号 $y_1$ 可表示为：

$$\begin{aligned} y_1(t) =& s_1(t) + n_1(t) + d_{11}(t) + \\ & NL\left[(y_1(t-\Delta t_1) + x(t)) \cdot G_2\right] * h_{21}(t) \end{aligned} \quad (4)$$

其中，$\Delta t_1$ 表示从设备1到设备2的系统延迟，$G_2$ 是设备2上放大器的增益，$NL(\cdot)$ 表示扬声器的非线性函数。播放信号 $d_{11}(t)$ 即为声学回声。$y_1(t)$ 与 $y_1(t-\Delta t)$ 之间的递归关系会导致播放信号被再次放大，产生令人烦恼的尖锐声音，这种现象称为声学啸叫。

在混合会议中，要实现全双工通信，必须同时解决声学回声消除和声学啸叫抑制问题。然而，任一问题的存在都可能妨碍对另一问题的准确检测与消除，导致缺乏有效的解决方案。

## 3. 提出的方法

### 3.1. 问题表述

为解决啸叫的递归特性，需要将深度神经网络模块集成到闭合声学回路中并进行递归训练。然而，由于计算成本高，这种做法并不实际。另一种方法是采用教师强制训练策略，将联合声学回声消除与声学啸叫抑制任务表述为一般的反馈抑制问题，如图2所示。该方法基于这样的假设：模型经过适当训练后，能够衰减所有反馈信号（$d_{11}$ 和 $d_{21}$），并且只传输目标语音 $s_1$。通过教师强制学习，在模型训练期间，实际输出 $\hat{s}_1$ 会被教师信号 $s_1$ 替代。因此，麦克风信号 (4) 不再是递归生成的，而是简化为目标信号、背景噪声、声学回声和由 $s_1$确定的一次性播放信号的混合：

$$\begin{aligned} y_1(t) =& s_1(t) + n_1(t) + d_{11}(t) + \\ & NL[(s_1(t-\Delta t_1) + x(t)) \cdot G_2] * h_{21}(t) \end{aligned} \quad (5)$$

因此，整个问题可以表述为模型训练期间的语音分离问题，任务是利用可获取的扬声器信号（$x_1$和/或 $x$，$x_{21}$）作为参考，从麦克风录音中分离出目标信号。

### 3.2. 输入与参考信号

适当的参考信号对于声学回声消除和声学啸叫抑制算法至关重要，因为它们能够准确估计回放信号。如图1和公式（2）所示，设备1的参考信号是两个信号的混合。最直接的方法是使用集成信号 $x_1$ 作为参考，以抑制 $y_1$中的两个反馈信号 $d_{11}$ 和 $d_{21}$ 。不过，这种方法抑制 $d_{21}$的效果可能较差。由公式（3）可知，回放信号包含来自不同声源的共同成分。根据音频系统的设计，除了集成扬声器信号 $x_1$之外，我们还可能获取 $x$ 和 $x_{21}$ 。使用分离后的扬声器信号（$x$ 和 $x_{21}$）作为参考，可以更有效地抑制两路反馈。

除了这些可直接从设备获取的参考信号外，我们还设计了网络，使其能够从输入中估计一些中间输出，并将其用作非线性参考信号，进一步提升反馈消除性能 [9, 20]。

### 3.3. 网络结构

所提方法的网络如图3所示。它将麦克风信号和一个或两个参考信号（分别表示为 $r_1$ 和 $r_2$）作为输入。输入信号以16千赫兹采样，并通过帧长为32毫秒、帧移为16毫秒的512点短时傅里叶变换（STFT）转换到频域。所得频域输入分别标记为 $Y$、 $R_1$和 $R_2$。

为了从输入中提取更多信息并便于抑制回放信号，我们遵循 [19, 14] ，将输入特征设计为以下内容的拼接：归一化对数功率谱（LPS）、跨时间帧和频率频点的相关矩阵，以及输入信号的通道协方差。随后，将这些特征拼接并输入线性层进行特征融合，再经过一个包含257个隐藏单元的门控循环单元（GRU）层和三个一维卷积层，以估计三个复值滤波器。然后，通过深度滤波 [21] 将这些滤波器应用于输入，以获得相应的中间信号 $\tilde{Y}$、 $\tilde{R}_1$和 $\tilde{R}_2$。这些信号作为额外的非线性参考信号，其LPS随后与原始融合特征拼接，并通过另一个线性层进行特征融合。

接下来，使用一个 SARNN模块来估计四通道增强滤波器随后应用于麦克风信号和三个学习得到的参考信号，以获得增强后的目标信号 $\hat{S}_1$。最后，使用逆短时傅里叶变换（iSTFT）获得波形 $\hat{s}_1$。有关特征设计和网络结构的更多详情，请参阅 [19]。

### 3.4. 损失函数

在本研究的初始阶段，模型训练采用时域尺度不变信号失真比（SI-SDR） [22] 损失与频域频谱幅度平均绝对误差（MAE）的组合，作为损失函数：

$$Loss_1 = -\text{SI-SDR}(\hat{s}, s) + \lambda \text{MAE}(|\hat{S}|, |S|) \quad (6)$$

鉴于反馈信号与目标信号高度相关，抑制它们可能很困难。为进一步抑制因回放信号衰减不足而引入的泄漏，我们提出加入相关性损失：

$$Loss_{corr} = [1 - corr(\hat{s}_1, s_1)] + corr(\hat{s}_1 - s_1, d_*) \quad (7)$$

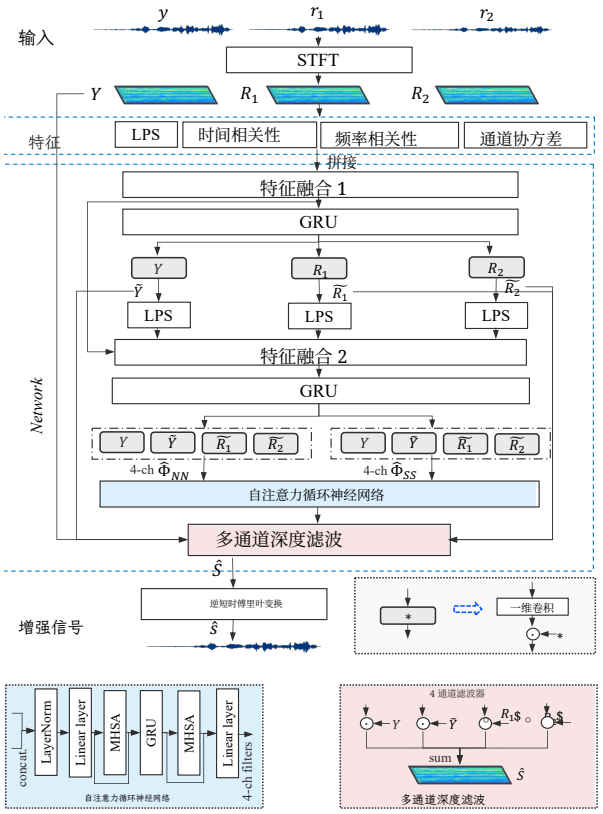



图3： 用于联合抑制声学回声和啸叫的深度神经网络模块架构。其中，灰色方块表示一维卷积层与深度滤波的组合。“一维卷积”输出一个复值比率滤波器，该滤波器随后通过深度滤波作用于信号 $*$，记作 $\odot$。

相关性损失由两项组成。第一项评估估计信号与目标信号之间的相似度，第二项衡量回放信号 $d_*$ 与估计目标中的残差信号之间的相似度。我们用于模型训练的修正损失函数为：

$$Loss_2 = Loss_1 + \beta Loss_{corr} \quad (8)$$

为确保不同损失之间保持平衡，我们在实现中分别将 $\lambda$ 和 $\beta$ 设为 10000 和 10。

## 4. 实验设置

### 4.1. 数据准备

实验使用 AISHELL-2 [23] 和 INTERSPEECH 2021 AEC 挑战赛 [24] 数据集。利用镜像法 [25]，生成了总计 10,000 组房间脉冲响应（RIR），其中纳入了随机的房间特性和 0 至 0.6 秒的混响时间（RT60）范围。每组房间脉冲响应包含 6 个房间脉冲响应，如图1所示。在数据生成过程中，随机选取一组房间脉冲响应，用于生成近端语音信号及对应的回放信号。系统延迟定义为 $[0.1, 0.3]$ 秒范围内的随机值；麦克风非线性失真则通过采用硬削波和 Sigmoid 函数 [26, 15, 8] 的饱和型非线性进行模拟。麦克风信号由一个随机选

表1： 关于输入/参考信号的探索 *ls.*

| $Net_{2-ch}$, $Loss_1$ | 尺度不变信号失真比（dB） | | | PESQ | | |
|---|---|---|---|---|---|---|
| 信号反馈比（dB） | -10 | -5 | 0 | -10 | -5 | 10 |
| 未处理 | -9.49 | -4.47 | 0.54 | 1.35 | 1.64 | 2.05 |
| $[y_1, x_1]$ | 4.59 | 7.78 | 10.72 | 2.26 | 2.59 | 2.88 |
| $[y_1, x]$ | 3.24 | 6.40 | 9.28 | 2.06 | 2.38 | 2.71 |
| $[y_1, x_{21}]$ | 4.67 | 7.69 | 10.43 | 2.55 | 2.88 | 3.12 |
| $[y_1, x, x_{21}]$ | 5.25 | 8.21 | 10.98 | 2.58 | 2.89 | 3.11 |
| $[y_1, x_{21}, x]$ | **5.31** | **8.53** | **11.42** | **2.69** | **2.99** | **3.23** |

表2： 关于损失函数的探索。

| $Net_{2-ch}$, $[y_1, x_1]$ | | 尺度不变信号失真比（dB） | | | PESQ | | |
|---|---|---|---|---|---|---|---|
| 信号反馈比（dB） | | -10 | -5 | 0 | -10 | -5 | 10 |
| 未处理 | | -9.49 | -4.47 | 0.54 | 1.35 | 1.64 | 2.05 |
| $Loss_1$ | | 4.59 | 7.78 | 10.72 | 2.26 | 2.59 | **2.88** |
| $Loss_2$ | $d^s_{21}$ | 3.95 | 6.58 | 8.52 | 2.13 | 2.43 | 2.67 |
| | $d_{21}$ | **4.96** | **7.92** | **10.83** | **2.27** | **2.60** | **2.88** |
| | $d^s_{21}$+$d^s_{11}$ | 4.32 | 6.98 | 9.03 | 2.26 | 2.56 | 2.79 |
| | $d_{21} + d_{11}$ | 4.74 | 7.65 | 10.32 | 2.27 | 2.58 | 2.86 |

取的信号反馈比（SFR）（范围为 $[-20, 5]$ dB）以及信噪比（SNR）（范围为 -10 dB 至 30 dB）混合生成。我们分别创建了 10,000、300 和 500 条话语，用于训练、验证和测试。用于生成测试数据的话语和房间脉冲响应与训练和验证数据所用的不同。模型训练了 60 个训练轮次，批次大小为 20。

### 4.2. 方法评估

采用尺度不变信号失真比和语音质量感知评估（感知语音质量评估）[27] 作为评估指标，用于展示播放衰减性能以及增强语音的质量。数值越高，表示性能越好。

## 5. 实验结果

### 5.1. 关于输入/参考信号的探索

为了研究参考信号对模型性能的影响，我们针对不同的信号反馈比（SFR）进行了实验，结果汇总于表1。为减小比较过程中模型规模差异的影响，我们从网络中移除了与 $R_2$ 相关的˜深度滤波分支，并使用由此得到的简化网络（记为“$Net_{2-ch}$”）对具有2或3个通道的输入进行训练。这意味着，使用三通道输入时，我们从前两个通道中提取中间信号，而第三个通道仅用于特征提取。使用集成扬声器信号 $x_1$ 作为参考信号是训练模型最直接的方法，使用 $[y_1, x_1]$ 训练得到的模型称为“初始”模型。在双通道输入的模型中，“初始”模型在大多数情况下取得了更好的尺度不变信号失真比（SI-SDR），而其语音质量（感知语音质量评估，PESQ）并不优于使用 $x$

$_{21}$ 作为参考信号。使用分离的参考信号（三通道输入）训练的模型，表现始终优于使用集成参考信号的模型 $x_1$。对于$Net_{2-ch}$，参考信号 $x$ 和 $x_{21}$ 的顺序决定从哪个 $\tilde{R}_1$ 中提取信息。以 $[y_1, x_{21}, x]$ 作为输入训练的模型取得了整体最佳表现。

### 5.2. 关于损失函数的探索

表 2 比较了使用$Net_{2-ch}$、输入 $[y_1, x_1]$，以及使用不同方法计算的 $Loss_2$ 训练的模型性能

表3： 用于反馈抑制的所提方法。

| 输入： $[y_1, x_{21}, x]$ | 尺度不变信号失真比 | | | PESQ | | |
|---|---|---|---|---|---|---|
| 信号反馈比 (dB) | -10 | -5 | 0 | -10 | -5 | 0 |
| 未处理 | -9.49 | -4.47 | 0.54 | 1.35 | 1.64 | 2.05 |
| $Net_{2-ch}$, $Loss_1$ | 5.31 | 8.53 | 11.42 | 2.69 | 2.99 | 3.23 |
| $Net_{3-ch}$, $Loss_1$ | 6.11 | 9.09 | 11.82 | **2.73** | **3.07** | **3.31** |
| $Net_{3-ch}$, $Loss_2$ | **6.47** | **9.24** | **11.87** | **2.73** | 3.03 | 3.28 |

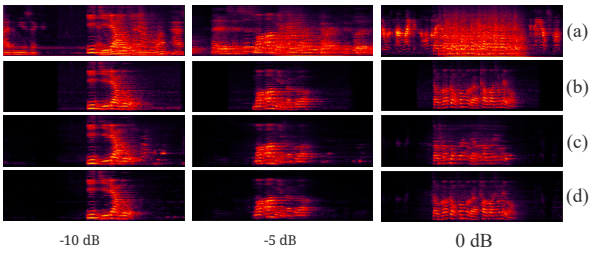



图4： 不同信号反馈比水平下获得的处理后的信号频谱图：*(a)* 未处理信号，*(b)* 目标信号，*(c)*“初始”模型的输出，以及 *(d)* 最佳模型的输出。

播放 $d_*$。结果表明，加入相关性损失并不总能提升性能，而基于 $d_{21}$ 计算的 $Loss_2$ 能取得最佳性能，且优于使用 $Loss_1$训练的模型，尤其是在尺度不变信号失真比方面。这是因为缺乏直接参考信号， $d_{21}$ 更难以抑制；在计算 $Loss_2$ 时使用它有助于进一步减轻泄漏。

### 5.3. 联合声学回声消除与声学啸叫抑制的所提方法

我们结合对输入和损失函数进行探索所得的结果，训练模型以实现最佳的反馈抑制。具体而言，我们采用包含 $d_{21}$ 的 3 通道输入 $[y_1, x_{21}, x]$、 $Loss_2$ ，以及图3所示的网络（记为“$Net_{3-ch}$”）进行模型训练。表3所示的性能比较结果表明，使用“$Net_{3-ch}$”比使用“$Net_{2-ch}$”能够取得更好的性能，而采用修改后的损失函数 $Loss_2$ 则可进一步提升播放衰减性能。我们还在图4中给出了使用“初始”模型和性能最佳模型所获得的处理后的信号频谱图，以进一步说明我们提出的联合声学回声与声学啸叫抑制方法的有效性。

## 6. 结论

我们提出了一种基于深度学习的方法，用于解决混合会议中的音频相关问题。我们提出的方法将声学回声和声学啸叫视为一个综合反馈问题，并利用教师强制学习策略同时实现声学回声消除与声学啸叫抑制。通过将递归反馈抑制问题转换为语音分离问题，采用 SARNN 模型从麦克风录音中提取目标语音，并将多个参考信号作为额外输入。我们研究了输入信号和损失函数对联合声学回声消除与声学啸叫抑制性能的影响。未来的工作包括考虑计算复杂度等实际问题，并研究使用级联网络逐步抑制声学回声和声学啸叫。

# 7. 参考文献


[1] B. Saatc¸i、R. R¨adle、S. Rintel、K. O'Hara 和 C. Nyland- sted Klokmose，“现代职场中的混合会议：成功与失败的故事”，载于《协作技术与社会计算：第25届国际会议》，CRIWG+Col- labTech 2019，日本京都，2019年9月4日至6日，论文集第25卷。施普林格，2019年，第45–61页。[2] B. Z. Hameed、Y. Tanidir、N. Naik、J. Y.-C. Teoh、M. Shah、M. L. Wroclawski、A. B. Kunjibettu、D. Castellani、S. Ibrahim、R. D. da Silva 等，“‘混合’会议会在 COVID-19 时代之后取代面对面会议吗？泌尿学界的看法与观点”，《泌尿学》，第156卷，第52–57页，2021年。[3]R. Carter，“什么是混合会议？简介”，https://www.uctoday.com/collaboration/what-is-a-hybrid-meeting-an-introduction/。[4] J. Benesty、T. G¨ansler、D. R. Morgan、M. M. Sondhi、S. L. Gay 等，“网络与声学回声消除的进展”，2001年。[5] G. Enzner、H. Buchner、A. Favrot 和 F. Kuech，“声学回声控制”，载于《学术出版社信号处理文库》。Elsevier，2014年，第4卷，第807–877页。[6] R. V. Waterhouse，“混响室中的回授啸叫理论”，《美国声学学会杂志》，第37卷，第5期，第921–923页，1965年。[7] T. Van Waterschoot 和 M. Moonen，“声学反馈控制五十年：现状与未来挑战”，《IEEE汇刊》，第99卷，第2期，第288–327页，2010年。[8] H. Zhang 和 D. Wang，“嘈杂和双讲场景下的声学回声消除深度学习”，《Interspeech 2018》论文集，第3239–3243页，2018年。[9] H. Zhang、S. Kandadai、H. Rao、M. Kim、T. Pruthi 和 T. Krist- jansson，“深度自适应声学回声消除：深度学习与自适应声学回声消除的混合方法”，载于《ICASSP 2022-2022 IEEE声学、语音与信号处理国际会议（ICASSP）》。IEEE，2022年，第756–760页。[10] H. Zhang 和 D. Wang，“深度多通道主动噪声控制：一种多通道主动噪声控制的深度学习方法”，《神经网络》，第158卷，第318–327页，2023年。[11] Z. Chen、Y. Hao、Y. Chen、G. Chen 和 L. Ruan，“一种面向实时通信应用的基于神经网络的啸叫检测方法”，载于《2022 IEEE声学、语音与信号处理国际会议（ICASSP）》。IEEE，2022年，第206–210页。[12] H. Gan、G. Luo、Y. Luo 和 W. Luo，“利用深度神经网络在时频域中进行啸叫噪声消除”，载于《第六届信息与通信技术国际大会论文集》。施普林格，2022年，第319–332页。[13] C. Zheng、M. Wang、X. Li 和 B. C. Moore，“助听器声学反馈系统边际稳定性问题的深度学习解决方案”，《美国声学学会杂志》，第152卷，第6期，第3616–3634页，2022年。[14] H. Zhang、M. Yu 和 D. Yu，“深度声学啸叫抑制：一种声学啸叫抑制的深度学习方法”，arXiv 预印本 arXiv:2302.09252，2023年。[15] C. M. Lee、J. W. Shin 和 N. S. Kim，“基于深度神经网络的残余回声抑制”，载于《国际语音通信协会第十六届年会》，2015年。[16] H. Zhang、K. Tan 和 D. Wang，“具有非线性失真的联合声学回声与噪声消除的深度学习”，载于《Interspeech》，2019年，第4255–4259页。[17] R. J. Williams 和 D. Zipser，“一种用于持续运行的全循环神经网络的学习算法”，《神经计算》，第1卷，第2期，第270–280页，1989年。[18] A. M. Lamb、A. G. ALIAS PARTH GOYAL、Y. Zhang、S. Zhang、A. C. Courville 和 Y. Bengio，“教授强制：一种训练循环网络的新算法”，《神经信息处理系统进展》，第29卷，2016年。

[19] M. Yu、Y. Xu、C. Zhang、S.-X. Zhang 和 D. Yu，“NeuralEcho：用于统一声学回声抑制与语音增强的自注意力循环神经网络”，*arXiv* 预印本 *arXiv:2205.10401*，2022。[20] Y. Zhang、M. Yu、H. Zhang、D. Yu 和 D. Wang，“KalmanNet：一种用于声学回声消除的可学习卡尔曼滤波器”，*arXiv* 预印本 *arXiv:2301.12363*，2023。[21] W. Mack 和 E. A. Habets，“深度滤波：使用复数时频滤波器进行信号提取与重建”，*IEEE* 信号处理快报，第 27 卷，第 61–65 页，2019。[22] J. Le Roux、S. Wisdom、H. Erdogan 和 J. R. Hershey，“SDR——半生不熟还是恰到好处？”，载于*2019*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019，第 626–630 页。[23] J. Du、X. Na、X. Liu 和 H. Bu，“AISHELL-2：将普通话自动语音识别研究推向工业级规模”，*arXiv* 预印本 *arXiv:1808.10583*，2018。[24] R. Cutler、A. Saabas、T. P¨arnamaa、M. Loide、S. Sootla、M. Purin、H. Gamper、S. Braun、K. Sørensen、R. Aichner 等，“INTERSPEECH 2021声学回声消除挑战赛”。载于*Interspeech*，2021，第 4748–4752 页。[25] 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。[26] A. Birkett 和 R. Goubran，“用于免提声学回声消除的扬声器非线性补偿”，电子快报，第 32 卷，第 12 期，第 1063–1064 页，1996。[27] A. W. Rix、J. G. Beerends、M. P. Hollier 和 A. P. Hekstra，“语音质量的感知评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法”，载于*2001*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 2 卷。IEEE，2001，第 749–752 页。