# 编码器–时序建模结构的时延估计及在回声抵消中的应用*

刘 杨[1,2] 杨飞然[1,2] 杨 军[1,2†]

(1 中国科学院噪声与振动重点实验室 (声学研究所) 北京 100190)
(2 中国科学院大学 北京 100049)

2022 年 7 月 22 日收到
2022 年 11 月 22 日定稿

**摘要** 提出了一种使用编码器–时序建模结构的时延估计方法来估计声学回声抵消中传声器信号相对远端信号的时延。该方法以短时傅里叶变换域的远端信号和传声器信号作为输入特征, 通过复数卷积神经网络构成的编码器提取带有相位信息的高维特征, 利用循环神经网络学习两输入信号之间的时延关系, 构建了从信号到时延的映射。仿真实验结果表明, 相比 WebRTC-DE 和 GCC-PHAT, 所提方法的优势有: (1) 模型的参数量和计算量不受时延长度影响; (2) 有效缩短了时延估计的收敛时间和跟踪时间; (3) 在长混响和双端对讲的情况下具有更小、更稳定的估计误差和标准差。将使用编码器–时序建模结构的时延估计方法与自适应回声抵消级联的实验验证了新方法的有效性。

**关键词** 回声抵消, 时延估计, 深度学习, 编码器–时序建模
**PACS 数** 43.60, 43.72 **DOI**：10.12395/0371-0025.2022045

# Delay estimation using encoder-temporal modeling structure for acoustic echo cancellation

LIU Yang[1,2] YANG Feiran[1,2] YANG Jun[1,2†]

(1 *Key Laboratory of Noise and Vibration Research*, *Institute of Acoustics*, *Chinese Academy of Sciences* Beijing 100190)
(2 *University of Chinese Academy of Sciences* Beijing 100049)

Received Jul. 22, 2022
Revised Nov. 22, 2022

**Abstract** A delay estimation method based on encoder-temporal modeling structure is proposed to estimate the delay of microphone signal relative to the far-end signal in acoustic echo cancellation. In the proposed method, the far-end signal and the microphone signal in the short-time Fourier transform domain are used as input features. High-dimensional features with phase information are extracted by an encoder composed of complex convolutional neural networks. The memory ability of recurrent neural network is used to learn the time delay relationship between two input signals. A mapping from signal to delay is constructed by the proposed method. The simulation results show that the proposed method has the following advantages over WebRTC-DE and GCC-PHAT: (1) the number of parameters and computational complexity of the model are not affected by the delay; (2) the convergence time and tracking time of delay estimation are effectively reduced; (3) smaller and more stable estimation error and standard deviation are achieved in the case of long reverberation time and double-talk. Experiments on adaptive echo cancellation cascaded with the proposed delay estimation module verify the effectiveness of the new method.

**Keywords** Acoustic echo cancellation, Time-delay estimation, Deep learning, Encoder-temporal modeling

* 国家自然科学基金项目 (62171438)、北京市自然科学基金-小米创新联合基金项目 (L223032)、中国科学院声学研究所自主部署“前沿探索”类项目 (QYTS202111) 资助。
† 通讯作者: 杨军, jyang@mail.ioa.ac.cn

# 引言

声学回声抵消广泛应用于语音通信和人机语音交互等方面, 旨在利用远端信号作为参考来消除近端传声器信号中的回声干扰[1–3]。回声抵消的解决方案包括自适应滤波、混合模型和端到端模型。自适应滤波通过辨识扬声器与传声器之间的回声路径来估计回声信号, 并从传声器信号中减去估计的回声得到近端信号[4]。为了进一步消除自适应滤波器收敛和功放、扬声器的非线性带来的残余回声, 混合模型在利用自适应滤波器消除线性回声后, 加入深度神经网络 (DNN) 作为后处理模块[5–6]。端到端的解决方案则直接利用远端信号和传声器信号经过深度神经网络预测干净的近端信号, 可以将回声抵消、去混响和降噪任务联合处理[7]。

一个典型的回声路径脉冲响应包含无声平坦区域和声学活动区域两部分[8]。无声平坦区域的长度由回声信号相对远端信号的时延决定, 声学活动区域的长度由房间混响时间决定。时延的主要来源包括网络传输、数模或模数转换、语音编解码和信号预处理等。由于网络传输的不稳定性, 远端信号与回声信号之间的时延问题变得更加严重, 在较差的条件下甚至能达到 1.5 s 的时延[9]。另外, 网络设备的不稳定还会带来抖动缓冲延迟 (JBD), 即时延是时变的。时变的长时延会带来计算量和模型参数的浪费, 使得回声抵消算法的性能下降。在使用自适应滤波的回声抵消中, 时变的长时延会使待估计的滤波器变长, 导致计算复杂度的增加和收敛跟踪性能的下降。在使用深度学习的回声抵消中, 时变的长时延使得模型花费更多的参数学习回声信号相对远端信号的时延, 用于回声抵消的参数变少, 从而导致回声抵消量的下降。

为解决上述问题, 需要在回声抵消前增加一个时延估计器。时延估计器首先估计远端信号与回声信号之间的时延, 然后利用估计的时延对齐远端信号[10–11]。时延估计方法主要有 3 种。第 1 种是广义互相关 (GCC-PHAT) 方法, 它在噪声环境下具有较好的鲁棒性[12]。但是在时延较大时, GCC-PHAT 的计算复杂度极高, 因而对于大时延估计而言优化 GCC-PHAT 并不是经济的选择。第 2 种是 WebRTC 中的时延估计方法 (WebRTC-DE), 它对远端信号和传声器信号进行帧级别的能量对比, 根据实验测得的时延更新概率来估计时延[13]。该方法能够实时地估计时延, 且在仅有回声的情况下具有较好的性能, 但在长混响和双端对讲场景下性能不稳定。第 3 种是参数模型法, 该方法通常先对远端信号和传声器信号进行降采样, 再利用自适应滤波器估计两个信号之间的传递函数, 进而得到时延值[5]。该方法思路简明, 但自适应滤波在非线性回声、双端对讲的情况下会出现性能下降[14]。

近年来, 编码器–时序建模–解码器的结构在深度学习语音增强、回声抵消中表现优异[14–16]。在到达时间差估计中, 利用深度学习对 GCC-PHAT 算法的优化也在仿真数据集上取得了更精确的估计结果[17–19]。基于以上讨论, 本文提出了一种使用编码器–时序建模结构的时延估计方法, 用于估计回声抵消中的时延。由频域分块自适应滤波器使用过去帧的形式的启发, 将时延估计设计为帧级的分类任务。该方法以短时傅里叶变换域的远端信号和传声器信号作为输入特征, 通过复数卷积神经网络构成的编码器提取带有相位信息的高维特征, 利用循环神经网络 (RNN) 的时序记忆能力学习两输入信号间的时延关系, 最后用全连接层进行分类[20]。将所提方法与 GCC-PHAT、WebRTC-DE 进行比较, 所提方法有效缩短了时延估计的收敛时间和跟踪时间, 且受混响时间和双端对讲的影响较小。在时延估计与回声抵消算法级联的实验中, 所提方法也有效提高了回声抵消量和近端语音质量。

# 1　模型描述

## 1.1　回声抵消中的时延估计

具有时延估计的回声抵消系统如图 1 所示。回环系统[9] 采集扬声器播放前的信号作为远端信号, 故扬声器所播放的信号相对远端信号有延迟。$n$时刻扬声器播放信号为

$$\widetilde{x}(n) = f[x(n-\Delta_1)], \tag{1}$$

其中, $x(n)$为远端信号, $f[\cdot]$为功放和扬声器等对信号的处理, $\Delta_1$为由回环系统、功放、扬声器等硬件引起的时延。扬声器播放出的信号经过房间反射传播到传声器, 得到回声信号。$n$时刻传声器接收信号为

$$y(n) = d(n) + e(n) = \widetilde{x}(n) * h(n-\Delta_2) + e(n), \tag{2}$$

其中, $d(n)$和$e(n)$为传声器接收到的回声和近端信号, 近端信号中包含近端语音 $s(n)$和噪声 $v(n)$, $h(n)$为从扬声器到传声器的回声路径, $\Delta_2$为物理距离、硬件接收和软件预处理引起的时延, $*$为卷积操作。根据卷积的性质, $d(n)$还可表示为

$$d(n) = f[x(n)] * h(n-\Delta) = f[x(n-\Delta)] * h(n), \tag{3}$$

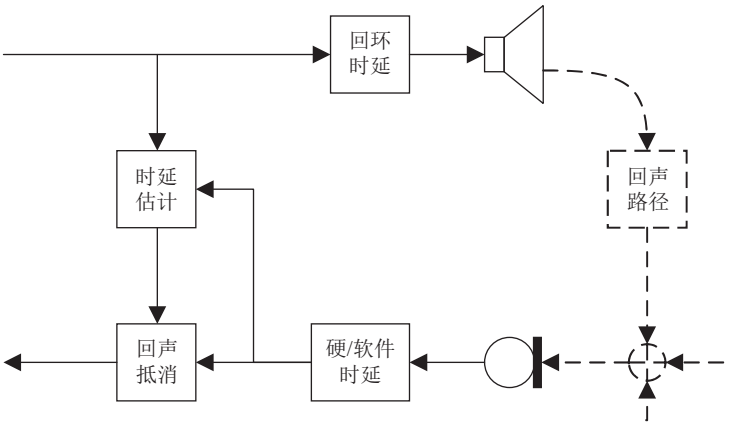


图 1 具有时延估计的回声抵消系统

其中, $\Delta=\Delta_1+\Delta_2$表示系统中从回环采集到算法处理前的总时延。首先对时延$\Delta$进行估计, 然后对远端信号进行延时操作$x_\Delta(n)=x(n-\Delta)$, 最后利用延时后的远端信号$x_\Delta(n)$与传声器接收信号$y(n)$进行回声抵消, 得到估计的近端信号$\widehat{e}(n)$。

### 1.2 基于深度学习的时延估计模块

现在给出本文所搭建的时延估计系统。正如前面所述, 基于自适应滤波器和基于深度学习的回声抵消都需要时延估计, 但是它们对时延估计的具体要求有所不同。具体说, 时延过估计将导致一个非因果的系统, 这会使得基于自适应滤波的回声抵消算法性能严重下降。基于深度学习的回声抵消网络由于其模式识别的能力, 能够容忍较小的时延过估计, 而对于估计误差的容忍能力则与训练集包含的时延信息有关。本文主要针对基于自适应滤波的回声抵消设计时延估计方法。该方法进行一些调整后也可用于基于深度学习的回声抵消系统, 但这不在本文的讨论范围内。将时延估计设计为分类任务, 利用 RNN 的记忆能力从训练集中学习传声器信号$y(n)$相对于远端信号$x(n)$的时延。

图 2 给出了深度学习时延估计模块的框图。时延估计网络由编码器、时序建模层和全连接层构成。编码器呈 Y 字形, 远端信号和传声器信号先经过 1 层编码层, 实部虚部分别拼接后一起传入 3 层编码层, 得到提取到的复数特征。每层编码层包含二维复数卷积 (Complex Conv2d) 层、复数批归一化 (Complex BN) 层和复数 PReLU (Complex PReLU) 层[15]。为了减小参数量, 使用门控循环单元 (GRU)[21] 作为时序层的组成单元, 设计了类似复数长短期记忆网络结构[15] 的复数 GRU (Complex GRU)。时序层包含 1 层复数 GRU 和 5 层 GRU, 其中复数 GRU 可以充分利用相位信息来提取时延特征。实数 GRU 比

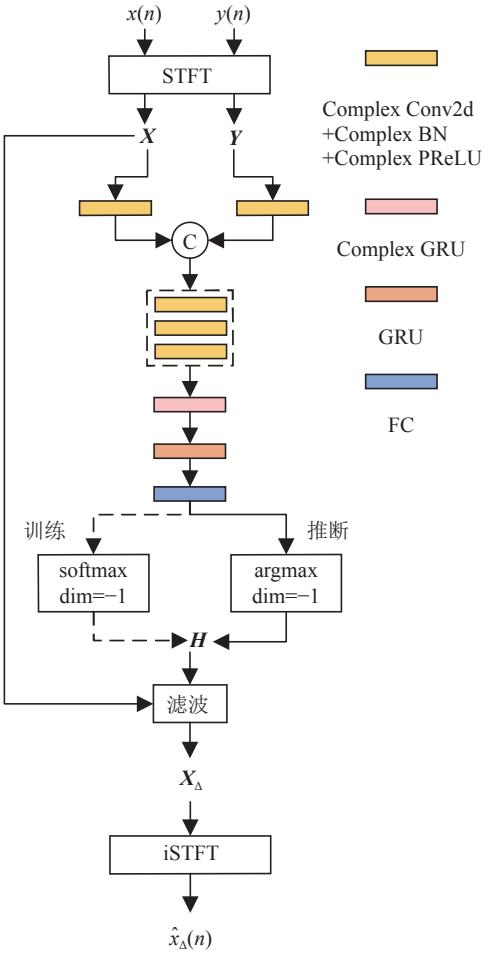


图 2 深度学习时延估计模块

复数 GRU 更容易训练, 在时序建模层中占比更大[22]。时序层后再接 1 层全连接 (FC) 作为输出层。输出特征的最后一维表示各类别的分类概率。

远端信号$x(n)$和传声器信号$y(n)$经过短时傅里叶变换 (STFT) 后得到时延估计网络的输入特征$\boldsymbol{X}\in\mathbb{R}^{2\times T\times(F/2+1)}$和$\boldsymbol{Y}\in\mathbb{R}^{2\times T\times(F/2+1)}$, 其中实部和虚部被拆成两个通道, $T$为计算当前帧所需时间帧数, 本文采用过去帧和当前帧共计$T=17$帧, $F$为傅里叶变换点

数。时延估计网络从输入特征中估计出时延滤波器 $\boldsymbol{H}\in\mathbb{R}^{(L+1)\times(F/2+1)}$, 其中 $L$ 表示可估计的最大时延帧数。所设计时延估计模块的最大可估计时延为 $LM/f_s$, 其中 $f_s$ 表示采样频率, STFT 的帧移为 $M$ 个采样点。使用时延滤波器 $\boldsymbol{H}$ 对远端信号 $\boldsymbol{X}$ 进行滤波, 第 $l$ 个时间帧第 $k$ 个频率块的对齐后的远端信号为

$$\widehat{X}_{\Delta}(l,k)=\boldsymbol{h}^{\mathrm{T}}(l,k)\boldsymbol{x}(l,k)=\sum_{i=0}^{L}H_i(l,k)X(l-i,k),\tag{4}$$

其中, $\boldsymbol{x}(l,k)=[X(l-L,k),\cdots,X(l,k)]^{\mathrm{T}}$ 是长度为 $L+1$ 的远端信号向量, $\boldsymbol{h}(l,k)=[H_L(l,k),\cdots,H_0(l,k)]^{\mathrm{T}}$ 是第 $l$ 个时间帧第 $k$ 个频率块对应的时延滤波器。各频率块之间的时延是一致的, 则每帧信号所需估计的时延滤波器可以省略表示为 $\boldsymbol{h}(l)=[H_L(l),\cdots,H_0(l)]^{\mathrm{T}}$, 所需估计的时延滤波器可降维为 $\boldsymbol{H}=\boldsymbol{h}(l)\in\mathbb{R}^{(L+1)}$。$H_i(l)$ 表示时延为 $i$ 个时间帧的概率, 因此在设置训练便签时理想的时延滤波器仅在真实时延对应的时间帧数 $I$ 的位置处有 $H_I(l)=1$, 其余位置为 $H_i(l)=0$。

将时延估计设计为一个分类任务时, 最大可估计时延帧数 $L$ 就等于总的分类数。采用加权的交叉熵 (WCE)[20] 作为损失函数:

$$\mathrm{WCE}=-\sum_{i=0}^{L}w_i c_i\log_{10}(\widehat{c}_i),\tag{5}$$

其中, $c_i$, $\widehat{c}_i$ 和 $w_i$ 分别表示第 $i$ 类所对应的标签、预测和权重。针对时延估计和自适应回声抵消器这一系统, 对损失函数做出了两点改进。第一是分类标签 $c_i$ 的设置。当时延的真实类别为 $I$ 时, 若与式 (4) 的滤波操作一致, 应令 $c_I=1$, 其他位置为 0。为了尽量降低时延过估计的概率, 使用有偏映射来设置分类标签 $c_i$, 使得标签时延相对真实时延小 20 ms。具体地, 当 $I\geqslant L-2$ 时, 令 $c_L=1$, 其他位置为 0; 当 $0\leqslant I<L-2$ 时, 令 $c_{I+2}=1$, 其他位置为 0。第二是分类权重 $w_i$ 的设置。普通交叉熵的分类权重为全 1, 即网络分错成各个其他类的成本是一致的。当 $I\geqslant L-2$ 时, 令 $w_L=1$, 其他位置为 2; 当 $0\leqslant I<L-2$, 将分类权重设置为

$$w_i=\begin{cases}1, & I\leqslant i\leqslant I+2,\\ 2, & \text{其他}.\end{cases}\tag{6}$$

即通过提高误分类代价来提高分类的准确率, 同时也起到降低估计方差的作用。

训练阶段, 对网络输出特征的最后一个维度 (滤波器所在维度) 做软最大化 (softmax)[14], 使得每个时间帧对应各时延概率之和为 1。这保证每帧只存在一个时延的估计类别 $\widehat{I}$。推断阶段, 计算滤波器所在维度最大值位置 (argmax), 得到时延的估计类别 $\widehat{I}$。将总类数与时延的估计类别相减, 即可得到估计的时延帧数 $L-\widehat{I}$, 从而得到时延估计值 $\widehat{\Delta}=(L-\widehat{I})M/f_s$。

### 1.3　时延估计级联回声抵消

选用变步长分块频域自适应滤波器 (VS-PBFDAF)[23] 作为回声抵消模块。自适应滤波器对回声路径 $h(n)$ 变化敏感, 因而设计一个稳定的时延估计算法非常重要。将时延估计设计为分类任务会导致所估计的时延随时间波动。在训练时, 仅使用损失函数式 (5) 对训练结果进行约束。在测试时, 进一步利用平滑和门限对时延估计结果进行稳定性控制。得到网络估计的每个时间帧 $l$ 对应的时延帧数 $\widehat{I}_0(l)$, 然后对估计结果进行平滑:

$$\widehat{I}_1(l)=0.5\widehat{I}_1(l-1)+0.5\widehat{I}_0(l),\tag{7}$$

其中, $\widehat{I}_1(l)$ 表示平滑后的时延帧数。设置一个门限来控制时延估计结果只在真实时延变化时才变化。设置时延对齐采用的时延帧数 $\widehat{I}_2$ 的更新门限为 3 个时延帧, 这对应着 20 ms 的估计误差。当连续 20 帧出现 $|\widehat{I}_1(l)-\widehat{I}_2|\geqslant 3$ 的情况时, 认为可以更新对齐操作采用的时延帧数 $\widehat{I}_2=\widehat{I}_1(l)$。该稳定性控制准则可以保证时延估计的准确率, 同时为后续自适应滤波提供帮助。

## 2　实验设置

### 2.1　数据集准备

所用的训练集每条样本长度为 10 s, 训练集和验证集时长分别为 500 h 和 100 h。远端和近端语音信号来自于 LibriSpeech[24], 其中 train-clean-100 和 dev-clean 分别用于训练集和验证集。噪声信号来自于 INTERSPEECH 2021 DNS-Challenge 的噪声数据集[25], 前 80% 和后 20% 分别用于训练集和测试集。房间脉冲响应采用镜像法生成[26], 房间尺寸在 [4, 4, 3] m 到 [10, 10, 4] m 间随机选取, 混响时间 ($\mathrm{RT}_{60}$) 在 [0.2, 1.0] s 范围内随机选取。用硬削波或软削波模拟功放引起的非线性, 用 tanh 和 sigmoid 函数来模拟扬声器引起的非线性[27–29]。将经非线性变换后的远端信号与房间脉冲响应卷积得到回声信号。回声信号相对远端信号的时延在 [0, 2.0] s 范围内随机选取。信号回声比 (SER) 和信噪比 (SNR) 分别在 [−15, 15] dB 和 [10, 30] dB 范围内随机选取。为了提高时延估计稳定性, 远端单讲样本和双端对讲样本分别占样本总数的 10% 和 90%。

测试集的远端、近端语音信号来自于 Libri-

Speech 的 test-clean。用镜像法生成测试集的房间脉冲响应时, 房间尺寸与训练集相同, 混响时间、时延根据具体测试条件设置。每种情况的测试集都有 100 条样本, 用多次实验求平均得到最终结果。从两个角度来评价时延估计模块的性能, 一是单独评价时延估计的性能, 二是评价时延估计对自适应回声抵消的增益。

### 2.2 模型和算法参数

信号采样率设置为 16 kHz。在时延估计模块中, STFT 的帧长为 20 ms, 帧移为 10 ms。模型训练时的优化器为 Adam[30], 初始学习率为 0.001, 每训练 2 轮, 学习率下降 98%。为了提高算法的跟踪性能, 在加载训练数据时, 在每条样本固有时延的基础上, 每 1 s 进行一次 4~10 个时间块内随机的时延抖动。将一段语音的训练分为两个阶段, 当$0 \leqslant l \leqslant L$时为时延估计收敛阶段, 当$l > L$时为时延估计跟踪阶段。在时延估计收敛阶段, 当$l < I$时分类标签为 0, 当$l \geqslant I$时标签按式 (5) 设置。在时延估计跟踪阶段, 当时延出现变化时, 当前帧的标签直接根据新时延按式 (5) 设置。通过这样的训练过程, 时序建模层可以自然地学到如何调整最优的记忆时长来适应时延的变化。

表 1 列出了时延估计网络的具体参数, 卷积层的通道数为实部虚部通道数之和, 超参数分别为卷积核大小、步长、输出通道数, 时序层的“ × 2”表示实部虚部分开存, 超参数分别为输入神经元个数、隐含神经元个数和层数, 全连接层的超参数表输入和输出神经元个数。在计算当前帧时, 总共考虑了 16 帧过去帧。

在回声抵消模块中, VS-PBFDAF 的帧长为 32 ms, 重叠保留率为 50%, 即傅里叶变换点数为 1024。滤波器所包含时间帧数为 8, 因此所能覆盖的混响时间范围为 512 ms。其他参数设置与文献 [23] 保持一致。

### 2.3 评价指标

利用时延估计误差、收敛速度、跟踪速度和过估计率来评价时延估计算法的性能。第$l$个时间帧时延估计误差由下式计算得到:

$$\widetilde{\Delta}(l) = \Delta(l) - \widehat{\Delta}(l), \tag{8}$$

其中,$\Delta(l)$和$\widehat{\Delta}(l)$分别为第$l$个时间帧的真实和估计时延值。当$\widetilde{\Delta}(l) \geqslant 0$时, 时延估计模块没有过估计, $\widetilde{\Delta}(l)$的值越小表示估计性能越好。当$\widetilde{\Delta}(l) < 0$时, 时延估计模块产生了过估计。算法的过估计率$\eta$定义为一段时间内$\widetilde{\Delta}(l) < 0$的时间帧数与该时间段总帧数的比值:

$$\eta = \frac{B_{\widetilde{\Delta}(l)<0}}{B}, \tag{9}$$

其中, $B_{\widetilde{\Delta}(l)<0}$为所统计的时间段内$\widetilde{\Delta}(l) < 0$的时间帧数, $B$为总时间帧数。算法的收敛速度$T_1$和跟踪时间$T_2$可用下式计算得到:

$$T_i = B_i R, \tag{10}$$

其中, $B_i$为时延估计算法收敛或跟踪所经过的时间帧数, 收敛和跟踪过程在时延估计误差小于 40 ms 时停止, $R$为时延估计精度。

在远端单讲阶段, 使用回声往返损耗增强 (ERLE) 来评价回声抵消量, 其计算方法如下:

$$\mathrm{ERLE} = 10\log_{10}\frac{\mathrm{E}\left\{[y(n)-e(n)]^2\right\}}{\mathrm{E}\left\{[\hat{e}(n)-e(n)]^2\right\}}. \tag{11}$$

在双端对讲阶段, 使用感知语音质量评价 (PESQ)[31] 来评价回声抵消后的近端语音质量。

## 3 实验结果和分析

在进行性能对比时, 首先对比了 WebRTC-DE、GCC-PHAT 和所提方法的时延估计性能, 然后将 3 种方法分别与 VS-PBFDAF 级联来比较各自对回

表 1 所提时延估计网络的参数配置

| 名称 | 输入尺寸 | 超参数 | 输出尺寸 |
|---|---|---|---|
| conv_mic | 2 × 17 × 161 | (5, 3), (1, 2), 4 | 4 × 13 × 80 |
| conv_ref | 2 × 17 × 161 | (5, 3), (1, 2), 4 | 4 × 13 × 80 |
| complex cat | (4 × 13 × 80) × 2 | 实部虚部分别拼接 | 8 × 13 × 80 |
| conv_1 | 8 × 13 × 80 | (5, 3), (1, 2), 16 | 16 × 9 × 39 |
| conv_2 | 16 × 9 × 39 | (5, 3), (1, 2), 32 | 32 × 5 × 19 |
| conv_3 | 32 × 5 × 19 | (5, 3), (1, 2), 64 | 64 × 1 × 9 |
| ComplexGRU | (288) × 2 | 576, 200, 1 | (100) × 2 |
| complex to real | (100) × 2 | 根据实部虚部求模 | 100 |
| GRU | 100 | 100, 100, 5 | 100 |
| FNN | 100 | 100, 200 | 200 |

声抵消算法的助益。具体参数设置: WebRTC-DE 的帧长 8 ms, 缓冲区帧数 250 帧, 时延估计精度 8 ms; GCC-PHAT 的帧长 2 s, 帧移 250 ms, 时延估计精度 0.0625 ms; 所提方法的最大可估计时延帧数 200, 帧长 20 ms, 帧移 10 ms, 时延估计精度 10 ms。

### 3.1　时延估计性能分析

时延估计性能评估测试集中每条样本长度为 20 s, 在第 5 秒处时延产生±50 ms 的随机抖动。回声具有非线性, SNR 为 20 dB。构建 3 个子测试集, 分别对比所提时延估计方法在不同时延、$RT_{60}$ 和 SER 下的性能。第 1 个子测试集为远端单讲情况, 固定 $RT_{60}$ 为 0.5 s, 时延范围为 [0.1, 1.5] s, 每个子类间隔步长 0.1 s。第 2 个子测试集也为远端单讲情况, 固定时延为 0.5 s, $RT_{60}$ 范围为 [0.2, 1.0] s, 每个子类间隔步长 0.1 s。第 3 个子测试集为双端对讲情况, 固定时延和 $RT_{60}$ 都为 0.5 s, SER 范围为 [−15, 15] dB, 每个子类间隔步长 5 dB。

对每条样本进行分段评估, 其中 0~5 s 用于计算收敛时间 $T_1$, 5~10 s 用于计算跟踪时间 $T_2$, 10~20 s 用于计算估计误差和过估计率。需要注意的是, 在评估所提方法的时延估计性能时, 并没有采用式 (7) 进行平滑。给出各评价指标在测试集上的均值、标准差。由于过估计率为正值, 表格中仅给出均值。

表 2 展示各种算法在远端单讲和双端对讲情况下的收敛时间 $T_1$、跟踪时间 $T_2$ 和过估计率, 其中加粗的为性能最优的结果。远端单讲结果为第 1 个和第 2 个子测试集的均值, 双端对讲结果为第 3 个子测试集的均值。对比远端单讲和双端对讲结果可知, 近端语音会造成时延估计性能的降低。所提算法的收敛时间 $T_1$ 和跟踪时间 $T_2$ 比另外两种方法短。这是因为神经网络能够从数据和标签中学习新、旧时间帧的权重, 减弱过去的时间帧对当前时间帧的影响。所提算法的过估计率较低, 这得益于式 (5) 的损失函数。值得注意的是, 如图 3 所示, 各算法的时延估计收敛时间的均值随真实时延值的增大而增大, 佐证了观测窗长需要比待估计时延长。

图 4 展示了远端单讲且时延真值为 0.5 s 时 3 种算法的估计误差随混响时间 $RT_{60}$ 的变化。WebRTC-DE 的估计误差始终为负值, 且误差绝对值随着混响时间的增长而增大。GCC-PHAT 估计误差的方差随着混响时间的增大而显著增大。这表明 WebRTC-DE 和 GCC-PHAT 在混响时间较长时, 时延估计的性能会显著下降。所提算法误差均值控制在式 (5) 设置的 20 ms 左右, 并且方差是 3 种方法中最小的且在长混响时仍具有可靠的性能。

表 2　各种算法的时延估计性能

| 算法 | | WebRTC-DE | GCC-PHAT | 所提方法 |
|---|---|---|---|---|
| 收敛时间 (s) | 远端单讲 | 1.40 ± 0.42 | 1.30 ± 0.43 | **1.07 ± 0.37** |
| | 双端对讲 | 3.08 ± 0.97 | 1.40 ± 0.26 | **1.07 ± 0.14** |
| 跟踪时间 (s) | 远端单讲 | 1.95 ± 0.23 | 0.62 ± 0.10 | **0.27 ± 0.12** |
| | 双端对讲 | 2.24 ± 0.66 | 0.81 ± 0.20 | **0.44 ± 0.17** |
| 过估计率 (%) | 远端单讲 | 86.75 | 43.53 | **0.03** |
| | 双端对讲 | 90.98 | 49.23 | **0.11** |

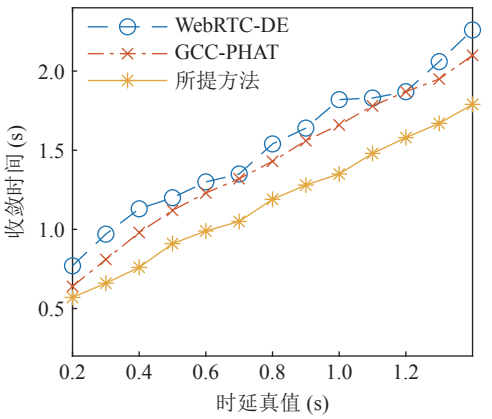


图 3　远端单讲且固定 $RT_{60}$ 为 0.5 s 时各算法时延估计收敛时间随时延真值的变化

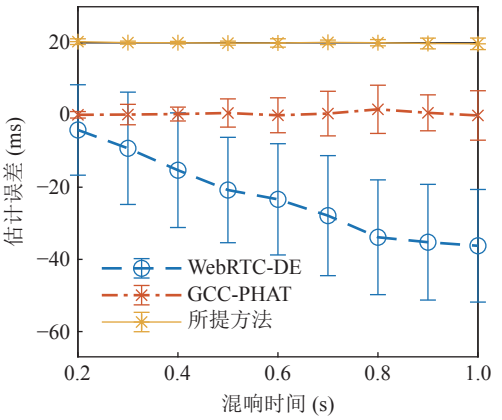


图 4　远端单讲且固定时延为 0.5 s 时各算法估计误差随 $RT_{60}$ 的变化

图 5 展示了远端单讲且混响时间为 0.5 s 时 3 种算法的估计误差随时延真值的变化。WebRTC-DE 的估计误差的均值和方差较为稳定, 但均值始终为负值。需要对 WebRTC-DE 的时延估计值进行进一步处理, 才能避免时延过估计。GCC-PHAT 在观测窗长固定时, 估计误差和方差随着时延的增大而增大。如果要使用 GCC-PHAT 估计大时延, 需要用更长时间的观测信号来求相关, 这会增加算法的复杂度。所提算法的估计误差和方差几乎不随着时延真值变化, 性能较为稳定。

图 6 展示了双端对讲阶段 3 种算法的估计误差随 SER 的变化, 时延真值和混响时间都固定为 0.5 s。3 种方法的估计误差的标准差都随着 SER 的增大而增大。在较低 SER (低于 5 dB) 时, GCC-PHAT 表现最好。在较高的 SER (高于 5 dB) 时, 所提方法的时延估计性能最好。双端对讲情况下, 近端语音干扰的统计特性与回声信号相似, 对基于信号相关特性的时延估计方法影响较大, SER 提高时 WebRTC-DE 和 GCC-PHAT 的性能相应降低。所提算法的训练集中双端对讲占 80%, 更易适应双端对讲情况。

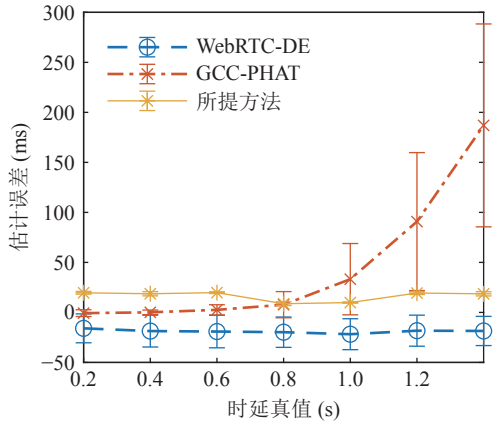


图 5 远端单讲且 $RT_{60}$ 为 0.5 s 时各算法估计误差随时延真值的变化

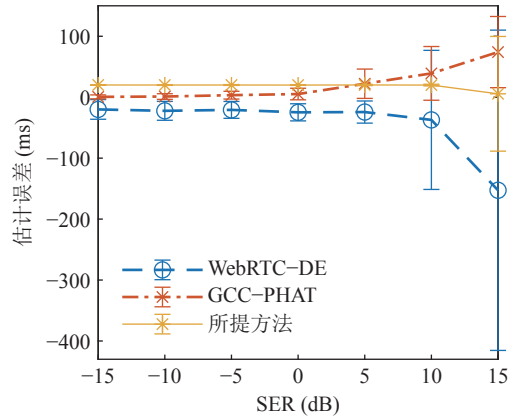


图 6 双端对讲且固定时延和 $RT_{60}$ 都为 0.5 s 时各算法估计误差随 SER 的变化

### 3.2 时延估计对回声抵消的增益

设计两种数据集, 分别评估时延估计算法在时延固定和时延变化时对回声抵消算法的影响。每条样本的 $RT_{60}$ 和时延分别在 [0.3, 0.5] s 和 [0.5, 1.5] s 范围内随机选取, 长度为 60 s, 每 20 s 为一段, 其中 0~40 s 为远端单讲, 40~60 s 为双端对讲且 SER 为 0 dB。测试集 A 控制时延不变, 回声路径在第 30 秒位置发生变化。测试集 B 控制时延改变, 时延在第 10 秒的位置相对原始值减少 50 ms, 在第 30 秒的位置相对原始值增加 50 ms。

表 3 展示了几种时延估计算法对回声抵消性能的增益, 其中加粗的是除了理想参考值之外的最好结果。表中, “未对齐”指参考信号未经时延对齐, “理想对齐”指采用真实时延值对齐参考信号。为消除自适应滤波收敛过程的影响, 未对前 10 s 回声抵消的性能进行评估。综合测试集 A 和 B 的结果来看, 所提时延估计方法在存在大时延、回声路径变化、双端对讲和时延变化的情况下, 相对 WebRTC-DE 和 GCC-PHAT 对提高回声抵消性能有更明显的帮助。比较表格中“所提方法 + VS-PBFDAF”和“所提方法 + 平滑 + VS-PBFDAF”的回声抵消结果可以知道, 对于自适应滤波来说, 稳定的时延估计值对提升回声抵消性能十分重要。

表 3 回声抵消性能展示

| | 测试集A | | | 测试集B | | |
|---|---|---|---|---|---|---|
| | ERLE (dB) | | PESQ | ERLE (dB) | | PESQ |
| | 10~20 s | 20~40 s | 40~60 s | 10~20 s | 20~40 s | 40~60 s |
| 不处理 | 0.00 | 0.00 | 1.245 | 0.00 | 0.00 | 1.285 |
| 未对齐 + VS-PBFDAF | −1.49 | −0.89 | 1.242 | −1.51 | −0.94 | 1.280 |
| 理想对齐 + VS-PBFDAF | 26.46 | 22.19 | 3.425 | 26.55 | 28.92 | 4.067 |
| WebRTC-DE + VS-PBFDAF | 7.18 | 6.60 | 1.770 | −1.24 | 4.24 | 2.210 |
| GCC-PHAT + VS-PBFDAF | 19.23 | 16.26 | 2.234 | **16.77** | 16.94 | 2.146 |
| 所提方法 + VS-PBFDAF | 20.46 | 18.21 | 2.972 | 16.17 | 16.48 | 2.508 |
| 所提方法 + 平滑 + VS-PBFDAF | **24.42** | **21.55** | **3.705** | 15.41 | **19.08** | **3.505** |

图 7 展示了当第 10 秒和第 30 秒处产生时延突变, 各时延估计算法级联 VS-PBFDAF 的时频图, 其中图 7(a)—图 7(f) 分别表示传声器接收信号、近端信号、“WebRTC-DE + VS-PBFDAF”、“GCC-PHAT + VS-PBFDAF”、“所提方法 + VS-PBFDAF”和“所提方法 + 平滑 + VS-PBFDAF”。对比图 7(d) 和图 7(e), 可以看出所提时延估计方法收敛时间 $T_1$ 和跟踪时间 $T_2$ 比 GCC-PHAT 稍快一些, 与 3.1 节结论一致。对比图 7(e) 和图 7(f) 可以看出“所提方法 + 平滑 + VS-PBFDAF”在 20~30 s 的残余回声要比“GCC-PHAT + VS-PBFDAF”和“所提方法 + VS-PBFDAF”

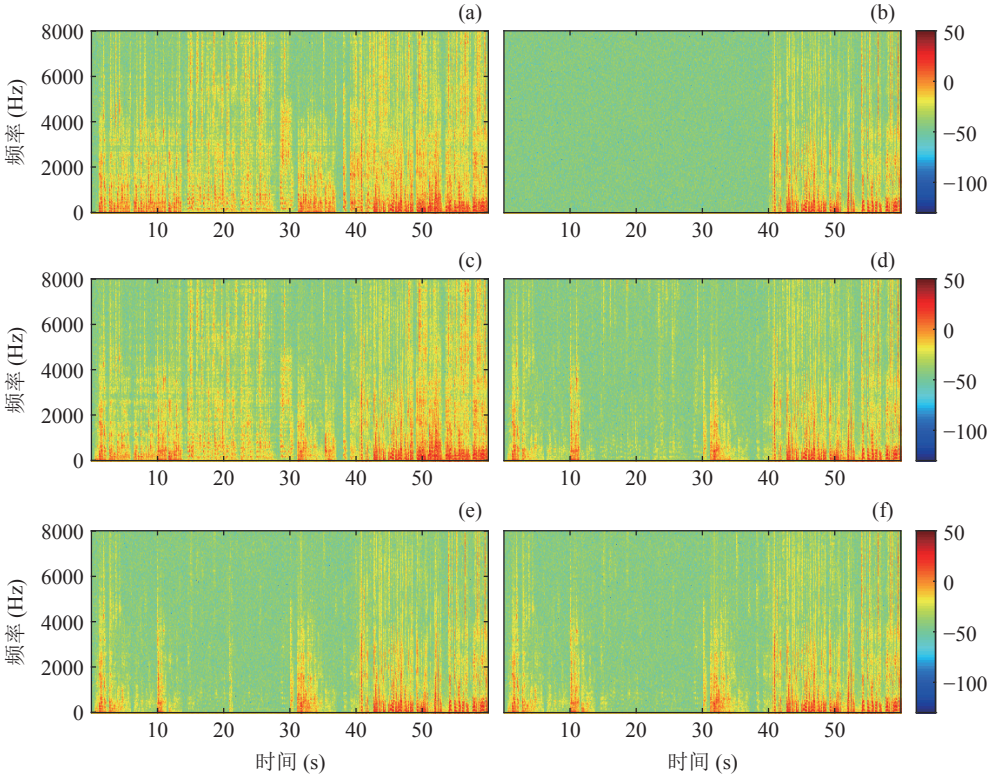


图 7　当第 10 秒和第 30 秒处产生时延突变时各时延估计算法级联 VS-PBFDAF 处理结果 (a) 传声器接收信号; (b) 近端信号; (c) WebRTC-DE + VS-PBFDAF; (d) GCC-PHAT + VS-PBFDAF; (e) 所提方法 + VS-PBFDAF; (f) 所提方法 + 平滑 + VS-PBFDAF

少。虽然平滑模块致使所提算法的收敛和跟踪时间变大, 但在估计到准确时延后, 这能够较好地保证自适应滤波的收敛和稳态效果。

## 4　结论

本文提出了一种应用于回声抵消的长混响、大时延和双端对讲鲁棒的时延估计方法。采用编码器–时序建模结构的深度学习网络, 建立远端信号和传声器信号与时延间的映射关系。时延时变的训练集, 能够降低时延估计的方差。通过改进加权的交叉熵损失的标签和权值, 预先根据真实时延值设置不均匀的误分类成本, 进一步减小时延估计的误差和方差, 并尽量避免时延过估计。与 GCC-PHAT 和 WebRTC-DE 相比, 所提算法的时延估计收敛时间、跟踪时间更短, 过估计率更低。将时延估计算法与自适应滤波级联, 当滤波器收敛到稳态后, 所提算法与自适应滤波级联的系统具有更好的性能。

## 参　考　文　献


1　周翊, 郑成诗, 李晓东. 一种用于立体声声学回波消除的新型鲁棒梯度法格梯形自适应滤波算法. 声学学报, 2010; **35**(2): 223—229

2　路阳, 程晓斌, 李晓东, 等. 结合房间声学特点的子带自适应滤波声学回音抵消算法. 电声技术, 2006(8): 54—56

3　陈智颖, 陈锴, 卢晶, 等. 双通道回声抵消系统中改进算法的定点化实现. 应用声学, 2009; **28**(3): 166—173

4　Benesty J, Morgan D R, Sondhi M M, *et al*. Advances in network and acoustic echo cancellation. New York: Springer-Verlag Berlin Heidelberg, 2001

5　Lee C M, Shin J W, Kim N S. DNN-based residual echo suppression. IEEE International Conference on Acoustic, Speech and Signal Processing, Dresden, ON, Germany, 2015: 1775—1779

6　Valin J M, Tenneti S, Helwani K, *et al*. Low-complexity, real-time joint neural echo control and speech enhancement based on PercepNet. IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 2021: 7133—7137

7　Westhausen N L, Meyer B T. Acoustic echo cancellation with the dual-signal transformation LSTM network. IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 2021: 7138—7142

8　Lu Y, Fowler R, Tian W, *et al*. Enhancing echo cancellation via estimation of delay. *IEEE Trans. Signal Process.*, 2005; **53**(11): 4159—4168

9　Cutler R, Saabas A, Parnamaa T, *et al*. INTERSPEECH 2021 acoustic echo cancellation challenge. Proc. Interspeech, Czechia, 2021: 4748—4752

10　王心一, 杜光. 降采样固定时延估算法在声回波对消系统中的应用. 山东大学学报(工学版), 2011; **41**(3): 42—45

11　陈华伟, 赵俊渭, 郭业才, 等. 一种维纳加权频域自适应时延估计算法. 声学学报, 2003; **27**(6): 514—517

12　Knapp C H K, Carter C. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, 1976; **24**(4): 320—327

13　Volcker B, Kleijn W B. Robust and low complexity delay estima-

tion. International Workshop on Acoustic Signal Enhancement, VDE, Aachen, Germany, 2012: 4—6

14 Peng R, Cheng L, Zheng C, *et al*. Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information. Proc. Interspeech, China, 2021: 4768—4772

15 Hu Y, Liu Y, Lv S, *et al*. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Proc. Interspeech, China, 2020: 2472—2476

16 武瑞沁, 陈雪勤, 俞杰, 等. 结合注意力机制的改进U-Net网络在端到端语音增强中的应用. 声学学报, 2022; **47**(2): 266—275

17 Comanducci L, Cobos M, Antonacci F, *et al*. Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 2020: 4945—4949

18 Pertilä P, Parviainen M, Myllylä V, *et al*. Time difference of arrival estimation with deep learning – from acoustic simulations to recorded data. IEEE 22nd International Workshop on Multimedia Signal Processing, Tampere, Finland, 2020: 1—6

19 Salvati D, Drioli C, Foresti G L. Time delay estimation for speaker localization using CNN-based parametrized GCC-PHAT features. Proc. Interspeech, Czechia, 2021: 1479—1483

20 Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: MIT Press, 2016

21 Chung J, Gulcehre C, Cho K, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Workshop on Deep Learning, 2014

22 Mönning N, Manandhar S. Evaluation of complex-valued neural networks on real-valued classification tasks. arXiv preprint: 1811.12351, 2018

23 Yang F, Yang J. Optimal step-size control of the partitioned block frequency-domain adaptive filter. *IEEE Trans. Circuits Syst. II*, 2018; **65**(6): 814—818

24 Panayotov V, Chen G, Povey D, *et al*. LibriSpeech: An ASR corpus based on public domain audio books. IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 2015: 5206—5210

25 Reddy C K A, Dubey H, Koishida K, *et al*. INTERSPEECH 2021 deep noise suppression challenge. Proc. Interspeech, Czechia, 2021: 2796—2800

26 Diaz-Guerra D, Miguel A, Beltran J R. gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimed. Tools Appl.*, 2021; **80**(4): 5653—5671

27 Nollett B S, Jones D L. Nonlinear echo cancellation for hands-free speakerphones. NSIP'97, 1997: 8—10

28 Comminiello D, Scarpiniti M, Azpicueta-Ruiz L A, *et al*. Functional link adaptive filters for nonlinear acoustic echo cancellation. *IEEE Trans. Audio Speech Lang. Process.*, 2013; **21**(7): 1502—1512

29 Shi K, Ma X, Zhou G T. An efficient acoustic echo cancellation design for systems with long room impulses and nonlinear loudspeakers. *Signal Process.*, 2009; **89**(2): 121—132

30 Kingma D P, Ba J L. Adam: A method for stochastic optimization. International Conference on Learning Representations, San Diego, USA, 2015

31 Rix A W, Beerends J G, Hollier M P, *et al*. Perceptual evaluation of speech quality (PESQ) − a new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 2001; **2**: 749—752