基于 ResUnet 和 TFGAN 网络的激光麦克风语音增强方法

代欣学[1,2], 范松涛[1], 周　燕[1,2*]

(1. 中国科学院半导体研究所 光电系统实验室，北京 100083；
2. 中国科学院大学，北京 100049)

摘　要：激光麦克风是一种利用光学多普勒效应获取远场语音信息的技术，其语音质量受到探测系统自身特性、光探测路径以及目标物等多个方面的影响。为了从远距离声场下的目标物获取更高质量的语音信息，文中通过单频声激励实验获得了 4 种典型目标物 (A4 纸片、A4 纸盒、瓦楞盒、塑料瓶) 的声致振动频率响应，发现了其在频率上的非均匀性。在此基础上，提出了一种基于 ResUnet 和 TFGAN 网络的激光语音增强方法，其通过 ResUnet 网络预测去噪梅尔谱图，并利用 TFGAN 网络由预测的梅尔谱图恢复出激光语音的时域波形。然后，利用实验室自制的激光麦克风在 4 种目标物上进行了远距离语音采集实验，采用文中提出的方法对采集到的激光麦克风语音进行了处理，并与非线性函数谐波重构法、DNN+谐波重构法进行了比较。最后利用客观语音质量评估 (PESQ) 和时域分段信噪比 (SNRseg) 对处理后的激光语音进行了量化评估。实验结果表明，在 4 种目标物上采集到的激光语音，经过非线性函数谐波重构方法和 DNN+谐波重构方法处理后，语音质量均无明显提升，其相应的 PESQ 和 SNRseg 分值无明显提高。而经过文中所提的 ResUnet+TFGAN 网络方法处理后，激光语音取得了更高的 PESQ 和 SNRseg 分值，语音质量明显提升。因此，文中提出的方法在激光麦克风应用中具有更好的激光语音增强效果。此外，由实验结果可知，此方法在频率响应一致性较差的目标物上，仍然可以较好地重建频谱，恢复出高质量的语音信息。

关键词：外差干涉；　语音增强；　神经网络；　声致振动

中图分类号：O439　　**文献标志码：**A　　**DOI：**10.3788/IRLA20230051

0　引　言

激光麦克风是一种利用光学多普勒效应获取声致振动信息 (语音) 的设备，与常规的麦克风相比，激光麦克风具有作用距离远、精度高、非接触的特点[1−3]。它可以定向地采集远处的声场信息，同时不受设备附近声场的干扰。但是，利用激光麦克风采集远距离声场语音信息时，获取的语音质量受到多方面因素的影响[4−5]，导致激光语音质量的严重下降。

目前，针对激光麦克风语音的语音增强算法的研究较为初步。李伟鸿[6] 等人提出了利用高斯带通滤波器和维纳滤波方法来增强激光语音信号。吕韬[7] 等人提出了基于最小控制递归平均的维纳滤波激光语音增强方法。屈直[8] 等人利用一种改进的小波阈值算法对激光语音信号进行增强。上述 3 种方法属于传统的单通道语音增强手段，要求信号和噪声满足平稳性或相关性条件，虽然简化了模型求解过程，但是限制了算法性能，在低信噪比、非平稳噪声等复杂情况下的性能显著下降。随着人工神经网络的发展，白涛[9] 等人提出了一种深度循环神经网络的激光语音增强算法。为了恢复语音中的谐波信息，Plapous[10] 提出了一种非线性函数谐波重构的方法。在此基础上，Shoji[11] 提出了一种基于深度神经网络和非线性谐波重构的激光语音增强方法。虽然上述两种深度神经网络的方法可以通过大量的语音数据的训练，学习含噪语音和清晰语音之间的复杂映射关系，获得性能优于传统方法的语音增强网络模型。但是，不同的目标物具有不同的频响特性，上述两种方法仅对来自特定的声致振动目标物的激光语音具有增强效果，对来自非预置环境下的复杂目标物的激光语音的泛化

收稿日期：2023−02−04；　**修订日期：**2023−03−27

作者简介：代欣学，男，博士生，主要从事激光外差干涉测振方面的研究。

导师(通讯作者)简介：周燕，女，研究员，博士生导师，博士，主要从事微弱光探测与成像方面的研究。

能力较差，而建立材质及形状一一对应的目标物的激光麦克风语音数据集是不现实的。因此，需要设计一种针对激光语音的语音增强方法，提高激光麦克风采集到的远场语音质量。

文中提出了一种基于 ResUnet 网络和 TFGAN 网络的激光麦克风语音增强方法，此方法参考人感知语音信息的过程[12−14]，将激光麦克风语音信息的处理分为特征分析和时域恢复两个阶段。在特征分析阶段，主要完成激光语音去噪后的梅尔谱图预测，在时域恢复阶段，则从预测的梅尔谱图恢复出时域波形。

1 ResUnet+TFGAN 方法

1.1 噪声模型

激光麦克风进行远距离语音采集时会受到探测系统自身特性、光探测路径以及目标物等方面的影响，获取的语音中主要包括以下几种噪声：

1) 以加性噪声为主的系统背景噪声：

$$d_{add}(x) = x + n \tag{1}$$

式中：x为清晰语音；n为加性噪声。

2) 声场多径传播导致的混响噪声：

$$d_{rev}(x) = x \times r \tag{2}$$

式中：r为空间脉冲响应滤波器。

3) 电子电路决定的截断噪声d_{trun}：

$$d_{trun}(x) = \max\left[\min(x,\eta),-\eta\right] \tag{3}$$

式中：η为电子电路决定的信号截断阈值。

4) 相位解调算法导致的卷绕噪声d_{wrap}经过微分后表现为毛刺信号：

$$d_{wrap}(x) = x + K \times \delta(t - t_0) \tag{4}$$

式中：$\delta(t)$为冲激信号；K为某个极大值。

5) 目标物特性决定的频域失真d_{res}：

$$d_{res}(x) = Resamp(x, ori_{fs}, low_{fs}) \tag{5}$$

式中：$Resamp(\cdot)$表示重采样操作；ori_{fs}为原始采样率；low_{fs}为重采样后的采样率。

文中根据上述各种噪声类型，建立了两种噪声模型$d(x)$和$h(x)$，用于生成两种网络的训练集，如公式(6)和(7)所示：

$$d(x) = rand(d_{add}, d_{rev}, d_{trun}, d_{wrap}) \tag{6}$$

$$h(x) = d_{res} \tag{7}$$

式中：$rand(\cdot)$为随机函数。

1.2 网络构建

如图 1 所示，文中利用训练好的网络对激光麦克风采集到的激光语音信号进行预测，依次完成特征分析和时域恢复两个阶段的处理，最终得到增强后的激光语音信号。

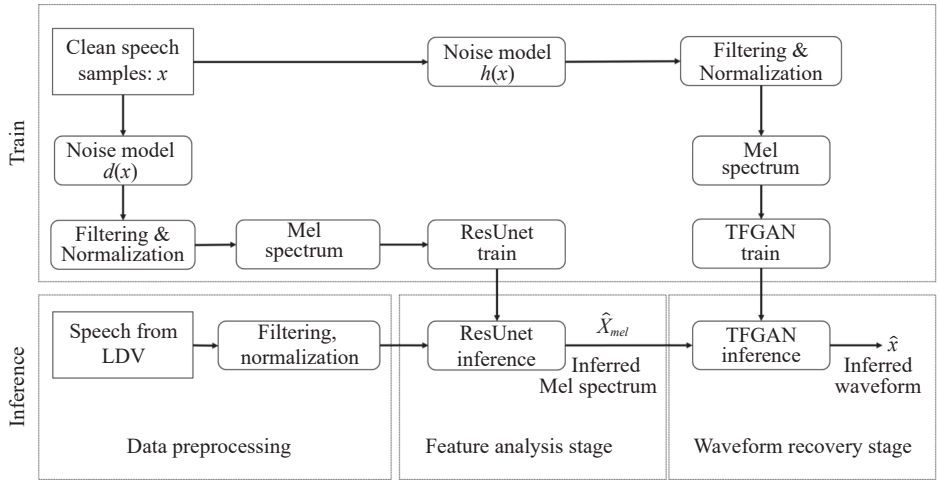

图 1 网络训练及预测的示意图

Fig.1 Schematic diagram of network training and prediction

1.2.1 ResUnet 网络

在特征分析阶段中，文中采用了一种改进的 Unet 网络结构[15−16]。

如图 2 所示，该网络包含 6 层编码及解码模块，编码与解码模块中均包含一个由 Batch normalization、LeakyReLu 以及线性卷积计算构成的残差卷

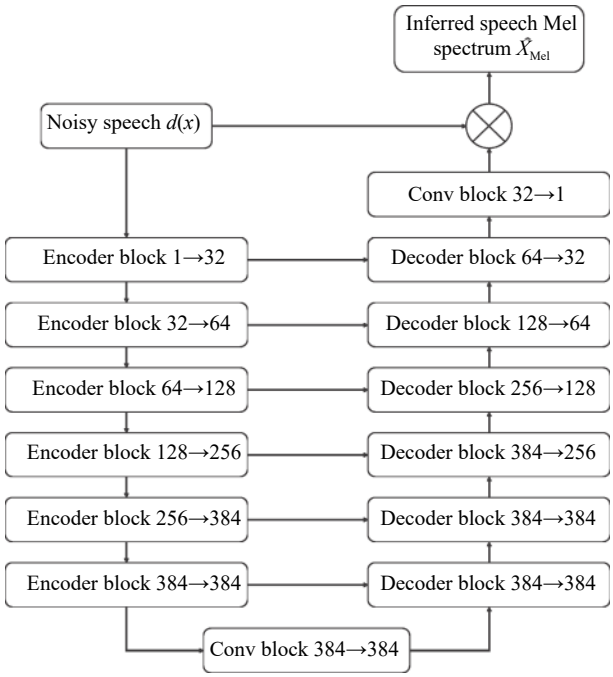

图 2 ResUnet 网络结构

Fig.2 ResUnet network structure

积层。编码层中采用平均池化层进行下采样，解码层中采用转置卷积完成上采样。同一层的编码与解码之间存在一个跳跃连接。

在该阶段，采用预测的语音梅尔谱图$\hat{X}_{Mel}$与原始清晰的语音梅尔谱图X_{Mel}的平均绝对误差 (Mean Absolute Error, MAE) 作为损失函数$\mathcal{L}_1$来优化网络模型：

$$\mathcal{L}_1 = \left\| \hat{X}_{Mel} - X_{Mel} \right\|_1 \tag{8}$$

1.2.2 TFGAN 网络

在时域恢复阶段，文中采用了 TFGAN 网络结构[17−18]，它是一种改进的生成对抗网络 (GAN) 结构。

生成器的网络结构如图 3 所示，包含条件网络模块、上采样以及残差堆模块、一维卷积层。在条件网络模块中，包含 3 个具有线性激活单元 (ELU) 的一维卷积层。在上采样模块中，输入数组首先通过带参数的线性整流单元 (LeakyReLu) 激活，然后利用正弦函

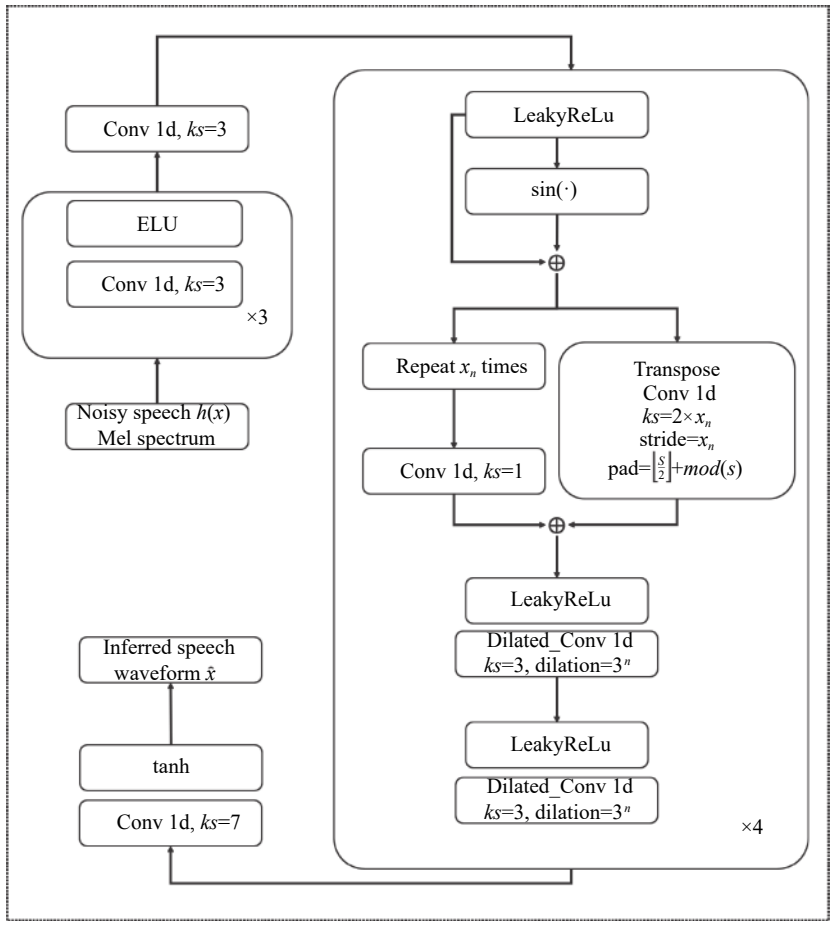

图 3 生成器的网络结构

Fig.3 Network structure of generator

数进行映射，该函数的输出与输入相加，进入两个分支。一个分支重复样本数次，再进行一维卷积。另一个分支按一定的步长进行转置卷积，最后将两个分支输出的和值经过由 LeakyReLu 层和扩展卷积 (Dilated_Conv 1d) 层组成残差堆。

鉴别器的组成及网络结构如图 4 和图 5 所示，其中时域鉴别器网络中只包含 LeakyReLu 层和卷积 (Conv 1d) 层，而频域鉴别器网络由一个卷积 (Conv 2d) 层和 8 个残差卷积层组成。

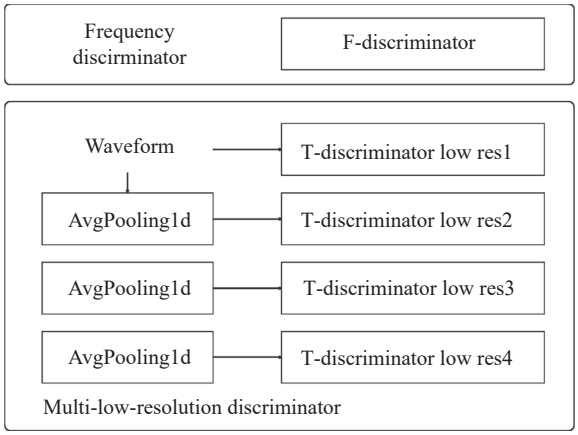

图 4 时域及频域鉴别器

Fig.4 Discriminators in the time and frequency domains

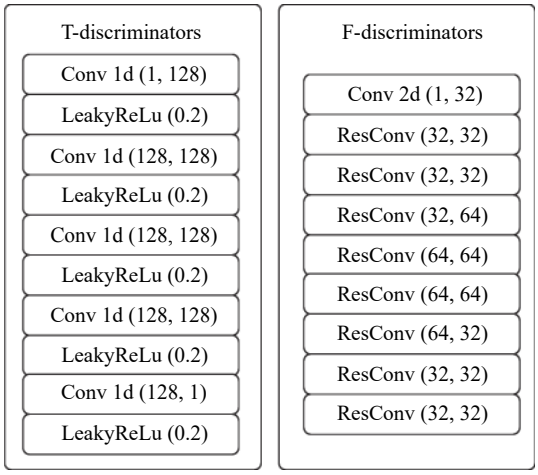

图 5 时域及频域鉴别器的网络结构

Fig.5 Network structure of time domain and frequency domain discriminator

在进行 TFGAN 网络训练时，采用的损失函数 $\mathcal{L}_2$ 为频域损失 $\mathcal{L}_F$、时域损失 $\mathcal{L}_T$ 以及加权判决器损失 $\mathcal{L}_D$ 的组合，如公式 (9)~(13) 所示：

$$\mathcal{L}_2 = \mathcal{L}_T + \mathcal{L}_F + 4\times\mathcal{L}_D \tag{9}$$

$$\mathcal{L}_T\left(\hat{X},X\right) = v\left(\hat{X}_\omega\right) - v\left(X_\omega\right)_1 \tag{10}$$

$$\mathcal{L}_F(\hat{X},X) = [\|\hat{X}_{\text{Mel}} - X_{\text{Mel}}\|_2 + \|\log(|\hat{X}|) - \log(|X|)\|_1] \tag{11}$$

$$\mathcal{L}_D(\hat{X},X) = \min_g \max_D \{\mathbb{E}_X[\log D(X)]\} + \mathbb{E}_{\hat{X}}\{\log[1 - D(X)]\} \tag{12}$$

$$D(X) = D_T(X) + D_F(X) \tag{13}$$

式中：$v(\cdot)$ 为音频分帧操作时的窗函数；ω 为语音信号分帧数；$\mathbb{E}$ 为期望概率；$D_T(X)$ 为时域鉴别器网络的输出值；$D_F(X)$ 为频域鉴别器的输出值。

2 实 验

2.1 实验装置

文中搭建的实验装置如图 6 所示，目标物与激光麦克风直线距离为 4 m，声源与目标物的直线距离为 0.1 m，声源不仅可以直接播放 SD 卡内的音频文件，还可以在信号发生器的调制下，播放特定频率及幅度的音频。

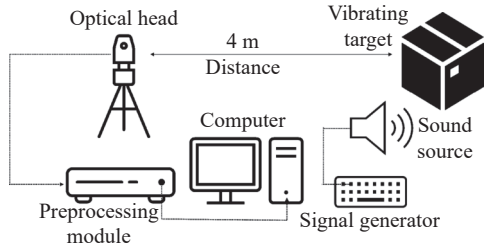

图 6 实验装置示意图

Fig.6 Schematic diagram of experimental apparatus

实验中使用的激光麦克风由实验室基于激光多普勒测振原理自制，包括光学头、信号预处理模块以及上位机 3 个部分。

其中，光学头主要负责实现本振光与目标物漫反射光的外差干涉。信号预处理模块由 NI-USRP2944 构成，负责外差干涉信号的采样滤波等预处理。上位机则采用 DELL 台式机，其处理器为 11th Gen Intel® Core(TM) i7-11700K@3.60 GHz，机带 RAM 32 GB，64 位操作系统。上位机显卡采用 NVIDIA GeForce RTX3070。在上位机上利用 Labview2019(×64) 软件实现信号的解调、显示及存储。

目标物采用 A4 纸片、A4 纸盒、瓦楞盒、塑料瓶等常见的典型物品，如图 7 所示，4 种目标物中，A4 纸片与 A4 纸盒材质均为长纤维木浆，瓦楞盒材质主要为草浆，A4 纸盒与瓦楞盒大小一致，均为 200 cm×140 cm×40 cm，空水瓶材质为 PET 塑料，容量为 380 mL。

图 7 实验中采用的 4 种目标物 (A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶)

Fig.7 The four objects used in the experiment (A4 sheet, A4 box, corrugated box, PET plastic bottle)

2.2 目标物的频响特性

目标物的频响特性受材质及形状的影响。如图 8 所示，黄色、紫色、绿色、橙色分别给出了 A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶在单频声场下的频响特性。

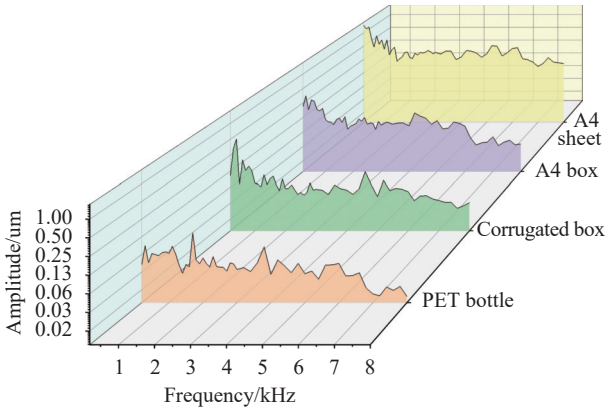

图 8 目标物的频率响应 (黄色: A4 纸片; 紫色: A4 纸盒; 绿色: 瓦楞盒; 橙色: PET 塑料瓶)

Fig.8 The frequency response of the target (yellow: A4 sheet; purple: A4 box; green: Corrugated box; orange: PET plastic bottle)

由图 8 可知，同一材质，不同形状的物体具有明显不同的频响曲线。在 3~8 kHz 内，A4 纸片的频响曲线更加平坦。在 5~8 kHz 内，A4 纸片 (黄色) 的振动响应相对于 A4 纸盒 (紫色) 更加强烈；同一形状，不同材质的物体频响特性也存在较大差异。例如，A4 纸盒与瓦楞盒形状大小几乎一致，但是，在 0~3 kHz 内，瓦楞盒的频响曲线更加陡峭。总的来看，4 种目标物的频响一致性较差，主要表现为高频响应明显弱于低频响应。

2.3 数据集

文中清晰语音样本来自于清华大学语音与语言技术中心发布的开源中文语音数据集 THCHS30，总时长大约 40 h，采样率为 16 kHz，采样大小 16 bit。文中将其中 10000 个语音样本用于训练集，2495 个语音样本用于测试集。ResUnet 网络的任务是预测去噪后的激光语音梅尔谱图，因此文中从加性噪声、混响、截断、卷绕等 4 种失真中随机地抽取 1 种与清晰语音数据集混合，形成 ResUnet 网络的训练集以及测试集。其中，加性噪声共 30 种，来自于 ESC50 环境的声数据集 (降采样至 16 kHz)，包括 10 种室内声音类型和 3 种信噪比 (−5、0、5 dB)。而为了模拟实验室内的混响效果，文中从开源脚本 rir_simulator_python 生成的 64 个空间脉冲滤波器中随机抽取，脚本参数：室内高度、宽度、长度 (m) 设置为 (4 m, 6 m, 10 m)，声场衰减 RT60 值设置为 0.4。为了实现截断失真效果，文中设定的截断阈值从 0.2、0.4、0.6 、0.8 中随机抽取。卷绕失真引入的毛刺信号的插入位置在输入的清晰语音数据采样点上随机选择，K 值在区间 (0.5, 2) 上服从均匀分布。TFGAN 网络的任务是由上一阶段输出的梅尔谱图还原出语音时域波形，其面临的主要问题是目标物频响特性导致的激光语音频域失真问题。因此，文中首先将 THCHS30(采样率为 16 kHz) 的清晰语音样本经过截止频率为 f_u/2且阶数为m的巴特沃斯低通滤波器，然后再分别重采样到 f_u和 16 kHz，前者重采样结果经过梅尔滤波器组后形成梅尔滤谱，并与后者形成 (频域失真，目标波形) 数据对。其中，f_u在区间(4 000, 8 000)上服从于均匀分布，m在区间(2, 10)上服从于均匀分布。

2.4 网络训练设置

文中采用 Pytorch 作为深度学习框架，使用 Adam 优化器来优化网络的训练，初始学习为 5×10^{-4}，β_1为 0.9，β_2为 0.999，训练批量大小为 16。其中，ResUnet 网络的迭代训练过程中选择 MAE 作为损失函数，如公式 (8) 所示。TFGAN 网络的迭代训练则采用时域损失、频域损失以及加权判决器损失的线性组合作为损失函数。此外，DNN+谐波重构方法中，DNN 网络的层数设置为 10 层，优化器、训练批量以及损失函数与 ResUnet 网络相同。由公式 (1)~(5) 中随机抽取 1 种失真效果与清晰语音数据集混合，形成 DNN 网络的训练集与测试集。

3 实验结果

3.1 语谱图分析

为了便于分析，文中利用短时傅里叶变换方法，将语音文件绘制成了语谱图，如图 9 所示，语谱图的纵坐标表示频率，横坐标表示加窗后的时间 (帧序)。人语音的语谱图上一般存在着明显的纹路 (声纹)，其中短横纹对应共振峰，竖直条则代表基音，语谱图中的颜色由暗紫色到亮黄色，对应着频率分量强度由低到高。

如图 9(a1)~(a4) 所示，ori 代表 SD 卡内存储的清晰语音，从其语谱图上可见明显的声纹。图 9(b1)~

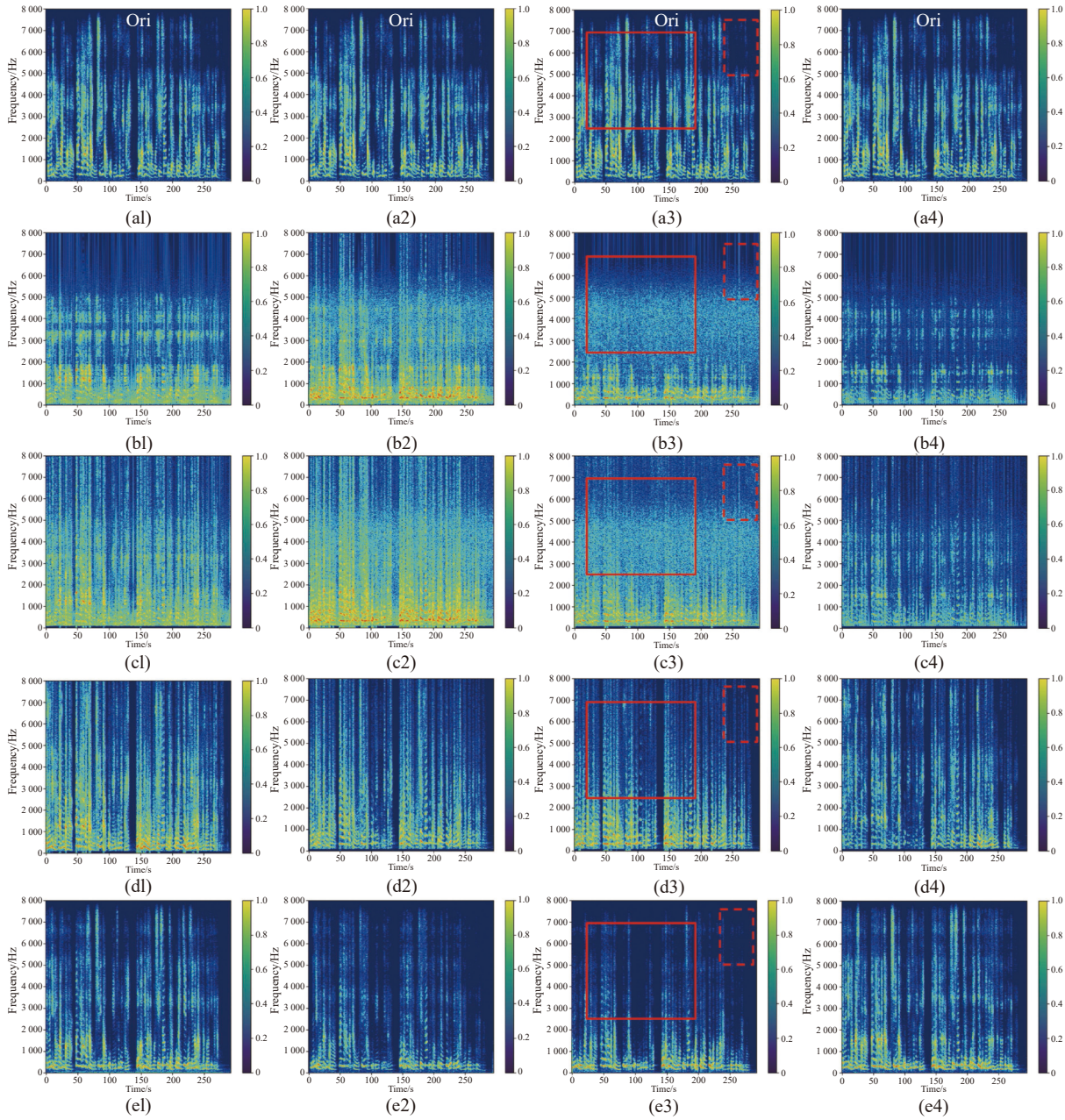

图 9 语谱图。(a) SD 卡内存储的清晰语音；(b) 激光麦克风采集到的激光语音；(c) 非线性函数谐波重构方法；(d) DNN+非线性谐波重构方法；(e) 文中所提方法；1: A4 纸片；2: A4 纸盒；3: 瓦楞盒；4: PET 塑料瓶

Fig.9 Spectrogram. (a) Clear original voice saved in SD card; (b) Laser speech collected by laser microphone; (c) Nonlinear function harmonic reconstruction method; (d) DNN+nonlinear harmonic reconstruction method; (e) Method of this article; 1-4: A4 sheets, A4 box, corrugated box, PET plastic bottles

(b4) 表示的是利用激光麦克风对声场内 4 种目标物分别采集所获得的激光语音。图 9(c1)~(c4) 表示的是对激光麦克风采集到的 4 种语音经过文献 [10] 中提出的非线性谐波重构处理后的结果。图 9(d1)~(d4) 表示的是按照文献 [11] 中提出的 DNN+谐波重构方法对激光麦克风采集到的 4 种语音进行处理后得到的语音。图 9(e1)~(e4) 表示的是利用文中所提的基于 ResUnet+TFGAN 的方法对激光麦克风采集到的 4 种语音进行处理后得到的语音。

观察语谱图可以发现，激光麦克风采集到的激光语音存在较强的背景噪声，低频段声纹较为模糊，高频段声纹则淹没在背景噪声中。其中，以 A4 纸片和塑料瓶为目标物时，高频段的声纹较为清晰。而以瓦楞盒为目标物时，几乎无法看到高频段声纹。

如果仅通过非线性函数进行谐波重构，如图 9(c3) 所示，虽然高频段被增强，但是此方法对语音信息和背景噪声产生了相同的效果，不能恢复出明显的高频段声纹，未实现激光语音音质的改善。如图 9(d3) 所示，利用 DNN+谐波重构方法对采集到的激光语音进行处理后，可以降低非语音段背景噪声，但是，如虚线红框中所示，此方法无法消除语音段内的背景噪声。

而如图 9(e3) 所示，利用文中所提方法对采集到的激光语音进行处理时，首先通过 ResUnet 网络预测了清晰语音的梅尔谱，降低了非语音段的背景噪声，然后还通过 TFGAN 网络，从预测的梅尔谱中直接解码还原了时域波形，最终获得的语音与 SD 卡内存储的清晰语音的语谱图在高频区纹理更加接近。

3.2 激光语音质量的客观评价方法

通过比较纯净语音和算法处理后的语音之间的“距离”可以量化语音的质量，从而判断算法的性能，并比较出算法之间的优劣。文中选取客观语音质量评估 (Perceptual Evaluation Of Speech Quality, PESQ) 和时域分段信噪比 (time-domian segmental SNR, SNRseg) 作为实验中激光语音质量的评价指标。

PESQ 算法的计算框图如图 10 所示，输入为一个参考信号 (纯净语音) 和一个待评价信号 (模糊语音)，输出的分值范围在−0.5~4.5 之间，得分越高表示语音质量越好[19]。

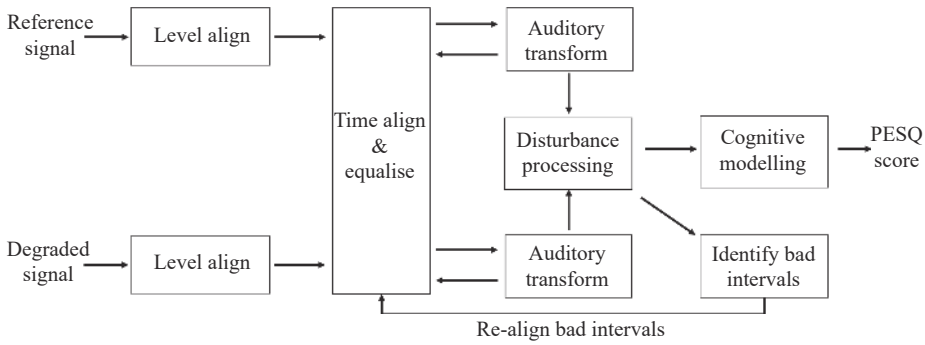

图 10 PESQ 评价方法的计算框图

Fig.10 Calculation diagram of PESQ

时域分段信噪比评价方法的计算如公式 (14) 所示：

$$\mathrm{SNRseg}=\frac{10}{M}\sum_{m=0}^{M-1}\lg\frac{\sum_{n=N_m}^{N_m+N-1}x^2(n)}{\sum_{n=N_m}^{N_m+N-1}(x(n)-\hat{x}(n))^2} \tag{14}$$

式中：$x(n)$为参考信号 (纯净语音)；$\hat{x}(n)$为待评价信号 (模糊语音)；N 为帧长 (文中为 30 ms)；M 为信号的帧数，计算结果值越大，语音质量越好。

3.3 客观评价结果

如图 11 所示，横坐标从左至右依次为激光麦克风采集的激光语音 (degraded)、经过非线性函数谐波重构法 (harm_rec) 处理后的激光语音、经过 DNN+谐波重构方法处理后的激光麦克风语音 (DNN + harm_rec)、经过文中所提的基于 ResUnet + TFGAN 网络的方法处理后的激光语音 (ours)。图 11(a)、(b) 中纵坐标分别为 PESQ 的评价得分和 SNRseg 的评价得分。

以 SD 卡内存储的原始清晰语音为参考信号，PESQ 评价方法的计算结果如图 11(a) 所示，激光麦克风设备在 A4 纸片、A4 纸盒、瓦楞盒上、PET 塑料瓶上直接采集到的激光语音 (degraded) 的 PESQ 分值分

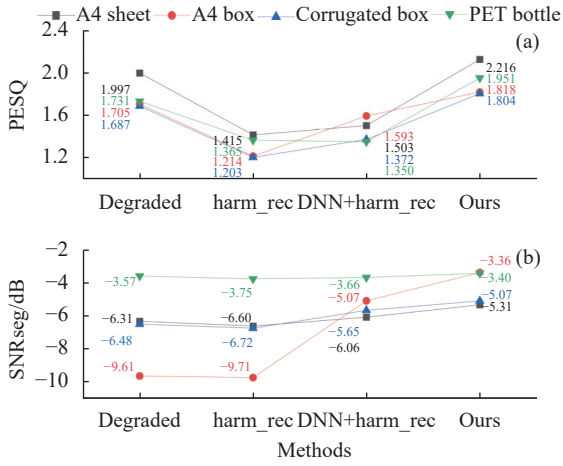

图 11 语音质量客观评价结果。(a) PESQ 方法; (b) SNRseg 方法。Degraded: 激光麦克风采集的激光语音; harm_rec: 非线性函数谐波重构方法; DNN+harm_rec: DNN+谐波重构方法; ours: 文中所提方法

Fig.11 Objective evaluation trsults of speech quality. (a) PESQ method; (b) SNRseg method. Degraded: laser speech collected by laser microphone; harm_rec: nonlinear function harmonic reconstruction method; DNN+harm_rec: DNN+harmonic reconstruction method; ours: the method proposed in this paper

别为 1.997、1.705、1.687、1.731; 经过非线性函数谐波重构方法处理后, 噪声同步发生重构, 产生了额外的谐波噪声, A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶所对应的经过此方法处理后的激光语音 PESQ 分值分别为 1.415、1.214、1.203、1.365; 而经过 DNN+谐波重构方法处理后, 音频中的白噪声被抑制。因此, 与单纯的谐波重构法相比, PESQ 分值略微上升, A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶所对应的经过此方法处理后的激光语音 PESQ 分值分别为 1.503、1.593、1.372、1.350。

与上述两种方法相比, 利用文中所提的方法对采集到的激光语音进行分阶段的增强处理后, 音频中的宽带噪声及脉冲噪声等得到了更好的抑制, 并且重建了较为准确的高频信息。A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶所对应的经过此方法处理后的激光语音 PESQ 分值分别为 2.126、1.818、1.804、1.951, PESQ 得分明显提高。

以 SD 卡内存储的原始清晰语音为参考信号, 时域分段信噪比评价方法的计算结果如图 11(b) 所示, 激光麦克风设备在 A4 纸片、A4 纸盒、瓦楞盒上、PET 塑料瓶上直接采集到的激光语音 (degraded) 的 SNRseg 分值分别为−6.31、−9.61、−6.48 、−3.57 dB。经过非线性函数谐波重构方法处理后, SNRseg 分值均相对下降, 分别下降 0.29、0.1、0.24、0.18 dB。经过 DNN+谐波重构方法处理后, SNRseg 分值变化不一致, 其中, 在 A4 纸片、A4 纸盒、瓦楞盒上分值相对上升 0.25、4.54、0.83 dB, 在 PET 塑料瓶上分值相对下降 0.09 dB。而通过文中所提的方法处理后, SNRseg 分值均明显上升, 在 A4 纸片、A4 纸盒、瓦楞盒、PET 塑料瓶上分别提高了 1、6.25、1.41、0.17 dB。

实验结果表明:

1) 仅通过非线性函数谐波重构处理, 无法改善激光麦克风采集到的激光语音质量, 由图 9 可以看出, 背景噪声同步发生谐波重构所产生的高频谐波噪声是语音质量进一步下降的原因。

2) 利用 DNN+谐波重构方法进行处理后, 虽然非语音段的背景噪声被消除, 但是无法抑制语音段内的高频噪声以及脉冲噪声, 与 SD 卡内存储的清晰语音相比, 激光语音质量没有明显的改善。

3) 经过文中所提的 ResUnet+TFGAN 网络处理后, 各个目标物所对应的激光语音质量得到了明显提升。并且, 当目标物频响一致性较差时, 可以在不引入新的高频噪声的情况下, 恢复出了较准确的频域信息。

4 结 论

文中提出了一种基于 ResUnet 和 TFGAN 网络的激光麦克风语音增强方法, 并通过实验室自制激光麦克风在多种目标物上采集了语音样本, 对文中所提方法进行了实验验证。实验结果表明, 此方法对来自于多种目标物的激光麦克风语音均具有较好的增强效果。与非线性函数谐波重构法、DNN+谐波重构法相比, 此方法的优势是通过 ResUet 和 TFGAN 网络分别实现激光语音的清晰梅尔谱预测和时域波形恢复, 避免了谐波重构方法在重建语音信号时所引入的高频噪声, 同时恢复了更加清晰的激光语音高频信息。客观评价方法 PESQ 和 SNRseg 的计算结果表明, 经过文中所提的方法处理后的激光麦克风语音具有更高的语音质量。此方法在一定程度上扩展了激光麦克风的适用范围, 今后将在材质及形状更复杂的目标物

上，进一步验证并改进此方法。

参考文献：

[1] Fan Hongxing, Zhou Yan, Fan Songtao, et al. Research on nanometer displacement telemetry based on digital zero intermediate frequency [J]. ***Infrared and Laser Engineering***, 2018, 47(11): 1117008. (in Chinese)

[2] Yan Chunhui, Wang Tingfeng, Zhang Heyong, et al. Arctangent compensation algorithm of laser speech detection system [J]. ***Infrared and Laser Engineering***, 2017, 46(9): 0906004. (in Chinese)

[3] Li Liyan, Fan Songtao, Zhou Yan. Eliminating light intensity disturbance algorithm based on phase demodulation carrier [J]. ***Infrared and Laser Engineering***, 2021, 50(9): 20210485. (in Chinese)

[4] Wang X. Research on several denoising methods for laser sound detection [D]. Hefei: Hefei University of Technology, 2018. (in Chinese)

[5] Luo Xinwei, Zhang Xi, Lin Benhai, et al. Experimental study on the response of soil nails with different materials and shapes to vibration wave [J]. ***Journal of Safety and Environment***, 2021, 21(4): 1712-1719. (in Chinese)

[6] Li Weihong, Liu Ming, Zhu Zhigang, et al. LDV remote voice acquisition and enhancement[C]//18th International Conference on Pattern Recognition (ICPR'06), 2006: 262-265.

[7] Lv Tao, Zhang Heyong, Guo Jin, et al. Acquisition and enhancement of remote voice based on laser coherent method [J]. ***Optics and Precision Engineering***, 2017, 25(3): 569-575. (in Chinese)

[8] Qu Zhi, Zhang Bohu. An improved wavelet threshold algorithm applied in laser interception [J]. ***Laser Technology***, 2014, 38(2): 218-224. (in Chinese)

[9] Bai Tao, Wu Jin, Li Minglei, et al. Application of DRNN in voice measurement system of laser Doppler vibrometer [J]. ***Laser Technology***, 2019, 43(1): 109-114. (in Chinese)

[10] Plapous C, Marro C, Scalart P. Speech enhancement using harmonic regeneration[C]//ICASSP'05. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. IEEE, 2005.

[11] Shoji U, Iwai, K, Fukumori T, et al. Sound quality improvement for speech acquisition based on deep learning and harmonic reconstruction with laser microphone[C]//Proceedings of the ICA Congress, 2019: 6937-6944.

[12] Bregman A S. Auditory scene analysis: The Perceptual Organization of Sound[M]. Cambridge: MIT Press, 1994.

[13] Griffiths T D, Warren J D. The planum temporale as a computational hub [J]. ***Trends in Neurosciences***, 2002, 25(7): 348-353.

[14] Dan K H. Neural and cognitive mechanisms affecting perceptual adaptation to distorted speech[D]. London: University College London, 2019.

[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer, 2015: 234-241.

[16] Choi H S, Park S, Lee J H, et al, Real-time denoising and dereverberation wtih tiny recurrent U-Net[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 5789-5793.

[17] Tian Q, Chen Y, Zhang Z, et al. TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis[EB/OL]. (2020-11-24)[2023-02-06]. https://doi.org/10.48550/arXiv.2011.12206.

[18] Liu H, Liu X, Kong Q, et al. VoiceFixer: A unified framework for high-fidelity speech restoration[EB/OL]. (2022-04-12)[2023-02-03]. https://arxiv.org/abs/2204.05841.

[19] 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[C]//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01 CH37221), 2001, 2: 749-752.

Speech enhancement method of laser microphone based on ResUnet and TFGAN network

Dai Xinxue[1,2], Fan Songtao[1], Zhou Yan[1,2*]

(1. Optoelectronics System Laboratory, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract:

Objective Laser microphone is a kind of equipment which employs optical Doppler effect to acquire acoustic vibration information (speech). Compared with conventional microphones, laser microphones have the characteristics of extended range, high precision and non-contact. It is capable of collecting distant sound field information in a directional fashion while avoiding interference from the sound field close to the equipment. However, when the laser microphone is used to collect the remote sound field speech information, the quality of the obtained speech is affected by many factors, which leads to the severe decline of the laser speech quality. At present, the research of speech enhancement algorithm for laser microphone speech is relatively preliminary. The traditional single-channel speech enhancement method requires the signal and noise to satisfy the conditions of stationarity or correlation, and its performance is significantly reduced under complex conditions such as low signal-to-noise ratio and non-stationarity noise. The method based on deep neural network can understand the complex mapping relationship between noisy speech and clear speech, and the performance is better than the traditional method. This technique, however, has poor generalizability for laser speech from complex targets in unpreset environments because different targets have different frequency response characteristics. Therefore, in order to increase the quality of far-field speech captured by laser microphones, a laser microphone speech enhancement method based on ResUnet network and TFGAN network is proposed in this paper.

Methods Using laboratory-made laser microphones, four different types of objects were used in this paper's remote speech acquisition tests (Fig.6). The technique described in this paper is used to process the recorded speech, and it is contrasted with methods for nonlinear function harmonic reconstruction and DNN+ harmonic reconstruction (Fig.9). Finally, objective speech quality assessment (PESQ) and time-domain segmented signal-to-noise ratio (SNRseg) were used to quantitatively evaluate the processed laser speech (Fig.11).

Results and Discussions Compared with the above two methods, the method proposed in this paper can better suppress the broadband noise and pulse noise and reconstruct the more accurate high-frequency information after the stepwise enhancement processing of the collected laser speech. The laser speech PESQ scores of A4 paper, A4 paper box, corrugated box and PET plastic bottle after this method are 2.126, 1.818, 1.804 and 1.951, respectively increased by 0.129, 0.113, 0.117 and 0.22. The corresponding SNRseg scores were −5.31 dB, −3.36 dB, −5.07 dB and −3.40 dB, which were increased by 1 dB, 6.25 dB, 1.41 dB and 0.17 dB, respectively. The experimental results show that the ResUnet+TFGAN network method proposed in this paper can effectively improve the laser speech quality of the above targets.

Conclusions In this study, a laser microphone speech enhancement method based on ResUnet and TFGAN network is proposed. Speech pieces are gathered on various targets by self-made laser microphones in the lab, and the proposed method is demonstrated through experiments. The experimental results show that this method can enhance the speech of laser microphone from a variety of objects. Compared with the nonlinear function harmonic reconstruction method and DNN+ harmonic reconstruction method, the advantages of this method are that ResUet and TFGAN networks can respectively realize the clear Mel spectrum prediction and time domain waveform recovery of laser speech, avoiding the high-frequency noise introduced by the harmonic reconstruction method in the reconstruction of speech signal, and at the same time recover the more clear high-frequency information of laser speech. PESQ and SNRseg results demonstrate that using the proposed method results in improved speech quality for the laser microphone. This method extends the application range of laser microphones to a certain extent, and we will further verify and improve this method on objects with more complex materials and shapes.

Key words: heterodyne interference; speech enhancement; neural network; acoustic vibration