文章编号：1003－1251(2024)02－0089－08

GAN在电动汽车主动发声系统中的应用研究

梁 凯[1]，张 巍[1]，赵海军[2]

(1. 洛阳理工学院 信息化技术中心，河南 洛阳 471023；2. 天津职业技术师范大学 智能车路协同与安全技术国家地方联合工程研究中心，天津 300222)

摘 要：为提高电动汽车引擎拟音的个性化效果和质量，引入生成对抗网络(GAN)模型，构建了电动汽车的GAN主动发声模型，设计了模型中各层网络的结构和卷积核大小，利用自适应时刻估计算法优化网络各层权重，并将模型用于样本生成试验。在模型训练中提出一种相位扰动操作，用于解决上采样操作产生音调噪声的问题；为证明GAN模型中不同输入信号的性能差异，构建了基于二维声谱图输入的GAN模型，并用于对照试验。试验结果表明：模型可准确地学习到原始音频信号的特征分布；人耳听觉测试结果显示，生成的声音样本真实度在90%以上；基于留一法(LOO)的1-NN分类评价结果显示，原生音频和二维声谱图GAN模型的LOO精度均大于或接近50%，表明模型训练未产生过度拟合，采用本文方法生成音效真实可靠。

关 键 词：电动汽车；主动发声；生成对抗网络；原生音频；声谱图

中图分类号：U469.72⁺2 **文献标志码**：A **DOI**：10.3969/j.issn.1003－1251.2024.02.014

Research on the Application of Generative Adversarial Network in the Sound Synthesis System of Electric Vehicles

LIANG Kai[1], ZHANG Wei[1], ZHAO Haijun[2]

(1. Information Technology Center, Luoyang Institute of Science and Technology, Luoyang 471023, China;
2. National Joint Engineering Research Center of Intelligent Vehicle Infrastructure Cooperation and Safety Technology, Tianjin University of Technology and Education, Tianjin 300222, China)

Abstract: To improve the personalization and quality of the sound imitation of electric vehicle engines, a generative adversarial networks (GAN) model was introduced to construct the GAN active sound model of electric vehicles. The structure of each layer of the network and the size of the convolution kernel in the model were designed. The adaptive moment estimation algorithm was used to optimize the weights of each layer in the network. The model was used for sample generation experiments. A phase perturbation operation was proposed in model training to solve the problem of pitch noise generated by the upsampling operation. In order to prove the performance of different input signals in the GAN model, a GAN model based on two-dimensional spectrogram input was constructed and used for controlled trials. The test results show that the model can accurately learn the feature distribution of the original audio signal. The human hearing test results show that the authenticity of the generated sound samples is more than 90%. The 1-NN classification evaluation results

收稿日期：2023－07－19

基金项目：国家自然科学基金项目(U1604141)；中国高校产学研创新基金项目(2021ITA07021)

作者简介：梁凯(1976—)，男，副教授，研究方向为噪声控制、深度学习、模式识别；赵海军(1974—)，通信作者，男，教授，博士，研究方向为车辆噪声控制、乘用车环境评估。

based on the leave-one-out method(LOO) show the LOO accuracy of the native audio and two-dimensional spectrogram GAN models are both greater than or close to 50%, indicating that model training does not produce overfitting, the method proposed in this paper is true and reliable in generating sound effects.

Key words: electric vehicles; initiative sound production; generative adversarial network; raw audio; spectrum

随着电动汽车产业的迅速发展,电动汽车在普及使用过程中逐渐暴露出一些问题,如低速行驶时因声音幅值较低而存在交通安全隐患、拟音不能满足用户个性化需求等,这些问题受到行业研究者的广泛关注。很多国家的相关协会起草了有关电动汽车低速行驶的噪声标准[1],很多研究人员设计了汽车主动发声系统,以更准确地拟合内燃发动机的声音。

文献[2]提出了基于采集数字音频信号高阶数据的多参数控制拟合算法,根据发动机转速、油门等参数,实现了模拟发动机声音的效果;文献[3]基于语音合成技术的叠加理论,设计了一种根据发动机转速自适应的引擎主动发声系统,应用效果较好;文献[4]提出了一种基于正弦波的发动机声音拟合系统,该系统能够较真实地模拟出特定发动机的声音;文献[5]采用短时傅里叶变换和 Kaiser 窗理论分析了多阶发动机谐波频率,合成了加速行驶时的车内声音,满足电动汽车主动发声系统的发动机阶次声音合成精度要求。上述研究中的发动机主动发声模型均基于特定内燃机的声音信号特点,采用矢量化信号处理算法,实现自适应声音拟合,但普遍存在三个方面问题:第一,声音信号的矢量化处理算法将原始声音划分为音频帧,求解对应帧之间的映射关系,可能出现声音特征参数缺乏连续性的情况,真实度较差;第二,未考虑电动汽车的动力整体特性,易导致引擎参数抖动时声音拟合平滑度变差;第三,上述模型不能满足不同种类内燃机声音拟合的多样化需求。

生成对抗网络(generative adversarial networks, GAN)在图像生成、语义分割、语音生成等方面已有广泛应用。文献[6]提出了基于 GAN 的图像生成算法,在 MNIST 等数据集上的训练试验结果显示,生成的图片人工辨识度很高;文献[7]将 GAN 应用于图像背景处理,可有效去除图像背景降雨条纹;文献[8]将 GAN 应用于无监督合成音频波形,直接生成人类语音片段。目前还未见 GAN 在电动汽车主动发声系统方面应用的研究。

本文以几种内燃机不同工况的声信号作为样本,建立基于 GAN 的主动发声模型,将原始音频信号和经过频域变换的梅尔谱特征的内燃机声信号样本输入模型训练,以获得最优的引擎声音合成模型。

1 生成对抗网络

1.1 GAN 原理

GAN 是一种深度学习模型,其通过构建判别器 D 和生成器 G 并使之协同工作完成训练[8]。生成器用于生成样本数据,并与真实样本一起送入判别器进行训练,目标是尽可能生成好的样本;判别器的目标是尽可能识别真实样本,拒绝生成样本。GAN 原理如图 1 所示。

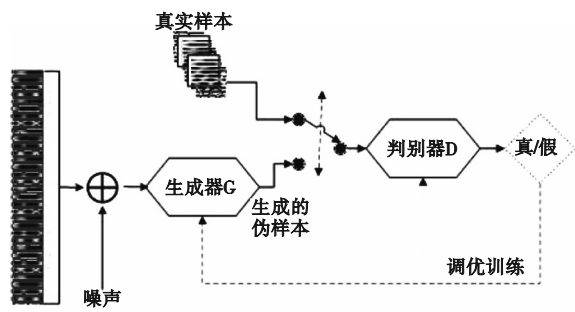

图 1 GAN 原理

Fig. 1 Principles of GAN

GAN 的底层模型是多层感知机,可以学习到低维潜在变量 $z(z \in \boldsymbol{Z}, \boldsymbol{Z}$ 为独立同分布样本的先验变量集合)到真实数据集 $\boldsymbol{X}$ 中点 x 的映射关系。判别器 D 的损失函数为二元分类器的正则交叉熵损失函数,其计算式为

$$l_{\mathrm{D}} = -[y\log p + (1-y)\log(1-p)] \quad (1)$$

式中:l_{D} 为判别器的损失函数;y 为样本真实度的期望值,对于真实样本,y 为 1,对于假样本,y 为 0;p 为真实样本的预测概率,$1-p$ 是假样本的预测概率。若用 $G(z)$ 表示生成样本 x,$D(x)$ 表示 p,则判别器的损失函数可表示为

$$l_{\mathrm{D}} = -[y\log(D(x)) + (1-y)\log(1-D(G(z)))] \quad (2)$$

生成器 G 的目标是最大化判别器 D 的损失

函数值,式(2)中第一项与生成器无关,故生成器 G 的损失函数 l_G 可表示为

$$l_G = (1-y)\log(1-D(G(z))) \tag{3}$$

生成器 G($\boldsymbol{Z}\mapsto\boldsymbol{X}$)和判别器 D($\boldsymbol{X}\mapsto[0,1]$)为竞争关系,两者的联合目标函数 $V(D,G)$ 可表示为

$$\min_G\max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{4}$$

式中:$p_{\text{data}}(x)$是 x 的概率分布;$p_z(z)$是 z 的概率分布;$\mathbb{E}_{x\sim p_{\text{data}}(x)}$和$\mathbb{E}_{z\sim p_z(z)}$分别表示 x 和 z 的数学期望。生成器 G 可采用全连接神经网络或卷积神经网络,其可通过噪点概率分布 $p_z(z)$得到一个生成数据概率分布 $p_g(x)$,使之尽可能接近 $p_{\text{data}}(x)$,即生成样本尽可能少被判别为假。判别器 D 的训练目标是最大限度鉴别生成样本为假样本,即最大化 $p_g(x)$和 $p_{\text{data}}(x)$之间的差距。

1.2　GAN 求解

式(4)中的目标函数 $V(D,G)$是连续的,采用积分形式表示期望,可得式(5)。

$$V(D,G) = \int_{-x}^{x} p_{\text{data}}(x)\log(D(x))\mathrm{d}x + \int_{-x}^{x} p_z(G^{-1}(x))\log(1-D(x))(G^{-1})'(x)\mathrm{d}x \tag{5}$$

$p_g(x)$为 z 的生成数据的概率分布,表示为

$$p_g(x) = p_z(G^{-1}(x))(G^{-1})'(x) \tag{6}$$

将式(6)代入式(5),可得式(7)。

$$V(D,G) = \int_{-x}^{x} p_{\text{data}}(x)\log(D(x))\mathrm{d}x + \int_{-x}^{x} p_g(x)\log(1-D(x))\mathrm{d}x \tag{7}$$

目标函数中 $D(x)$的最优解推导过程如式(8)~(10)所示。

$$\frac{\partial}{\partial D(x)}(p_{\text{data}}(x)\log(D(x)) + p_g(x)\log(1-D(x))) = 0 \tag{8}$$

即

$$\frac{p_{\text{data}}(x)}{D(x)} - \frac{p_g(x)}{1-D(x)} = 0 \tag{9}$$

解得

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)} \tag{10}$$

式中 $D^*(x)$表示 $D(x)$的最优解。当 $p_{\text{data}}(x) = p_g(x)$时,$D^*(x) = 1/2$。将该最优解表达式代入式(7),优化求解生成器 G,此时目标函数 $C(G)$表达式为

$$C(G) = \int_{-x}^{x}\left[p_{\text{data}}(x)\log\left(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}\right) + p_g(x)\log\left(\frac{p_g(x)}{p_{\text{data}}(x)+p_g(x)}\right)\right]\mathrm{d}x \tag{11}$$

将式(11)整理为连续函数的 KL 散度形式为

$$C(G) = \mathrm{KL}\left[p_{\text{data}}(x)\,\|\,\frac{p_{\text{data}}(x)+p_g(x)}{2}\right] + \mathrm{KL}\left[p_g(x)\,\|\,\frac{p_{\text{data}}(x)+p_g(x)}{2}\right] - \log 4 \tag{12}$$

KL 散度为非负数,由式(12)可知,当 $p_g(x) = p_{\text{data}}(x)$时,$C(G)$取得全局最小值 $-\log 4$。

2　主动发声系统设计

本文提出一个基于自动评价效果的 GAN 模型用于生成内燃机引擎声音,其主要流程为:将声音样本和标签预处理后分为训练集和测试集两部分,将训练集数据和对应的标签作为原始输入训练 GAN 模型,使用测试集验证训练后模型,从而筛选、保存最优模型,最后保存的 GAN 生成器模型用来生成新的声音样本。主动发声系统原理如图 2 所示。

2.1　原生音频 GAN 模型设计

受深度卷积对抗生成网络(DCGAN)的启发,为保证 GAN 模型训练的可控性和成功率,模型中的生成器和判别器均采用卷积神经网络结构,生成器的卷积层称为转置卷积(Trans Conv)层[9],是对特征图的上采样过程,类似普通卷积层的反向梯度计算。根据原生音频的特点将原 DCGAN 中 5×5 二维卷积核调整成长度为 25 的一维卷积核,步长也同样从 2×2 调整为 4。综合考虑音频样本长度和时间效率,将原生音频的 GAN 网络结构调整为:生成器包含输入层、全连接层和 5 个卷积层,卷积层之间均使用 ReLU 激活函数。由于原生音频输入为一维向量,故卷积核仅用宽度表示。原生样本 GAN 生成器具体结构如表 1 所示。

原生样本判别器包含 5 个卷积层和 1 个输出层,输出层为全连接层。各层之间的激活函数使用 LReLU 函数,各卷积层之间增加相位转换与调整操作,最后为重构层和全连接层。原生样本 GAN 判别器具体结构如表 2 所示。表中:α 为 Leaky 值;n 表示相位转换因子。

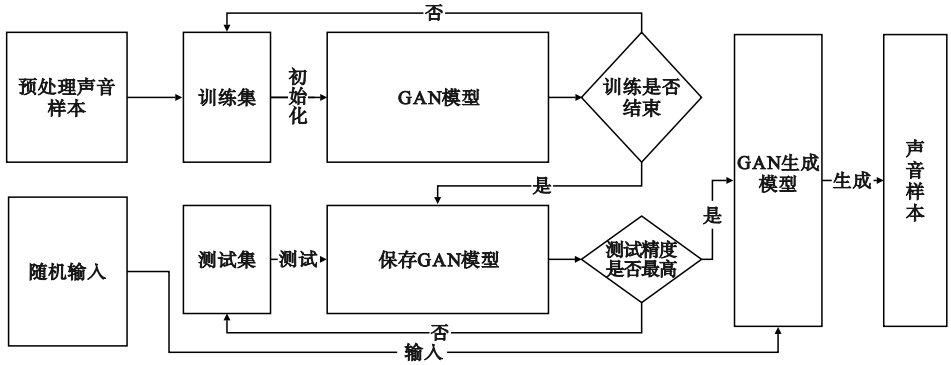

图 2 主动发声系统原理

Fig. 2 Schematic diagram of active generation system

表 1 原生样本 GAN 生成器结构

Table 1 Native sample GAN generator structure

层名称	类型	卷积核大小	卷积核数量
输入层	输入层	输入层	—
全连接层	全连接	1 × 100	16 384
ReLU	激活层	—	—
Conv1D（步长为 4）	卷积层	1 × 25	512
ReLU	激活层	1 × 25	
Conv1D（步长为 4）	卷积层	1 × 25	256
ReLU	激活层	—	—
Conv1D（步长为 4）	卷积层	1 × 25	128
ReLU	激活层	—	—
Conv1D（步长为 4）	卷积层	1 × 25	64
ReLU	激活层	—	—
Conv1D（步长为 4）	卷积层	1 × 25	1
ReLU	激活层	—	—

表 2 原生样本 GAN 判别器结构

Table 2 Native sample GAN discriminator structure

层名称	类型	卷积核大小	卷积核数量
输入层 $G(z)$	输入层	—	—
Conv1D（步长为 4）	卷积层	1 × 25	64
LReLU（$\alpha = 0.2$）	激活层	—	—
相位转换	相位调整	—	—
Conv1D（步长为 4）	卷积层	1 × 25	128
LReLU（$\alpha = 0.2$）	激活层	—	—
相位转换（$n = 2$）	相位调整	—	—
Conv1D（步长为 4）	卷积层	1 × 25	256
LReLU（$\alpha = 0.2$）	激活层	—	—
相位转换（$n = 2$）	相位调整	—	—
Conv1D（步长为 4）	卷积层	1 × 25	512
LReLU（$\alpha = 0.2$）	激活层	—	—
相位转换（$n = 2$）	相位调整	—	—
Conv1D（步长为 4）	卷积层	1 × 25	1 024
LReLU（$\alpha = 0.2$）	激活层	—	—
Reshape	重构层	—	—
全连接层	全连接	16 384 × 1	64

2.2 声谱图 GAN 模型设计

根据声谱图的二维特性，仍使用 DCGAN 的原生结构，卷积核的大小为 5 × 5，步长为 2。综合考虑音频样本时长和性能，确定样本声谱图的 GAN 网络结构为：生成器包含全连接层和 5 个卷积层，卷积层之间均使用 ReLU 激活函数，最后一层使用 Tanh 激活函数，由于声谱图输入为二维数据，故卷积核大小采用长度 × 宽度来表示。基于样本声谱图 GAN 生成器具体结构如表 3 所示。

判别器包含 5 个卷积层、1 个重构层和 1 个输出层，卷积层之间为激活层，激活函数使用 LReLU 函数，最后为重构层和输出层，输出层为全连接层。每层的卷积核尺寸均为 5 × 5，卷积核的数量与表 2 中对应的每层卷积核数量相等。

3 样本生成试验

3.1 声音样本处理

音频样本的主成分分析表现为周期性特点，因此对于长音频信号样本处理时分解为相应组成频带。文献[10 – 11]的研究表明，在半监督的音频分类中使用音频原生时域信号来完成训练，分类精确度未受影响。车辆发动机启动时产生的声音信号是随着时间变化的非平稳信号，可能包含了启动马达和发动机转速的音频特征，因此在短时音频样本处理中，采用二维时频域声谱图信号和原生时域信号处理方式做对照试验。

表 3　基于样本声谱图 GAN 生成器结构

Table 3　Structure of GAN generator based on sample spectrogram

层名称	类型	卷积核大小	卷积核数量
输入层	输入层	—	—
全连接层	全连接	1×100	16 384
ReLU	激活层	—	—
Trans Conv2D(步长为 2)	卷积层	5×5	64×8
ReLU	激活层	—	—
Trans Conv2D(步长为 2)	卷积层	5×5	64×4
ReLU	激活层	—	—
Trans Conv2D(步长为 2)	卷积层	5×5	64×2
ReLU	激活层	—	—
Trans Conv2D(步长为 2)	卷积层	5×5	64×1
ReLU	激活层	—	—
Trans Conv2D(步长为 2)	卷积层	5×5	1
Tanh	激活层	—	—

内燃机工作声音样本录制环境为普通的实验室环境,共采集 1 200 个声音样本作为试验用样本集。样本库包含丰田 HR16DE 汽油机、现代 D4BH 柴油机、三菱 4G6 MIVEC 汽油机共三种型号的内燃机启动、转速从 800 r/min 到4 500 r/min 共 40 组稳态声音信号。每个声音样本处理为采样率 16 kHz、长度 1 s、单声道。

原生时域声音样本处理时将真实的样本信号波形转化为一维向量,并采用最值归一法归一化,如式(13)所示。

$$\boldsymbol{X}^{*}=\frac{\boldsymbol{X}}{\max(\boldsymbol{X})} \tag{13}$$

式中:$\boldsymbol{X}$ 代表声音样本集合;$\boldsymbol{X}^{*}$ 代表归一化后的声音样本集合。

二维时频域声谱图处理采用短时傅里叶变换(STFT)方法转化音频信号[12],表达式为

$$\mathrm{STFT}_{s}(t,f)=\int_{-\infty}^{+\infty}s(u)w^{*}(u-t)\mathrm{e}^{-\mathrm{j}2\pi fu}\mathrm{d}u \tag{14}$$

式中:u 表示时间 t 的偏移量;f 为输入频率;$s(u)$ 表示 u 时刻的连续信号;$w^{*}(u-t)$ 为随时间 t 变化的窗函数,上标"*"表示复共轭。式(14)最终计算结果为任意时刻下该信号所包含各频率的幅值、相位等数据。$w(t)$ 窗函数计算如式(15)所示。

$$w(t)=\frac{1}{2}\left[1-\cos\left(\frac{2\pi t}{N-1}\right)\right]R_{N}(t) \tag{15}$$

$$R_{N}(t)=\begin{cases}1, 0\leqslant t\leqslant N-1\\ 0, \text{其他}\end{cases} \tag{16}$$

式中 N 为窗口长度。

信号预处理过程中以 16 ms 间隔加窗分帧,重叠量为 8 ms,经 STFT 获得信号为 129×1 999 的二维矩阵。为保证得到的声音频谱大小合适,还需使用梅尔滤波器[13]处理二维时频信号,梅尔滤波器的计算如式(17)所示。

$$\mathrm{mel}(f)=2\ 595\times\lg\left(1+\frac{f}{700}\right) \tag{17}$$

式中 $\mathrm{mel}(f)$ 代表 f 的梅尔值。

通过梅尔滤波器处理后的信号能较好反映人耳对频率的敏感性,即在低频区域梅尔值增长速度较快,在高频区域梅尔值增长速度缓慢。处理后的梅尔谱图如图 3 所示。

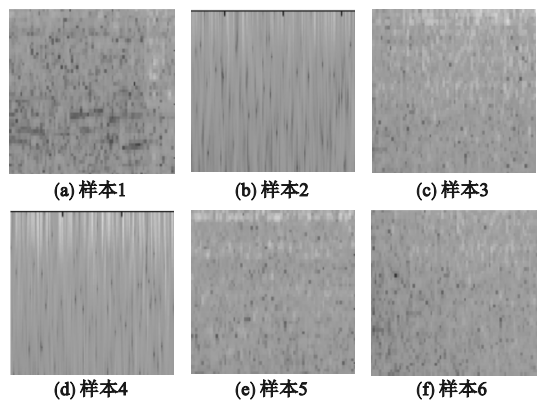

图 3　部分样本的梅尔谱图

Fig. 3　Mel spectra of some samples

采用式(18)对样本数据归一化处理,使数据集的均值为 0、标准差为 1,处理后的数据值均在[-1,1]之间。

$$X_{i}^{*}=\frac{X_{i}-\mu}{\sigma} \tag{18}$$

式中:X_i 和 X_i^{*} 分别表示归一化前后样本 i 的数据;μ 为所有样本数据的均值;σ 为所有样本数据的方差。处理过的数据经人耳听力测试,样本没有产生听觉差异。

3.2　模型训练与调优

为保证试验结果的可比较性,原生声音和声谱图输入均采用 1 s 时长的样本,生成器的输入均为 100 维的潜在向量。原始样本的生成器训练采用表 1 所示结构,全连接层将输入 100 维噪声向量样本变换为 16×1 024 的特征图,然后按照相应的卷积核长度和数量进行反卷积操作,经过 5 次反卷积操作和激活操作后得到16 384维向量,作

为判别器的输入,进行下一步操作。原生样本的判别器训练采用表 2 所示结构,将生成器产生的 16 384 维向量和真实样本读取的 16 384 维数据统一作为输入,进行卷积操作和相应的激活操作。激活函数采用 LReLU 函数[14],该函数在输入小于 0 时仍有梯度变化,可以减轻普通 ReLU 函数的稀疏性,此处 α 取为 0.2。判别器经过 5 次卷积操作后,连接全连接层和二元分类器,判别样本的真伪。

原生样本判别器训练过程中,由于真实数据存在频率重叠,上采样操作不可避免地会产生音调噪声,类似二维图像反卷积操作造成“棋盘”伪影现象[15]。由于音调噪声常出现在特定阶段,判别器很容易学习到拒绝这些噪声样本的规则,从而抑制整体优化,降低判别器性能。为解决该问题,在训练中提出一种相位扰动操作,通过 m 个样本随机扰动每一激活层数据的相位,使特征图在输入到下一层之前统一化,从而提升判别器抗噪能力。

相位扰动操作是在判别器的每一层中将特征图左或右边界映射填充到上采样后样本缺失的部分,从而得到均匀的样本。特征图相位扰动操作示意如图 4 所示,图中表示了 5 个特征图在 $m=2$ 时所有的输出可能。

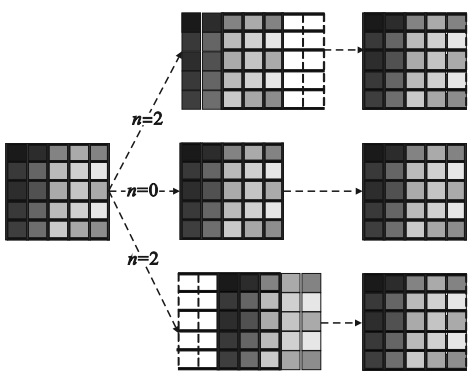

图 4 特征图相位扰动操作示意

Fig. 4 Schematic diagram of feature map phase perturbation operation

声谱图样本输入训练过程中,生成器训练采用表 3 所示结构,全连接层将输入 100 维噪声向量样本变换为 4 × 4 × 1 024 的特征图,然后采用 5 × 5 的二维卷积核(卷积核数量逐层递减)进行反卷积操作,经过 5 次反卷积操作和激活操作,最后得到 128 × 128 的二维声谱图作为判别器的生成样本输入。同样,判别器训练中将生成器产生的样本声谱图和真实样本读取的声谱图统一作为输入,进行卷积操作和相应的激活操作,采用与原生样本判别器训练中相同的激活函数。判别器输出处理与原生样本判别器相同。

为保证两种模型方案的可比较性,训练中每批均选取 64 个样本预测梯度,学习率均选择为 0.000 2。为快速处理凸函数的稀疏梯度问题,训练的梯度下降优化均采用自适应时刻估计(Adam)算法[16],Adam 优化器参数均设置为 0.5。

3.3 试验过程

试验采用 NVIDIA GeForce GTX 1070 GPU 和 CUDA9.0 计算环境,选择样本集中 80% 的样本作为训练集,10% 作为验证集,其余 10% 作为测试集。训练 80 个批次,约 9 h 后趋于收敛。损失曲线如图 5 所示,图中 G_W 和 G_S 分别代表原始样本和声谱图样本的生成器训练损失曲线,D_W 和 D_S 分别代表原始样本和声谱图样本的判别器训练损失曲线。由图 5 可见,训练 20 批次时,两对生成器和判别器的训练损失函数值均基本趋于稳定。

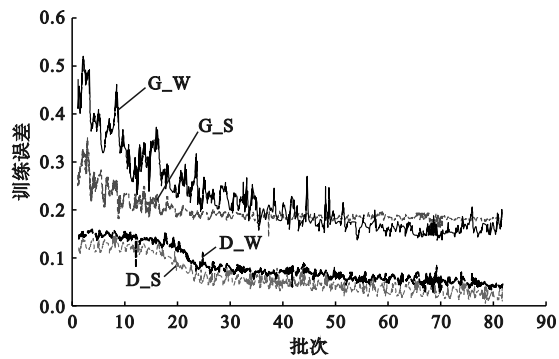

图 5 损失曲线

Fig. 5 Loss curves

4 模型评估与分析

为准确验证训练的模型质量,对于训练生成的样本分别采用定性和定量的评价方法。

4.1 人工评价

在定性评价中采用人工听音评价方法。随机抽取由原生音频 GAN 模型和声谱图 GAN 模型生成的声音样本各 10 个,与 20 个真实声音混合,分两组放置在声品质评价系统中,由试听者分别投票辨识各组声音样本的真实性,评价中共有 25 人参与,评价结果如图 6 所示。样本属性值为 0 表示生成样本,样本属性值为 1 表示真实样本,投票结果显示大部分样本的真实度都在 90% 以上。因此,普通人的听觉不能较好区分出两种样本之间

的差异,也不能有效评价出原始音频模型和声谱图模型之间的差别。

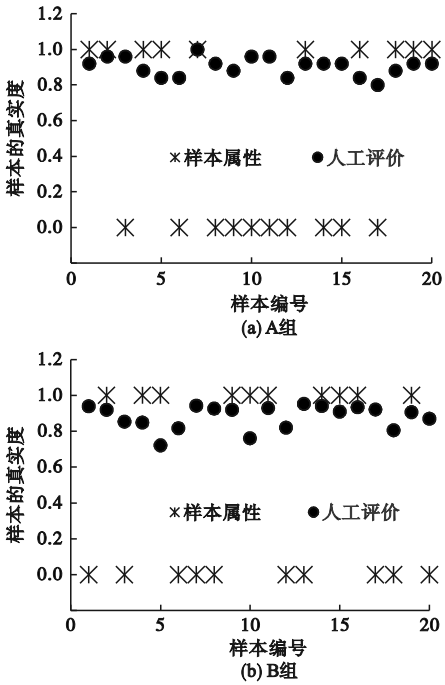

图 6　定性评价结果

Fig. 6　Qualitative evaluation results

4.2　基于 1-NN 分类器评价

样本质量的定量评价采用基于 1-NN 分类器的留一法(LOO)精度[17]评价方法。如果模型生成样本足够好,生成样本分布与真实样本分布完美匹配,经 1-NN 分类器得到的 LOO 精度应约为 50%,即无论如何分配验证集和训练集,1-NN 分类器都只有 50% 概率预测正确。本文定量评价中使用所有 120 个真实样本作为正样本,120 个生成样本作为负样本,使用 LOO 法循环训练 1-NN 分类器。评价结果如图 7 所示,可见两种模型所有验证集样本的 LOO 精度均处于增长趋势,说明模型可靠性较高。

两种模型生成的样本详细评价数据如表 4 所示。由表 4 可见:所有验证集的 LOO 精度均大于或者接近 0.5,说明原生音频 GAN 模型和声谱图 GAN 模型的训练过程都没有产生过度拟合;原生音频 GAN 模型和声谱图 GAN 模型的真实样本验证集 LOO 精度都低于生成样本验证集,表明两种 GAN 模型都能够捕捉到发动机引擎的音频特征。

由于大多数真实样本周围都充满着生成样本,导致两种 GAN 模型真实样本验证集的 LOO 精度较低。由于生成样本倾向于聚集在一起,即测出的真实样本和生成样本的最近邻分布实际都存在于生成样本和生成样本之间,在其各自作为

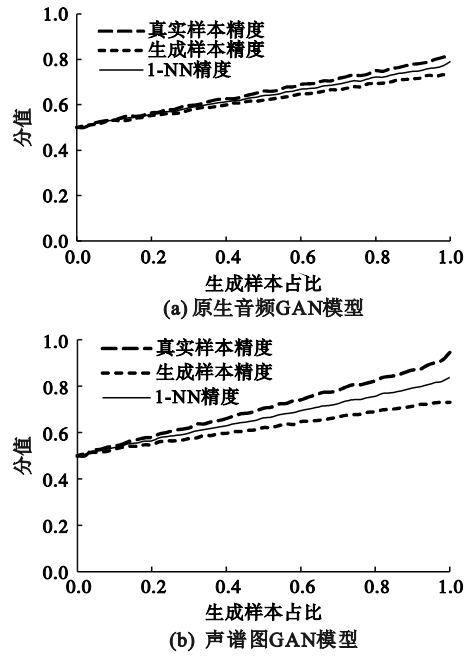

图 7　定量评价结果

Fig. 7　Quantitative evaluation results

表 4　生成样本的对比评价结果

Table 4　Comparison evaluation results of generated samples

验证集	真实样本	原生音频 GAN	声谱图 GAN
混合样本	0.485	0.791	0.813
真实样本	0.495	0.735	0.881
生成样本	0.504	0.814	0.945

验证集时,判别器均会做出负样本的正确判定[18],因此生成样本验证集的 LOO 精度较高。由表 4 中还可以看出,声谱图 GAN 模型的生成样本作为验证集时 LOO 精度为 0.945,较原生音频 GAN 生成样本验证集 LOO 精度高 0.131,说明声谱图 GAN 模型在训练时可能产生了部分模型坍塌,没有完全学习到所有样本的真实分布。虽然人耳听觉无法区分生成的声音真假,但存在生成的声音样本类型不足的问题,在一定程度上与训练样本数量、类型过少有一定的关系,今后研究中将考虑增加训练样本数量和类型。

5　结论

1)提出了一种基于 GAN 的电动汽车主动发声模型,合理设计了生成器网络和判别器网络的层次结构。试验中将采集到的内燃机启动声音样本作为模型的输入训练 GAN 网络模型,经 1-NN 分类评价显示,模型可准确地学习到原始音频信

号的特征分布,经人耳听觉测试显示,生成的声音样本仿真度较高。

2)GAN 模型的原生音频样本输入与二维声谱图输入的对照试验显示,声谱图 GAN 模型在训练时可能产生了模型坍塌,原生音频 GAN 模型的生成样本质量高于二维声谱图 GAN 模型。

参考文献(References):

[1] GENUIT K,BRAY W R. Prediction of sound and vibration in a virtual automobile[J]. Sound & Vibration,2002,36(7):12 - 19.

[2] CAO Y T,HOU H S,LIU Y J,et al. Engine order sound simulation by active sound generation for electric vehicles[J]. SAE International Journal of Vehicle Dynamics,Stability,and NVH,2020,4(2):151 - 164.

[3] JAGLA J,MAILLARD J,MARTIN N. Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-and-add method[J]. The Journal of the Acoustical Society of America,2012,132(5):3098 - 3108.

[4] MIN D,PARK B,PARK J. Artificial engine sound synthesis method for modification of the acoustic characteristics of electric vehicles[J]. Shock and Vibration,2018,2018:1 - 8.

[5] 曹蕴涛. 电动汽车主动发声系统设计与评价方法研究[D]. 长春:吉林大学,2019.

[6] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks[J]. Communications of the ACM,2020,63(11):139 - 144.

[7] JIN X,CHEN Z B,LI W P. AI-GAN:asynchronous interactive generative adversarial network for single image rain removal[J]. Pattern Recognition,2020,100:107143.

[8] VAN DEN OORD A,DIELEMAN S,ZEN H,et al. WaveNet:a generative model for raw audio[EB/OL]. arXiv:1609.03499(2016 - 09 - 19)[2023 - 07 - 01]. https://arxiv.org/abs/1609.03499.

[9] LI X,MAKIHARA Y,XU C,et al. Gait recognition invariant to carried objects using alpha blending generative adversarial networks[J]. Pattern Recognition,2020,105:107376.

[10] 郭川磊,何嘉. 基于转置卷积操作改进的单阶段多边框目标检测方法[J]. 计算机应用,2018,38(10):2833 - 2838.
GUO C L,HE J. Improved single shot multibox detector based on the transposed convolution[J]. Journal of Computer Applications,2018,38(10):2833 - 2838. (in Chinese)

[11] WANG X,YAMAGISHI J. Investigating self-supervised front ends for speech spoofing countermeasures[EB/OL]. arXiv:2111.07725(2022 - 02 - 04)[2023 - 07 - 01]. https://arxiv.org/abs/2111.07725.

[12] VIETING P,LÜSCHER C,MICHEL W,et al. On architectures and training for raw waveform feature extraction in ASR[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU). Cartagena,Colombia:IEEE,2022:267 - 274.

[13] 梁凯,赵海军,宋伟志. 基于卷积神经网络的内燃机声品质评价方法研究[J]. 内燃机工程,2019,40(2):67 - 75.
LIANG K,ZHAO H J,SONG W Z. Research on evaluation method of internal combustion engine sound quality based on convolutional neural network[J]. Chinese Internal Combustion Engine Engineering,2019,40(2):67 - 75. (in Chinese)

[14] 吴新忠,夏令祥,张旭,等. 基于谱熵梅尔积的语音端点检测方法[J]. 北京邮电大学学报,2019,42(2):83 - 89.
WU X Z,XIA L X,ZHANG X,et al. Voice activity detection method based on MFPH[J]. Journal of Beijing University of Posts and Telecommunications,2019,42(2):83 - 89. (in Chinese)

[15] DAUBECHIES I,DEVORE R,FOUCART S,et al. Nonlinear approximation and(deep)ReLU networks[J]. Constructive Approximation,2022,55(1):127 - 172.

[16] REN G L,GENG W J,GUAN P Y,et al. Pixel-wise grasp detection via twin deconvolution and multi-dimensional attention[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology,2023,33(8):4002 - 4010(2023 - 01 - 18)[2023 - 07 - 01]. https://ieeexplore.ieee.org/document/10019258. DOI:10.1109/TCSVT.2023.3237866.

[17] IRFAN D,ROSNELLY R,WAHYUNI M,et al. Perbandingan optimasi SGD,ADADELTA,dan Adam dalam klasifikasi hydrangea menggunakan CNN[J]. Journal of Science and Social Research,2022,5(2):244 - 253.

[18] NAVIDAN H,MOSHIRI P F,NABATI M,et al. Generative adversarial networks(GANs)in networking:a comprehensive survey & evaluation[J]. Computer Networks,2021,194:108149.

(责任编辑:宋颖韬)