文章编号：1003-0530(2022)09-1988-11

基于HiFi-GAN的改进型高效声码器

唐 君 张连海 李嘉欣 李宜亭
（中国人民解放军战略支援部队信息工程大学信息系统工程学院，河南郑州 450001）

摘 要： HiFi-GAN声码器通过采用缩减网络层的通道数或层数的方式来有效减少模型参数、提高推理速度，但此种方式也严重损害了生成语音的质量。针对此问题，提出了两点改进措施：1. 采用多尺度卷积策略对输入Mel谱进行处理来有效表征特征信息；2. 采用一维深度可分离卷积替换生成器网络中的标准一维卷积。实验结果表明，多尺度卷积策略有效提升了模型性能，提高了生成语音的质量，而一维深度可分离卷积显著减少了模型参数量并加快了模型推理速度。通过将这两者结合，有效提升了HiFi-GAN模型的性能，具体来说，模型参数量约减少了67.72%，在GPU、CPU上的推理速度分别提升了11.72%、28.98%。此外，语音质量也得到略微提升，平均主观意见分（Mean Opinion Score，MOS）提升了0.07，客观语音质量评估（Perceptual Evaluation of Speech Quality，PESQ）得分提升了0.05。

关键词： 语音合成；声码器；HiFi-GAN；深度可分离卷积；多尺度卷积策略
中图分类号： TN912.33 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2022.09.021

引用格式： 唐君，张连海，李嘉欣，等. 基于HiFi-GAN的改进型高效声码器[J]. 信号处理，2022，38(9)：1988-1998. DOI：10.16798/j.issn.1003-0530.2022.09.021.
Reference format： TANG Jun，ZHANG Lianhai，LI Jiaxin，et al. Improved high-efficiency vocoder based on HiFi-GAN [J]. Journal of Signal Processing，2022，38(9)：1988-1998. DOI: 10.16798/j.issn.1003-0530.2022.09.021.

Improved High-efficiency Vocoder Based on HiFi-GAN

TANG Jun ZHANG Lianhai LI Jiaxin LI Yiting
（School of Information System Engineering，PLA Strategic Support Force Information Engineering University，Zhengzhou，Henan 450001，China）

Abstract: The HiFi-GAN vocoder effectively reduces model parameters and improves inference speed by reducing the number of channels or layers of the network layer，but this method also seriously damages the quality of the generated speech. To solve this problem，two improvement measures are proposed：1. Multi-scale convolution strategy is used to process the input Mel spectrum to effectively characterize the feature information；2. One-dimensional depthwise separable convolution is used to replace standard one-dimensional convolutions in the generator network. Experimental results show that the multi-scale convolution strategy effectively improves the model performance and the quality of the generated speech，while the one-dimensional depthwise separable convolution significantly reduces the amount of model parameters and speeds up the inference speed. By combining the two，the performance of the HiFi-GAN model is effectively improved. Specifically，the amount of model parameters is reduced by approximately 67.72%，and the inference speed on the GPU and CPU are increased by 11.72% and 28.98%，respectively. In addition，voice quality has also been slightly improved，the Mean Opinion Score（MOS）increased by 0.07 and Perceptual Evaluation of Speech Quality（PESQ）score increased by 0.05.

Key words: speech synthesis；vocoder；HiFi-GAN；depthwise separable convolution；multi-scale convolution strategy

收稿日期：2022-01-17；修回日期：2022-03-15
基金项目：国家自然科学基金资助项目(61673395，62171470)

1　引言

近年来，基于神经网络的语音合成技术得到了迅速发展，合成语音的质量甚至已接近了人类录音的水平。这类语音合成系统通常可分为两个模块实现，第一个模块，根据输入文本预测声学特征，通常称为合成器模块；第二个模块，根据预测的声学特征生成语音波形，该模块通常称为声码器模块，本文聚焦于声码器模块的研究。传统的声码器，如STRAIGHT[1]、WORLD[2]等，它们生成语音的质量通常很低并且听起来很不自然，而基于神经网络的声码器（简称神经声码器）在利用梅尔（Mel）谱生成语音波形方面取得了显著的成绩，其生成语音的质量远远超过传统声码器生成语音的质量。因此，神经声码器技术已经成为了目前的主流声码器技术。

根据神经声码器结构的不同，该类声码器可分为两大类：自回归模型和非自回归模型。起初，神经声码器研究主要集中在自回归模型，如WaveNet[3]、WaveRNN[4]和SampleRNN[5]等。自回归模型的基本思想是将语音波形的分布分解为多个条件分布的乘积进行建模，即当前的采样点依赖于先前生成的采样点来生成，以建模语音波形间的长期相关性。虽然这类模型能够生成高质量的语音，但这种顺序推理过程导致其生成语音的速度非常慢、效率低下，无法满足实时应用的要求。

为解决自回归模型因其结构而带来的局限性，非自回归模型应运而生。非自回归模型直接建模语音波形的联合分布，因此这类模型具有高度的并行性，其推理速度比自回归模型快得多。近几年来，非自回归声码器得到了迅速发展，根据其方法，可分为四类：一、基于知识蒸馏的模型，如Parallel WaveNet[6]和ClariNet[7]等。在这个模型框架下，自回归教师模型的知识被转移到基于逆自回归流[8]（Inverse Autoregressive Flow，IAF）的学生模型，虽然基于IAF的学生模型能够快速生成合理感知质量的语音，但是这种方法不仅需要训练有素的教师模型，而且还需要一些策略来优化复杂的概率密度蒸馏过程，如使用基于Kulback-Leiber（KL）散度的概率蒸馏目标以及附加的感知损失项来训练学生模型。此外，由于学生模型参数庞大，还需要GPU进行推理才能达到实时性；二、基于流的模型，如WaveGlow[9]、WaveFlow[10]等，它们由一个可逆网络实现，仅通过最小化训练数据的负对数似然损失来直接学习，该类方法通常需要庞大的参数量和繁重的计算量，其生成语音质量才能与自回归模型相当，虽然它们在GPU上推理速度很快，但该类方法通常不适用于内存受限、硬件条件不足的场景下；三、基于扩散概率模型，如DiffWave[11]和WaveGrad[12]等。这类模型以迭代的方式利用具有固定步数的马尔可夫链将白噪声信号转换成结构化语音信号，因此它们需要一定的模型参数和迭代次数才能生成较高质量的语音。虽然这类模型的推理速度相比自回归模型快得多，但是相比基于流的模型却要慢得多，也只是能在GPU上实现较低的实时性；四、基于生成对抗网络（Generative Adversarial Network，GAN）的模型，它们是目前最有前景的方法之一。Parallel WaveGAN[13]和MelGAN[14]是GAN在声码器上的早期尝试，相比于其他模型，它们显著提高了模型推理速度（Parallel WaveGAN在GPU上实现较高的实时性，而MelGAN不仅在GPU上实现了很高的实时性，在CPU上也能实时生成语音），然而它们生成的语音质量上并不令人满意。为此，研究者作了进一步的努力，如Multi-Band MelGAN[15]改进了MelGAN，进一步提高了MelGAN生成语音的质量和速度，VocGAN[16]改善了MelGAN中存在的生成语音质量不足以及生成语音的Mel谱与输入Mel谱的声学特征不一致的问题。LVCNet[17]则改进了Parallel WaveGAN，提出了一种位置变量卷积用于建模语音波形序列的长期相关性，在不降低音质和增加参数的情况下，合成速度相比Parallel WaveGAN提升了4倍左右。以上基于GAN的声码器研究，虽然实现以较高的实时速度生成语音波形，甚至在CPU上也能实现实时推理，但是这些模型生成语音的质量始终与自回归模型有所差距，但HiFi-GAN[18]的出现打破了这种桎梏，它实现了高效和高保真的语音波形生成。从本质上来说，HiFi-GAN的成功主要得益于其不仅有效建模了语音波形的长期相关性，更重要的是其有效建模了语音波形的周期模式，这也是之前基于GAN的声码器所欠缺的。此外，HiFi-GAN相比于其他模型在端到端语音合成系统中兼容性更好，很多端到端语音合成系统采用HiFi-GAN模型作为声码器来将前端预测的Mel谱转换为语音波形。

HiFi-GAN作为目前最先进的声码器网络之一，但其仍存在一些不足：无法在语音质量和模型参数、推理速度上进行很好的权衡，通常采用缩减网络层的通道数或层数的方式来减少模型参数、提高推理速度，但这种方式需要牺牲较大的语音质量去换取模型参数的减少和推理速度的提升。

为了更好权衡HiFi-GAN模型在语音质量和模型参数、推理速度上的关系，本文在HiFi-GAN的基础上，引入了多尺度卷积策略和深度可分离卷积，以期望在不明显降低其生成语音质量的情况下，显著减少了HiFi-GAN的参数量，并进一步提高了其推理速度。

2 模型框架

在本节中，首先介绍一维深度可分离卷积的原理，然后介绍生成器的结构原理并将多尺度卷积策略和一维深度可分离卷积引入其中，接着再介绍判别器的结构原理，最后介绍用于训练生成器和判别器的损失函数。

2.1 一维深度可分离卷积

在这一小节中，先介绍标准一维卷积的原理，再借此引入深度可分离卷积，并对两者的参数量和计算量进行了比较。

标准一维卷积的原理如图1所示，对于大小为$T_1 \times C_1$的输入（如Mel谱，T_1代表帧数，C_1代表Mel谱的维度，即通道数），卷积核的尺寸为K（标准一维卷积默认对输入的所有通道进行处理，因此卷积核的实际参数大小为$K \times C_1$），若共有C_2个卷积核（滤波器），则输出大小为$T_2 \times C_2$，T_2的值由卷积核的尺寸、卷积步长、填充等参数所决定。因此，这里标准一维卷积的参数量P_{std}、乘法计算量$O_{std\times}$以及加法计算量O_{std+}分别如式（1）、（2）、（3）所示。

$$P_{std} = (K \times C_1 + 1) \times C_2 \tag{1}$$

$$O_{std\times} = K \times C_1 \times T_2 \times C_2 \tag{2}$$

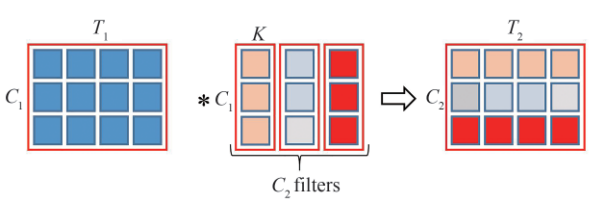


图1 标准一维卷积原理图

Fig. 1 Schematic diagram of standard one-dimensional convolution

$$O_{std+} = ((K \times C_1 - 1) + 1) \times T_2 \times C_2 \tag{3}$$

其中，式（1）、（3）中的+1表示偏置。

深度可分离卷积（Depthwise Separable Convolution，DSC）由两个过程组成，分别为深度卷积和逐点卷积。

① 深度卷积（Depthwise Convolution）

一维深度卷积的原理如图2所示，对于大小同样为$T_1 \times C_1$的输入，卷积核的尺寸为K（这里每个卷积核只对一个特定的通道进行处理，因此每个卷积核的实际参数大小为$K \times 1$），共有C_1个卷积核，即输出通道与输入通道相等。由于深度卷积逐通道处理的特性，因此其也被称为逐通道卷积，这里深度卷积的参数量P_{dw}、乘法计算量$O_{dw\times}$以及加法计算量O_{dw+}分别如式（4）、（5）、（6）所示。

$$P_{dw} = (K \times 1 + 1) \times C_1 \tag{4}$$

$$O_{dw\times} = K \times 1 \times T_2 \times C_1 \tag{5}$$

$$O_{dw+} = ((K \times 1 - 1) + 1) \times T_2 \times C_1 \tag{6}$$

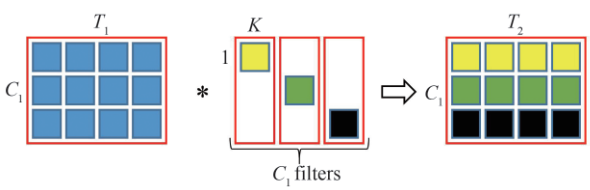


图2 一维深度卷积原理图

Fig. 2 Schematic diagram of one-dimensional depthwise convolution

其中，式（4）、（6）中的+1表示偏置。

②逐点卷积（Pointwise Convolution）

一维逐点卷积的原理如图3所示，它是标准一维卷积的特例（卷积核的尺寸为1），共有C_2个卷积核，由于卷积核的尺寸为1，因此输出的时间维度与输入的时间维度保持不变。逐点卷积主要对输出通道起到升维和降维的作用（通道维度由参数C_2控

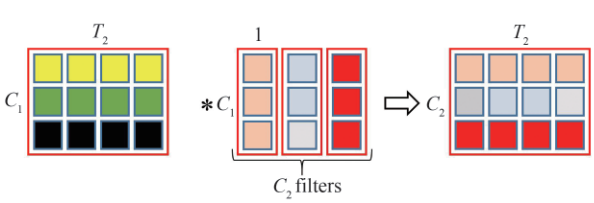


图3 一维逐点卷积的原理图

Fig. 3 Schematic diagram of one-dimensional pointwise convolution

制），并将各通道之间信息进行了整合。因此由于逐点卷积的存在，深度可分离卷积并没有丢失通道间的信息，这里逐点卷积的参数量P_{pw}、乘法计算量$O_{pw\times}$以及加法计算量O_{pw+}如式(7)、(8)、(9)所示。

$$P_{pw} = (1 \times C_1 + 1) \times C_2 \tag{7}$$

$$O_{pw\times} = 1 \times C_1 \times T_2 \times C_2 \tag{8}$$

$$O_{pw+} = ((1 \times C_1 - 1) + 1) \times T_2 \times C_2 \tag{9}$$

其中，式(7)、(9)中的+1表示偏置。

综上所述，一维深度可分离卷积的总参数量为$P_{dw} + P_{pw}$，总乘法计算量为$O_{dw\times} + O_{pw\times}$，总加法计算量为$O_{dw+} + O_{pw+}$。相比标准一维卷积，两者参数量、乘法计算量以及加法计算量的比值分别如式(10)、(11)、(12)所示。

$$\frac{P_{dw} + P_{pw}}{P_{std}} = \frac{(K + 1) \times C_1 + (C_1 + 1) \times C_2}{(K \times C_1 + 1) \times C_2} \approx \frac{1}{C_2} + \frac{1}{K} \tag{10}$$

$$\frac{O_{dw\times} + O_{pw\times}}{O_{std\times}} = \frac{1}{C_2} + \frac{1}{K} \tag{11}$$

$$\frac{O_{dw+} + O_{pw+}}{O_{std+}} = \frac{1}{C_2} + \frac{1}{K} \tag{12}$$

通常情况下C_1、$C_2 \gg 1$且$K \geqslant 3$，因此一维深度可分离卷积相比标准一维卷积而言，参数量、乘法计算量以及加法计算量将会得到明显减少，这将有助于模型的压缩与加速。

从本质上来讲，标准一维卷积将时间相关性和通道相关性进行联合映射，而一维深度可分离卷积将时间相关性和通道相关性进行分开映射。从语音波形中提取Mel谱的过程来看，先对语音分段，再对每段语音分别进行计算提取一帧的Mel谱，因此本文采用一维深度可分离卷积将Mel谱中的时间相关性和通道相关性进行分开学习，这种分而治之的思想更有助于网络利用Mel谱还原出语音波形，也相对减轻网络的学习负担，有助于网络的快速学习。

2.2　生成器

生成器结构如图4所示，它使用Mel谱作为输入，并通过转置卷积不断对其进行上采样，直到输出序列的长度与原始波形的时间分辨率相匹配。每个转置卷积后面都连接一个多接受场融合(Multi-Receptive Field Fusion，MRF)模块，MRF的具体结构如图5所示。

MRF模块累积多个残差块(ResBlock)的输出之和，每个残差块由一系列的一维卷积构成，这些卷积

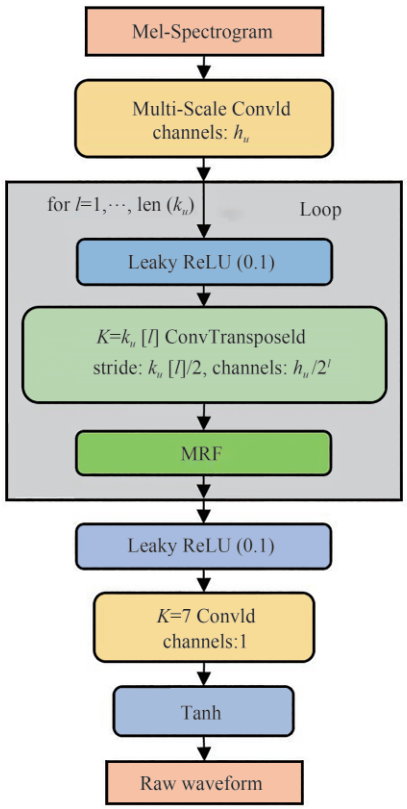


图4　生成器的结构原理图(K表示卷积核的尺寸)

Fig. 4　Schematic diagram of the structure of the generator (K represents the size of the convolution kernel)

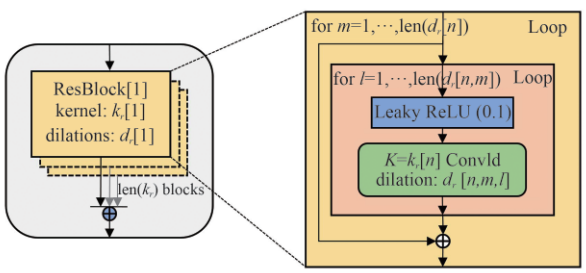


图5　MRF的结构原理图(右边为残差块的具体结构，当表示第i个残差块时，则$n = i$)

Fig. 5　The schematic diagram of the structure of the MRF (The right side is the specific structure of the residual block, when the i-th residual block is represented, then $n = i$)

有着不同大小的卷积核和扩张率，以形成不同大小的接受场，有效建模语音波形的短期和长期相关性。

Mel谱作为语音的一种低分辨率表示，其帧与帧之间存在着强烈的短期和长期相关性，这些相关性对于建模语音分布是至关重要的。因此，本文为了捕获Mel谱的局部特性和远程相关性，采用一个

多尺度卷积(Multi-Scale Convolution,MSC)策略对输入Mel谱进行处理。

如图6所示,MSC策略是指采用多个不同尺寸的卷积核对输入Mel谱进行处理并返回这些处理结果之和,卷积核的具体尺寸如图中所示。这里每个卷积层的输出通道(channels)是一致,即每个卷积层的输出的通道维度是相同的,同时为保证每个卷积层的输出在时间维度上保持不变,将输入(Mel谱)送入卷积层处理之前,根据各卷积层的卷积核大小对输入进行补零填充操作,即在其时间维度上填充一定数量值为0的Mel谱帧。因此,经过不同尺寸的卷积核处理后的输出的纵向和横向维度大小均相同,这将保证各个输出结果可以直接相加。

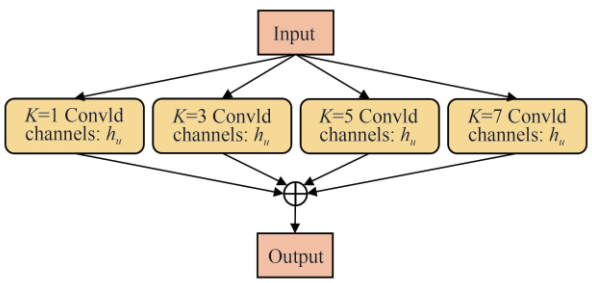


图6 多尺度卷积策略原理图

Fig. 6 Schematic diagram of multi-scale convolution strategy

不同尺寸的卷积核能够捕获不同尺度的全局和局部特征,与原生成器网络采用固定尺寸的卷积核来处理Mel谱相比,本文提出的多尺度卷积策略能充分表征从Mel谱中提取的信息,有助于后续网络利用这些信息进行学习以建模原始语音的分布。此外,为了不显著降低生成器生成语音的质量情况下,有效减少模型参数量并提高模型推理速度,受图像中的深度可分离卷积[19]启发,本文采用一维深度可分离卷积去替代生成器中原有的标准一维卷积。在训练过程中,生成器除了转置卷积层外,其他所有卷积层均采用权重归一化[20],值得注意的是深度可分离卷积采用权重归一化等同于深度卷积和逐点卷积均采用权重归一化。

2.3 判别器

对于生成对抗网络来说,判别器主要对生成器起一个对抗训练的作用,引导生成器能产生更接近真实的数据。这里模型的判别器仍采用HiFi-GAN原有的配置,其包含两个判别器:多尺度判别器和多周期判别器。

2.3.1 多周期判别器

多周期鉴别器(Multi-Period Discriminator,MPD)的结构如图7所示,它由多个网络结构相同的子鉴

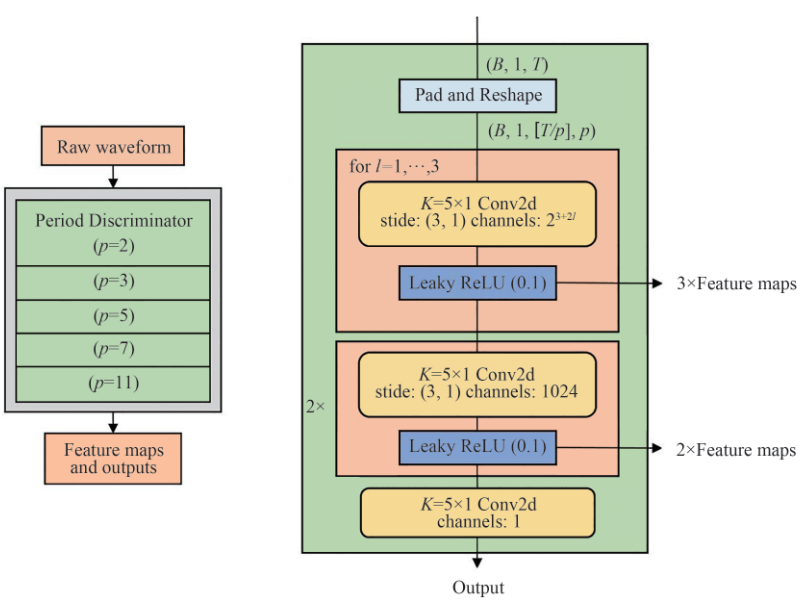


图7 多周期鉴别器原理图(左为整体结构,右为子鉴别器的网络结构,Feature map为每层网络的特征输出,用于下节提到的特征匹配损失)

Fig. 7 Schematic diagram of the multi- period discriminator (the left is the overall structure, the right is the network structure of the sub-discriminator, and "Feature map" is the feature output of each layer of the network, which is used for the feature matching loss mentioned in the next section)

别器组成，每个子鉴别器用于捕获输入语音的一部分周期信号，以识别语音数据中潜在的各种周期模式。

为实现子鉴别器捕获语音信号中的周期模式，每个子鉴别器并不直接处理语音波形，而是将语音波形进行填充与整形，如图 8 所示，以保证每个子鉴别器只接受输入语音波形的等间距采样点，间隔由周期参数 p 控制。通过这样的方式，长度为 T 的一维原始语音被处理成高度为 T/p、宽度为 p 的二维数据，因此，MPD 需要采用二维的卷积神经网络来处理这些数据，除最后一层网络外，其他层均采用二维跨步卷积（只在高度上进行跨步），并且每层卷积均采用权重归一化。在 MPD 的每层卷积层中，卷积核的宽度轴的大小被限制为 1，从而独立处理宽度轴方向的周期语音采样点。因此，每个子鉴别器可以通过处理语音波形不同的部分来捕获语音中彼此不同的潜在周期模式。

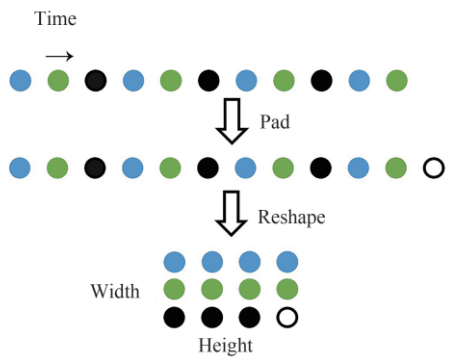


图 8　语音波形的填充与整形示意图（周期参数 p = 3）

Fig. 8　Schematic diagram of pad and reshape of speech waveform (period parameter p = 3)

2.3.2　多尺度判别器

多尺度鉴别器（Multi-Scale Discriminator，MSD）的结构如图 9 所示，MSD 是三个网络结构相同但工作在不同尺度上的鉴别器组合，即分别处理原始语音、经过×2 平均池化的原始语音、经过×4 平均池化的原始语音。子鉴别器通过采用分组卷积来保证使用较大尺寸的卷积核，同时保持较小的参数量。除了对原始语音处理的第一个子鉴别器应用谱归一化[21]，其他两个子鉴别器应用权重归一化，这里采用谱归一化有助于稳定训练。

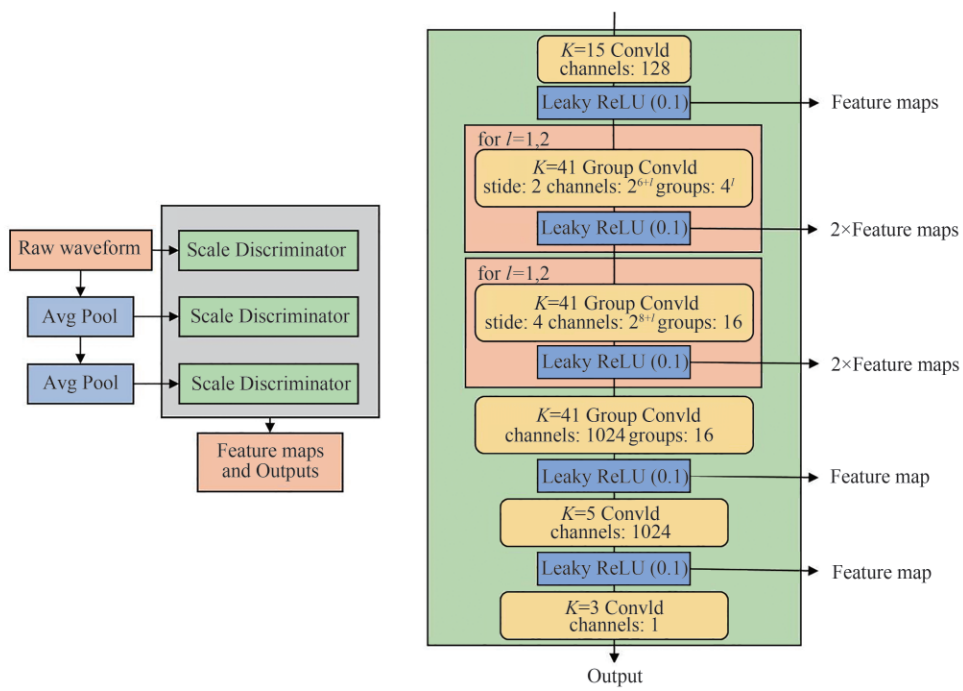


图 9　多尺度鉴别器结构原理图（左为整体结构，右为子鉴别器的网络结构，Feature map 为每层网络的特征输出，用于下节提到的特征匹配损失）

Fig. 9　Schematic diagram of the multi-scale discriminator structure (the left is the overall structure, the right is the network structure of the sub-discriminator, and "Feature map" is the feature output of each layer of the network, which is used for the feature matching loss mentioned in the next section)

2.4 损失函数

损失函数由三部分组成，分别为对抗损失、特征匹配损失以及Mel谱损失。

2.4.1 对抗损失

对于生成器和鉴别器的对抗训练目标，遵循LSGAN[22]的设置。判别器负责对语音样本进行分类，即将真实语音分类为1、生成器生成的语音分类为0，而生成器则根据输入条件生成语音以欺骗判别器，即判别器错将生成语音分类为1。最后通过生成器和判别器的相互博弈过程，直至生成器能够做到以假乱真的效果。因此，生成器和判别器的对抗损失函数分别如式(13)、(14)所示。

$$\mathcal{L}_{\text{Adv}}(G;\ D)=\mathbb{E}_{s}\left[(D(G(s))-1)^2\right] \tag{13}$$

$$\mathcal{L}_{\text{Adv}}(D;\ G)=\mathbb{E}_{(x,s)}\left[(D(x)-1)^2+(D(G(s)))^2\right] \tag{14}$$

为了简洁，这里将MSD和MPD描述为一个鉴别器，其中x代表真实语音，s表示输入条件(对应真实语音提取的Mel谱)。

2.4.2 特征匹配损失

为提高生成器的能力，采用了MelGAN中的特征匹配损失(Feature Matching Loss，FML)，FML通过比较真实语音和生成语音在判别器每层网络的输出特征之间的差异来提高生成器的伪造能力，采用L1距离来衡量这种差异，特征匹配损失函数如式(15)所示。

$$\mathcal{L}_{\text{FM}}(G;\ D)=\mathbb{E}_{(x,s)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(x)-D^i(G(s))\right\|_1\right] \tag{15}$$

其中，T表示鉴别器中的卷积层数，D^i和N_i分别表示鉴别器的第i层中的特征和特征数量。

2.4.3 Mel谱损失

Parallel WaveGAN通过联合优化多分辨率频谱损失和对抗损失，有效捕获了真实语音波形的时频分布。类似于多分辨率频谱损失，HiFi-GAN根据人耳听觉特性，采用Mel谱损失，以期望提高生成语音的感知质量。具体来说，Mel谱损失是生成器生成语音的Mel谱与真实波形的Mel谱之间的L1损失，如式(16)所示。

$$\mathcal{L}_{\text{Mel}}(G)=\mathbb{E}_{(x,s)}\left[\left\|\phi(x)-\phi(G(s))\right\|_1\right] \tag{16}$$

其中，$\phi(\cdot)$表示从语音中提取Mel谱的函数，注意这里提取的Mel谱是全频带的(最低频率为0 Hz，最高频率为语音采样率的一半)，不同于作为输入条件的带限Mel谱，采用全频带的Mel损失有助于模型学习语音的全频带信息。

2.4.4 总损失

特征匹配损失和Mel谱损失作为辅助损失用于稳定模型训练并加速收敛，因此训练生成器和判别器最终的损失函数如式(17)、(18)所示。

$$\mathcal{L}_G=\sum_{k=1}^{K}\left[\mathcal{L}_{\text{Adv}}(G;\ D_k)+\lambda\mathcal{L}_{\text{FM}}(G;\ D_k)\right]+\mu\mathcal{L}_{\text{Mel}}(G) \tag{17}$$

$$\mathcal{L}_D=\sum_{k=1}^{K}\mathcal{L}_{\text{Adv}}(D_k;\ G) \tag{18}$$

其中，D_k表示MPD和MSD中第k个子鉴别器，λ和μ为超参数用于控制各项损失的比重，在实验中其值分别设置为2和45。

3 实验及结果

3.1 实验设置

实验数据采用公开的LJSpeech数据集，该数据集由13100个英语语音片段和相应的文本组成，语音总时长约为24个小时，语音格式为16比特PCM编码，采样率为22050 Hz，由一名专业的美国女性说话者录制。实验中将数据集随机分成两部分：12800个音频样本用于训练集，300个音频样本用于测试集。实验在单个GPU和CPU(NVIDIA Tesla V100 GPU用于训练，Xeon(R) E5-2620 v4 2.10 GHz CPU和NVIDIA GTX 1080Ti GPU用于测试)上进行，模型网络架构基于PyTorch搭建，模型采用80维(0~80 kHz频率范围)的Mel谱作为输入，其中FFT长度、帧长和帧移分别设置为1024、1024和256个采样点，并采用汉宁窗减少频谱能量泄露。

模型采用AdamW[23]优化器进行训练，其中$\beta_1=0.8$、$\beta_2=0.99$、$\varepsilon=1e^{-6}$，初始学习率设置为$2e^{-4}$，每经过一个训练周期(Epoch)，学习率衰减为0.999倍。批处理每条语音的长度和批处理大小分别设置为16384个采样点和12个音频样本。此外，生成器网络的具体参数如表1所示，部分参数的表示方式类似数组形式，如$k_r[2]=7$，$d_r[3,2,1]=3$。为了进行对比实验，本文也训练了WaveNet、WaveGlow声码器，模型及训练设置保持与原有设置相同。

3.2 Mel谱损失对比

Mel谱是根据人耳的听觉特性设计的，生成语音的Mel谱与真实语音的Mel谱之间的差距能够一

表 1　生成器的网络参数配置

Tab. 1　Network parameters configuration of the generator

网络参数	参数值
h_u	512
k_u	[16, 16, 4, 4]
k_r	[3, 7, 11]
d_r	[[[1, 1], [3, 1], [5, 1]], [[1, 1], [3, 1], [5, 1]], [[1, 1], [3, 1], [5, 1]]]

定程度上反映生成语音与真实语音在语音质量上的差异。为比较多尺度卷积(MSC)策略和一维深度可分离卷积(DSC)在 HiFi-GAN 模型中的有效性，本文首先展示了它们结合 HiFi-GAN 模型在不同训练周期时，测试集上的 Mel 谱损失情况，如图 10 所示。

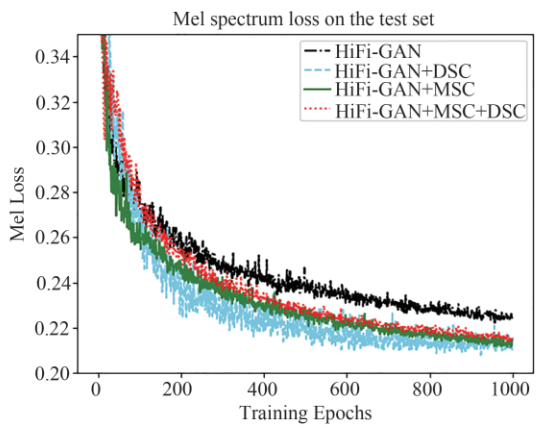


图 10　不同训练阶段的测试集上的 Mel 谱损失

Fig. 10　Mel spectral loss on the test set at different training stages

根据图中结果可以看出，MSC 策略的确有助于生成器学习，产生更接近真实语音的生成语音。采用 DSC 能使模型的 Mel 谱损失快速下降，但是在 800 Epochs 左右，其 Mel 损失就开始趋于稳定，而其他三个模型(包括基线 HiFi-GAN)的损失仍有下降的趋势。不过 MSC 解决了这个问题，提高了模型进一步的学习能力，如在 1000 Epochs 的时候，HiFi-GAN+MSC+DSC 模型不仅在损失上与 HiFi-GAN+DSC 模型基本持平，而且其损失仍有下降的趋势，模型还有继续学习的能力。

综上所述，DSC 具有加速模型学习的能力，但这种能力是不稳定的，即其损失在图中震荡比较大(如在 950 Epochs 左右，Mel 谱损失甚至跳回基线水平)；而 MSC 策略能够有效提高模型学习能力并且稳定训练过程，如采用 MSC 的两个模型在损失上更稳定(毛刺较少，浮动幅度较小)，尤其是与 DSC 结合的情况下更明显。

3.3　语音质量、模型参数及推理速度

为比较不同模型间的语音质量(模型均训练到其损失不再明显下降)，本文采用了一个客观指标：客观语音质量评估[24](Perceptual Evaluation of Speech Quality，PESQ)和一个主观评价：平均主观意见分(Mean Opinion Score，MOS)。生成器生成的语音是以真实语音的 Mel 谱作为输入条件，因此生成语音有着对应真实语音作为参考，即可以计算两者的 PESQ 值。这里需要注意的是进行 PESQ 计算前，需要保证生成语音和原始语音的长度一致，由于生成语音的时间分辨率是输入 Mel 谱的时间分辨率的 256(帧移)倍，因此测试集每条语音的长度需要处理成帧移的整数倍才能计算两者的 PESQ 值，本文是通过丢弃每条语音尾部多余的采样点(即最多丢弃 255 个采样点)来实现的，因为一条语音有几万甚至十几万个采样点，其尾部二百多个采样点基本是语音中的静音段，截去并不会影响语音质量和内容。随后则利用处理后的测试集来提取真实 Mel 谱输入模型中得到生成语音，生成语音与处理后的测试集中的对应真实语音均被下采样至 16 kHz，并采用 python 里面的 pypesq 库进行 PESQ 值的计算；而 MOS 测试是从测试集中随机选取 30 条语音作为评估集，由 15 位精通英语的听众通过耳机试听并根据语音的质量，由差(1)到好(5)采用 5 分制进行打分，MOS 得分的置信区间(Confidence Intervals，CI)为 95%。

此外，本文也比较了各模型间的参数量与分别在 GPU、CPU 上的推理速度，这里的推理速度采用实时因子(Real Time Factor，RTF)的倒数来衡量，RTF 表示模型生成一秒时长的语音波形所需的时间(单位秒)，即生成一秒波形需一秒的时长则认为模型刚好达到了实时性。因此，RTF 的倒数值表示该模型的推理速度是实时的倍数值。

不同模型间的参数量、推理速度、MOS 和 PESQ 得分如表 2 所示。根据表中结果可以看出，DSC 的引入显著减少了 HiFi-GAN 模型的参数量，提升了模型在 GPU、CPU 上的推理速度，并且未对 HiFi-GAN 生成语音的质量造成明显影响。MSC 则仅仅在 HiFi-GAN 的基础上增加了 2.65% 的参数量，却

表2 不同模型的参数、推理速度、MOS和PESQ得分

Tab. 2 Parameters、inference speed、MOS and PESQ scores of different models

模型	MOS(CI)	PESQ	参数(百万)	推理速度(GPU)	推理速度(CPU)
真实语音	4.58±0.08	4.50	—	—	—
WaveNet	4.05±0.07	3.28	24.75	× 0.002	—
WaveGlow	3.99±0.08	3.21	87.88	× 5.10	× 0.13
HiFi-GAN	4.32±0.08	3.70	13.94	× 70.55	× 2.45
HiFi-GAN + DSC	4.30±0.10	3.69	4.37	× 80.72	× 3.18
HiFi-GAN + MSC	4.41±0.07	3.78	14.31	× 66.63	× 2.39
HiFi-GAN + DSC + MSC	4.39±0.05	3.75	4.50	× 78.82	× 3.16

有效提升了HiFi-GAN生成语音的质量。本文通过将DSC和MSC结合,使得改进后的HiFi-GAN模型不仅参数量约减少了67.72%,语音质量也略有提升(MOS提升了0.07,PESQ提升了0.05),此外在GPU、CPU上推理速度约分别提升了11.72%和28.98%。与基于流(WaveGlow)和基于自回归(WaveNet)的模型相比,改进后的HiFi-GAN在各项指标上均优于它们。

3.4 多说话人条件下的性能测试

为验证改进后的HiFi-GAN模型在不可见说话人条件下的通用性,本文另外采用了公开的VCTK数据集,该数据集包括由109名说话人录制约44200个语音片段,总时长约为44个小时,语音格式为16比特PCM编码,采样率为44 kHz,这里为了训练方便(保持LJSpeech-1.1数据集下的训练设置),将语音采样率下降为22 kHz。选取100名说话者的语音片段作为训练集,剩下9名说话者的语音片段作为测试集。为了对比实验,本文也利用该数据集训练了WaveNet、WaveGlow模型,模型均训练到损失不再明显下降。PESQ评价方式如上节相同,即计算测试集中真实语音与利用其真实Mel谱生成语音之间的PESQ得分。此外,为了进行MOS测试,随机从测试集中9个说话人的语音片段中共选取45个片段作为评估集,即每个说话人选取5个片段,评价方式和打分规则与4.2节相同。注意这里测试集里面的9个说话人是在模型训练过程未出现过的,对于模型来说是不可见的。

在不可见说话人条件下,各模型生成语音的PESQ和MOS得分如表3所示,可以看出改进后的HiFi-GAN声码器的MOS和PESQ得分均相比于原模型有所提高,与在单说话人数据集上的得分结果相似,这表明改进后的模型对于不可见说话人上仍具有不错的通用性。

表3 不可见说话人条件下的MOS和PESQ得分

Tab. 3 MOS and PESQ scores in unseen speaker condition

模型	MOS(CI)	PESQ
真实语音	4.16±0.07	4.50
WaveNet	3.83±0.08	3.22
WaveGlow	3.81±0.09	3.19
HiFi-GAN	4.07±0.07	3.67
HiFi-GAN + DSC	4.04±0.10	3.64
HiFi-GAN + MSC	4.10±0.08	3.71
HiFi-GAN + DSC + MSC	4.09±0.08	3.69

3.5 端到端语音合成

为了验证改进的HiFi-GAN模型作为端到端语音合成任务上的声码器的有效性,采用相同数据集(LJSpeech-1.1)训练了基于Tacotron 2[25]的声学特征预测模型作为前端,注意用于训练声学特征预测模型的数据集切分也与训练声码器网络的数据集切分保持一致,以避免在评估期间泄露测试数据而夸大实验结果。基于Tacotron 2的声学特征预测模型采用文本序列作为输出,输出预测的Mel谱,然后再将其输入声码器中以生成语音波形。采用MOS评估生成语音的质量,方式与4.2节中保持相同。

不同声码器在端到端语音合成条件下的MOS得分如表4所示,改进后的HiFi-GAN模型在端到端语音合成任务中取得了最好的结果,表明改进后的HiFi-GAN模型在端到端语音合成任务上有很好的兼容性。

表4　端到端语音合成条件下的MOS得分

Tab. 4　MOS scores under end-to-end speech synthesis conditions

模型	MOS
真实语音	4.58±0.08
Tacotron 2 + WaveNet	3.83±0.09
Tacotron 2 + WaveGlow	3.85±0.07
Tacotron 2 + HiFi-GAN	4.12±0.06
Tacotron 2 + 改进的HiFi-GAN	4.15±0.06

4　结论

本文在HiFi-GAN声码器的基础上，提出一种参数更少、推理速度更快、语音质量更高的改进型HiFi-GAN声码器。具体来说，首先通过多尺度卷积策略来有效表征输入Mel谱特征，为后续网络提供更充足的信息以生成更高质量的语音，其次利用深度可分离卷积有效地减少了模型参数，提升了推理速度。根据实验结果表明，本文的方法在不降低生成语音质量的前提下（示例语音可以在网址：https://pan. baidu. com/s/1rZ4fAuLs65Qz-NlLbp4AWA?pwd=naiv上进行试听），有效减少了HiFi-GAN模型约67. 72%的参数，并提升了模型推理速度（在GPU上提升了11. 72%，在CPU上提升了28. 98%），这对于将模型部署到硬件条件不足、内存受限的应用场景中是十分有意义的，并且改进后的HiFi-GAN模型在端到端语音合成任务上表现优异，有很好的兼容性。

参考文献

[1] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNÉ A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3/4): 187-207.

[2] MORISE M, YOKOMORI F, OZAWA K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE Transactions on Information and Systems, 2016, E99.D(7): 1877-1884.

[3] OORD A V D, DIELEMAN S, ZEN HEIGA, et al. WaveNet: A generative model for raw audio[EB/OL]. 2016: arXiv: 1609.03499[cs.SD]. https://arxiv.org/abs/1609.03499.

[4] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[EB/OL]. 2018: arXiv: 1802.08435[cs.SD]. https://arxiv.org/abs/1802.08435.

[5] MEHRI S, KUMAR K, GULRAJANI I, et al. SampleRNN: an unconditional end-to-end neural audio generation model[EB/OL]. 2016: arXiv: 1612.07837[cs.SD]. https://arxiv.org/abs/1612.07837.

[6] OORD A V D, LI Yazhe, BABUSCHKIN I, et al. Parallel WaveNet: Fast high-fidelity speech synthesis[EB/OL]. 2017: arXiv: 1711.10433[cs. LG]. https://arxiv.org/abs/1711.10433.

[7] PING Wei, PENG Kainan, CHEN Jitong. ClariNet: parallel wave generation in end-to-end text-to-speech[EB/OL]. 2018: arXiv: 1807.07281[cs. CL]. https://arxiv.org/abs/1807.07281.

[8] KINGMA D P, SALIMANS T, JOZEFOWICZ R, et al. Improving variational inference with inverse autoregressive flow[EB/OL]. 2016: arXiv: 1606.04934[cs. LG]. https://arxiv.org/abs/1606.04934.

[9] PRENGER R, VALLE R, CATANZARO B. WaveGlow: A flow-based generative network for speech synthesis[C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK. IEEE, 2019: 3617-3621.

[10] PING W, PENG K, ZHAO K, et al. WaveFlow: A compact flow-based model for raw audio[C]// Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 7706-7716.

[11] KONG Zhifeng, PING Wei, HUANG Jiaji, et al. Diffwave: A versatile diffusion model for audio synthesis[EB/OL]. 2020: arXiv: 2009.09761[eess.AS]. https://arxiv.org/abs/2009.09761.

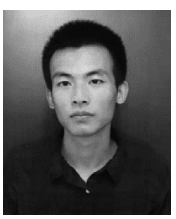
[12] CHEN Nanxin, ZHANG Yu, ZEN Heiga, et al. WaveGrad: Estimating gradients for waveform generation[EB/OL]. 2020: arXiv: 2009.00713[eess.AS]. https://arxiv.org/abs/2009.00713.

[13] YAMAMOTO R, SONG E, KIM J M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain. IEEE, 2020: 6199-6203.

[14] KUMAR K, KUMAR R, BOISSIERE T D, et al. MelGAN: Generative adversarial networks for conditional waveform synthesis[EB/OL]. 2019: arXiv: 1910.06711[eess.AS]. https://arxiv.org/abs/1910.06711.

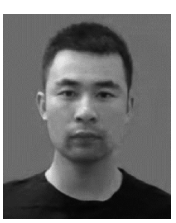
[15] YANG Geng, YANG Shan, LIU Kai, et al. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech [C]//2021 IEEE Spoken Language Technology Workshop. Shenzhen, China. IEEE, 2021: 492-498.
[16] YANG J, LEE Junmo, KIM Y, et al. VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network [EB/OL]. 2020: arXiv: 2007.15256 [eess.AS]. https://arxiv.org/abs/2007.15256.
[17] ZENG Zhen, WANG Jianzong, CHENG Ning, et al. LVC-Net: Efficient condition-dependent modeling network for waveform generation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, ON, Canada. IEEE, 2021: 6054-6058.
[18] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[EB/OL]. 2020: arXiv: 2010.05646[cs.SD]. https://arxiv.org/abs/2010.05646.
[19] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. 2017: arXiv: 1704.04861 [cs.CV]. https://arxiv.org/abs/1704.04861.
[20] SALIMANS T, KINGMA D P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks[EB/OL]. 2016: arXiv: 1602.07868[cs.LG]. https://arxiv.org/abs/1602.07868.
[21] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks [EB/OL]. 2018: arXiv: 1802.05957 [cs.LG]. https://arxiv.org/abs/1802.05957.
[22] MAO Xudong, LI Qing, XIE Haoran, et al. Least squares generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy. IEEE, 2017: 2813-2821.
[23] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization [EB/OL]. 2017: arXiv: 1711.05101 [cs.LG]. https://arxiv.org/abs/1711.05101.
[24] 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]//ICASSP 2001-2001 IEEE International Conference on Acoustics, Speech and Signal processing. Proceedings (Cat. No. 01CH37221). Salt Lake City, UT, USA. IEEE, 2001, 2: 749-752.
[25] SHEN J, PANG Ruoming, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions[C]//ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. IEEE, 2018: 4779-4783.

作者简介



唐 君 男,1996年生,江西九江人。中国人民解放军战略支援部队信息工程大学硕士研究生,主要研究方向为智能信息处理、人工智能、语音合成。
E-mail: 2433548528@qq.com

张连海 男,1971年生,山东单县人。中国人民解放军战略支援部队信息工程大学教授,主要研究方向为语音信号处理、智能信息处理、人工智能、信号分析等。
E-mail: llhhzz163@163.com



李嘉欣 男,1998年生,湖南湘乡人。中国人民解放军战略支援部队信息工程大学硕士研究生,主要研究方向为智能信息处理、人工智能、语音合成。
E-mail: 414171817@qq.com

李宜亭 男,1993年生,甘肃兰州人。中国人民解放军战略支援部队信息工程大学硕士研究生,主要研究方向为智能信息处理、语音识别、模型压缩。
E-mail: 904890536@qq.com