计算机科学
COMPUTER SCIENCE

融合门控循环单元及自注意力机制的生成对抗语音增强

张德辉, 董安明, 禹继国, 赵恺, 周酉

引用本文

张德辉, 董安明, 禹继国, 赵恺, 周酉. 融合门控循环单元及自注意力机制的生成对抗语音增强[J]. 计算机科学, 2023, 50(11A): 230200203-9.

ZHANG Dehui, DONG Anming, YU Jiguo, ZHAO Kai and ZHOU You. Speech Enhancement Based on Generative Adversarial Networks with Gated Recurrent Units and Self-attention Mechanisms [J]. Computer Science, 2023, 50(11A): 230200203-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200203

融合门控循环单元及自注意力机制的生成对抗语音增强

张德辉[1] 董安明[1,2] 禹继国[1,2] 赵 恺[3] 周 酉[4]
1 齐鲁工业大学(山东省科学院)计算机科学与技术学院 济南 250353
2 齐鲁工业大学(山东省科学院)大数据研究院 济南 250353
3 中国科学院自动化研究所 北京 100190
4 山东海看新媒体研究院有限公司 济南 250013
(anmigndong@qlu.edu.cn)

摘 要 因其通过两种网络对抗训练并不断提升网络映射能力的特性,生成对抗网络(Generative Adversarial Networks,GAN)具有强大的降噪能力,近年来被应用于语音增强领域。针对现有生成对抗网络语音增强方法未充分利用语音特征序列中的时间相关性和全局相关性这一不足,提出一种融合门控循环单元(Gated Recurrent Unit,GRU)和自注意力机制(self-attention)的语音增强GAN网络。该网络利用串联和并联两种方式构建了时间建模模块,可捕获语音特征序列的时间相关性和上下文信息。与基线算法相比,所设计的新型GAN网络语音质量听觉估计分数(PESQ)提高了4%,且在语音信号分段信噪比(SSNR)和短时客观可懂度(STOI)等多个客观评价指标上表现更优。该研究结果表明,融合语音特征序列中的时间相关性和全局相关性有助于提升GAN网络语音增强的性能。

关键词:语音增强;生成对抗网络;门控循环单元;自注意力机制;特征融合
中图法分类号 TP391

Speech Enhancement Based on Generative Adversarial Networks with Gated Recurrent Units and Self-attention Mechanisms

ZHANG Dehui[1],DONG Anming[1,2],YU Jiguo[1,2],ZHAO Kai[3] and ZHOU You[4]
1 School of Computer Science and Technology,Qilu University of Technology(Shandong Academy of Sciences),Jinan 250353,China
2 Big Data Research Institute,Qilu University of Technology(Shandong Academy of Sciences),Jinan 250353,China
3 Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China
4 Shandong HiCon New Media Institute Co.LTD,Jinan 250013,China

Abstract Generative adversarial networks(GAN) have strong noise reduction ability and have been applied in the field of speech enhancement in recent years due to their ability to use two kinds of network adversarial training and constantly improve the network mapping ability. In view of the shortcomings of existing generative adversarial network speech enhancement methods,which do not make full use of temporal and global dependencies in speech feature sequences,this paper proposes a speech enhancement GAN network that integrates gated recurrent units and self-attention mechanism. The network constructs a time modeling module in series and parallel to capture the time dependence and context information of speech feature sequences. Compared with the baseline algorithm,the proposed new GAN network speech quality auditory estimation score(PESQ) improves by 4%,and performs better on several objective evaluation indexes such as segmental signal-to-noise ratio(SSNR) and short-term objective intelligibility(STOI). The results show that the integration of temporal correlation and global correlation in speech feature sequences is helpful to improve the performance of GAN network speech enhancement.

Keywords Speech enhancement,Generative adversarial network,Gated recurrent unit,Self-attention mechanism,Feature fusion

---

基金项目:国家重点研发计划(2019YFB2102600);国家自然科学基金(62272256);山东省科技型中小企业创新能力提升工程项目(2022TSGC2180,2022TSGC2123);济南市"高校20条"自主培养创新团队(202228093);齐鲁工业大学(山东省科学院)科教产融合试点工程项目(基础研究类)先导项目(2022XD001)
This work was supported by the National Key Research and Development Program of China(2019YFB2102600),National Natural Science Foundation of China(62272256),Innovation Capability Enhancement Program for Small and Medium-sized Technological Enterprises of Shandong Province(2022TSGC2180,2022TSGC2123),Piloting Fundamental Research Program for the Integration of Scientific Research,Independent Training Innovation Team of Jinan(202228093),and Piloting Fundamental Research Program for the Integration of Scientific Research,Education and Industry of Qilu University of Technology(Shandong Academy of Sciences)(2022XD001).
通信作者:董安明(anmingdong@qlu.edu.cn)

1 引言

随着近年来人工智能技术在自动翻译、移动通信[1]、自动驾驶、智能机器人、智慧医疗、智能教育等场景中的大规模应用，通过语音进行人机交互[2]变得越来越广泛。然而，各种智能语音设备所采集的语音信号不可避免地受到环境噪声的影响。语音增强[3]是指从受到噪声污染的语音信号中恢复出纯净语音的信号处理技术，其目的是抑制语音信号的背景噪声，改善语音信号的主观感知质量和可懂度[4]。语音增强成为改善未来智能人机交互系统语音识别可靠性和精准性的重要保障。

传统的语音增强方法一般通过假设目标语音与噪声服从某种分布来对带噪语音进行增强，如谱减法[5]、维纳滤波算法[6]、卡尔曼滤波算法[7]和信号子空间算法[8]等。这些基于统计模型的语音增强方法需要依赖于噪声的统计分布，适用于平稳噪声环境。当环境噪声非平稳时，其降噪能力会大幅下降[9]。

近年来，随着人工智能技术的发展，各种基于深度学习的语音增强算法相继被提出，如深度神经网络[10](Deep Neural Network，DNN)、卷积神经网络[11-12](Convolutional Neural Network，CNN)、循环神经网络[13-15](Recurrent Neural Network，RNN)等。与传统基于统计模型的语音增强方法相比，基于深度学习的语音增强方法具有抑制非平稳噪声的潜力，在复杂的声学环境下也能够有效地提取纯净语音[16-18]。

上述基于深度学习的语音增强方法均将带噪语音信号送入一个多层前馈或者递归神经网络，并通过训练数据利用误差反向传播机制进行权值的更新，即模型的训练。与这些经典深度学习框架不同，近年来出现了一种被称为生成对抗网络(Generative Adversarial Net，GAN)的新型深度学习框架[19]。GAN 网络由两个神经网络模型组成，分别被称为生成模型(Generative Model)和判别模型(Discriminative Model)。其中，生成模型用于捕获输入数据的统计分布；判别模型用于估计出一个样本来自于训练数据(而非生成模型输出)的概率。GAN 通过两个网络的对抗来训练网络，即判别器尽力提升自己分辨真实数据和生成数据的能力；而生成器以最大化判别器判决判错概率为目标，尽力提高自己所生成数据的伪装能力，最终让判别器难辨真假。由于能够用来生成逼真的图像并具有高维分布推广能力，GAN 网络成为计算机视觉领域一个里程碑式的发展，为解决各种图像预测问题提供了新模式。

近期，GAN 网络也开始被应用于语音增强领域。文献[20]率先将 GAN 应用于语音增强，提出了一种称为 SEGAN(Speech Enhancement Generative Adversarial Network)的语音增强网络。SEGAN 利用生成对抗的训练方式对语音在时域进行端到端的增强映射，从而实现有效的降噪。随后，各种新型的基于 GAN 的语音增强方法被相继提出[21-24]。这些基于 GAN 的语音增强方法根据不同的处理特征方式，一般可以分为时域和时频域两种模式。时域语音增强模式通过训练神经网络直接寻找带噪语音与纯净语音的映射关系，端到端地进行语音增强。例如，文献[21]中提出了一种迭代生成器的 GAN 网络，通过多生成器迭代学习语音时域特征增强语音；文献[22]将自注意力机制(self-attention)融入到 SE-GAN 网络，通过关注语音时域的上下文信息提升增强语音的质量。而时频域语音增强模式则是通过训练网络预测语音的纯净幅值谱，利用傅里叶逆变换将得到的语音幅值谱恢复为增强语音。例如，文献[23]利用全卷积的 GAN 网络预测语音的幅值谱进行语音增强；文献[24]将 LSTM 融入 GAN 网络预测语音的时频掩码，然后将掩码与噪声幅值谱相乘，得到增强语音的幅值谱。在这些工作中，GAN 网络都能够生成接近于纯净语音的增强语音。

基于 GAN 的语音增强研究目前依然处于初步发展阶段，有很多问题需要进一步探索。例如，上述基于 GAN 的端到端语音增强方法中，SEGAN[20]，SASEGAN[22]，ISEGAN[21]和 DSEGAN[21]等算法，生成网络和鉴别网络都采用卷积神经网络结构处理语音特征。这些卷积神经网络层与层之间都利用有限尺度的卷积核进行映射，每一个特征点都只蕴含上层特征图相应卷积核大小的特征信息，感受野受限，且无法与特征图中其他特征产生关联，即便随着卷积层数加深，感受野加大，如果不将感受野扩大到整个特征图，也仍然无法将整个特征图关联在一起。这让卷积神经网络难以关注序列特征的时间序列相关性以及全局相关性，因此不能充分利用语音特征的多样性进行时间依赖性建模。针对语音特征的时间序列相关性，门控循环单元(GRU)可以利用独特的处理机制提取特征序列相关性。GRU 中含有和特征序列时间长度相同的单元数，特征图按照时间序列输入到 GRU 层，每个 GRU 单元的门控机制将特征图中的时间关联特征保留并输入给下一个单元。GRU 的独特处理机制能够很好地从时间序列的角度捕捉语音特征的序列相关性。针对语音特征的全局特征，自注意力机制可以进行全局特征的提取。Self-attention 将特征图转化为 3 个矩阵，即查询矩阵(Query，Q)、键矩阵(Key，K)、值矩阵(Value，V)，利用查询矩阵和键矩阵进行相似度计算得到独有的注意力矩阵，注意力矩阵与值矩阵相乘可以很好地提取特征图的全局特征。虽然也有相关工作采用自注意力机制提取语音的全局特征[22]，但是语音的时域表示作为一种时间序列信息，不仅需要关注全局特征，时间序列之间的相关性同样不容忽视，所以单一地从时间顺序或特征全局进行时间依赖性建模并不能充分地考虑到语音信息的各种特征。

针对上述问题，本文面向语音增强提出一种融合门控循环单元和自注意力机制的新型 GAN 网络结构。与 SASEGAN 只用 self-attention 提取语音全局特征不同，所提 GAN 结构将 GRU 和 self-attention 联合，构建一个时间建模结构，进行语音特征的时间依赖性建模。具体地，将 GRU 和 self-attention 采用串联和并联两种联合方式融入到全卷积的生成对抗网络结构中。上述描述中，串联方式将 GRU 和 self-attention 上下链接，语音特征经过 GRU 处理之后进入 self-attention 层；并联方式中语音特征同时进入 GRU 和 self-attention 并行处理，之后的输出融合输入到下一个网络层。因而，所提 GAN 网络融合了 GRU 对时间序列特征的提取能力和自注意力机制的全局特征提取能力，这样可以充分学习语音时序信息的特征，从不同方面联合对语音信息的时间依赖性进行建模。实验结果证明，通过关注语音信息的多种特征，显著提高了语音增强 GAN 的性能。

2 生成对抗网络语音增强

GAN 网络框架由两个神经网络模型组成，一个被称为生成器（简称 G），另一个被称为判别器（简称 D）。GAN 网络利用生成对抗的思想进行训练，即判别器最大化自己鉴别伪造数据的能力，而生成器最大化自己所生成数据的欺骗性，其最终的目标是使得判决器无法将生成模型所生成的数据与真实数据区分开来。

基于 GAN 网络的语音增强模型的基本框架如图 1 所示。其中，生成器包括编码器和解码器两部分，编码器由多层卷积层组成，用于提取特征；解码器由多层反卷积层组成，用于恢复语音。鉴别器除了多层卷积层用于提取特征之外，最后还需要一个全连接层和一个 softmax 层分类真假。

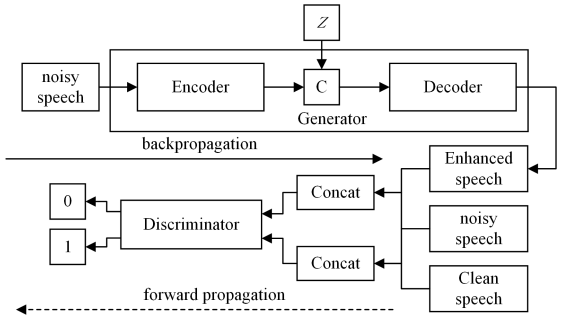


图 1 基于生成对抗网络的语音增强模型框架图

Fig. 1 Framework diagram of speech enhancement model based on generative adversarial network

基于 GAN 网络的语音增强模型采用生成对抗的机制进行训练。通过生成器与鉴别器相互对抗的方式，学习带噪语音与纯净语音之间的映射关系。

在训练过程中，将带噪语音信号输入到生成器。生成器中的编码器利用多层卷积处理从带噪语音中提取语音的特征，并编码为一个中间特征矩阵 C。该特征矩阵再进一步与一个同维度的随机噪声向量 Z 拼接起来形成新的拼接矩阵，进而送入解码器，以生成降噪后的增强语音信号。

鉴别器通过多层卷积神经网络提取输入信号的特征，以鉴别所输入信号是否为生成器生成的信号。鉴别器输入的信号有两种类型。其中，第一种为纯净语音信号和带噪语音信号的拼接；第二种则是生成器输出的增强语音信号和带噪语音信号的拼接。对于第一种拼接信号，其训练标签为 1，表示所输入的信号是真实语音拼接而成；对于第二种拼接信号，其训练标签为 0，表示其为生成语音拼接而成。

3 融合门控循环单元与自注意力机制的生成对抗网络

3.1 网络结构

本文所提出的生成对抗网络语音增强算法依然遵循图 1 所示的整体结构，其中具体的生成网络和鉴别网络结构分别如图 2 和图 3 所示。

图 2 所示生成网络的编码器由 11 个卷积层和 1 个时间建模模型构成，解码器由 11 个反卷积层和 1 个时间建模模型构成。本文方法区别于文献中 SASEGAN 方案的地方，就在于时间建模模型的引入。

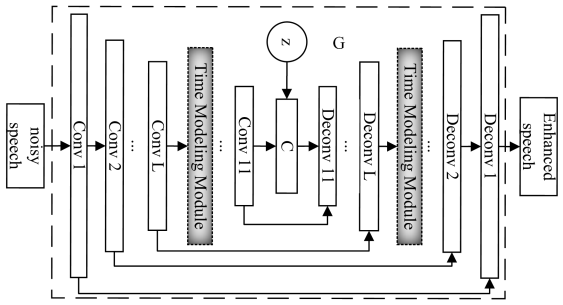


图 2 生成器结构图

Fig. 2 Generator rstructure diagram

生成网络输入长度为 len 个采样点的带噪语音信号。本文设置 $len=16384$，对应 16 kHz 采样率下时间约为 1 s 的信号。编码器的每一个卷积层中的卷积核尺寸都为 31，步幅为 2。11 层的卷积核数量分别为{16，32，32，64，64，128，128，256，256，512，1024}，得到大小分别为 8192×16，4096×32，2048×32，1024×64，512×64，256×128，128×128，64×256，32×256，16×512，8×1024 的特征映射。编码器输出的特征映射图 $C\in\mathcal{R}^{8\times1024}$ 与 噪声样本 $Z\in\mathcal{R}^{8\times1024}$ 堆叠后，形成解码器的输入。

解码器结构与编码器结构镜像相反，具有相同数量的滤波器和网络配置参数，通过反卷积来反转编码过程。G 网络的每一层卷积层之后，都有一个参数整流线性单元（PreLU）[25] 作为激活函数。为了让信息从编码阶段流进解码阶段，使用跳跃连接[26] 将编码器中的每个卷积层与解码器中的对应反卷积层连接起来。

鉴别器作为一个二分类器来判别语音信息的真假，结构类似于生成器的编码器部分。不同的是，鉴别器接收一对语音信息片段作为输入，每一层卷积层之后是虚拟批规范层和 LeakyReLU 激活层，激活层参数 $\alpha=0.3$。11 层一维卷积层提取语音特征之后，通过增加一个 1×1 卷积层对语音特征谱进一步处理，将 8×1024 的语音特征谱减少到 8 个特征，最后经过全连接层映射为 1 个特征并送入 softmax 层进行分类。鉴别器结构如图 3 所示。

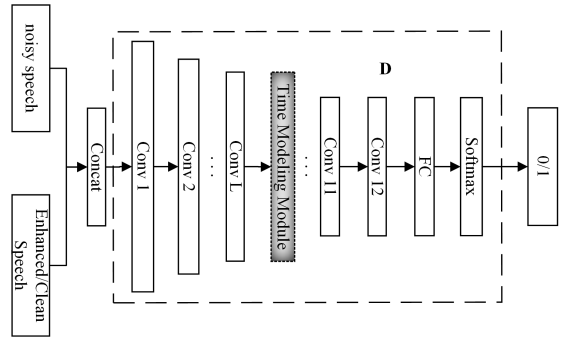


图 3 鉴别器结构图

Fig. 3 Discriminator structure diagram

需要说明的是，图 2 和图 3 所展示的是时间建模模块与第 L 个卷积层耦合的生成器和鉴别器示例。一般来说，如果内存容量足够的话，时间建模模块可以与任意数量甚至所有卷积层组合使用。

3.2 时间建模模块

语音信号在时间上是相关的，而且语音具有短时平稳性，所以相邻语音帧的音素之间存在明显联系。另外，很多语音

音节之间也存在相似性，所以远程的语音帧之间同样存在联系。由于语音相邻帧之间和远程帧之间都具有相关性，所以语音信号的特征关联不仅具有局部相关性，还有全局相关性。

为了更好地捕获语音的时间序列信息和全局上下文信息等特征，本文利用门控循环单元和自注意力机制组成时间建模模块来体现这种特征的时间相关性。其中，GRU 具有顺序捕获时间序列局部关联性的特性，但是缺乏对全局上下文信息的关注能力。具有自注意力机制的神经网络能够全局提取输入信息的特征关联，快速捕获特征的上下文关联，但是对时间序列之间的特征关联性学习不够充分。

根据 GRU 和自注意力模块的不同组合形式，本文设计了串联和并联两种不同结构的时间建模模块(Time Modeling Module)，分别如图 4 和图 5 所示。

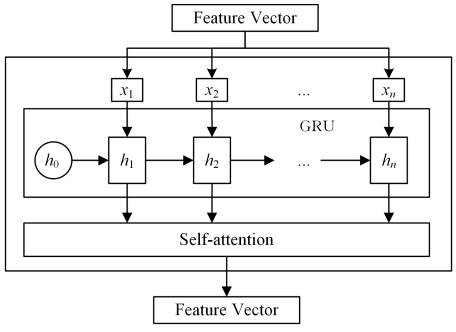


图 4 时间建模模块 1 结构图

Fig. 4 Structure diagram of time modeling module 1

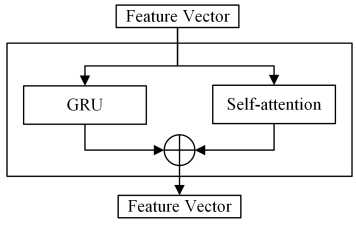


图 5 时间建模模块 2 结构图

Fig. 5 Structure diagram of time modeling module 2

本文将融合了 GRU 与 self-attention 的时间建模模块的语音增强生成对抗网络命名为 GSA-SEGAN。根据所采用的如图 4 和图 5 所示的不同时间建模模块，又细分为 GSA-SEGAN-1 和 GSA-SEGAN-2 两种方案。

3.3 门控循环单元

门控循环单元是经典的长短时记忆神经网络(LSTM)的演进版[27]，能够克服传统的 RNN 在长时序列建模时容易发生梯度消失和梯度爆炸的问题。与 LSTM 相比较，GRU 将 LSTM 中的遗忘门和输入门融合输入一个单独的更新门，并合并了输出状态，因而参数量少，易于训练，收敛速度快且过拟合风险低。

GRU 的结构如图 6 所示，由更新门 z_t 和重置门 r_t 组成。更新门决定了有多少前一时间步的信息和当前时间步的信息要被继续传递到未来，重置门控制要遗忘多少过去的信息。

GRU 中各个网络节点的输出分别表示为：

$$z_t=\sigma(W_z x_t+U_z h_{t-1}+b_z) \tag{1}$$

$$r_t=\sigma(W_r x_t+U_r h_{t-1}+b_r) \tag{2}$$

$$\tilde{h}_t=\tanh(W_h x_t+U_h(r_t\otimes h_{t-1})+b_h) \tag{3}$$

$$h_t=(1-z_t)\otimes h_{t-1}+z_t\otimes\tilde{h}_t \tag{4}$$

其中，x_t 和 h_t 分别表示当前时刻的输入和输出向量，h_{t-1} 为 $t-1$ 时刻的输出向量，$\tilde{h}_t$ 为备用激活状态；W_z，W_r，W_h，U_z，U_r，U_h 为权值矩阵，$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别为 sigmoid 和双曲正切激活函数。若 z_t 输出趋近于 1，表示有更多的当前信息能够传递下去，反之表示更多过去的信息能够传递到未来；若 r_t 趋于 1，表示当前时间步信息中保留更多过去的信息，反之忘记更多过去的信息。

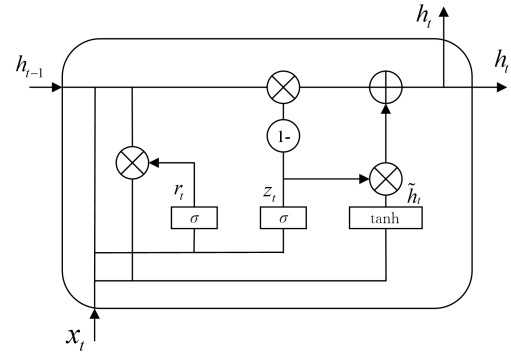


图 6 门控循环单元结构图

Fig. 6 Structure of gated recurrent unit

GRU 层中蕴含的单元数取决于输入特征序列的数量。由于语音时域特征信息按照时间序列排序，所以一层 GRU 网络层的单元数等于输入的特征时间序列个数，当第 t 单元在时序建模时，t 时刻的输入特征序列 x_t 和 $t-1$ 时刻的单元态特征 h_{t-1} 通过门控结构更新本单元的状态特征 h_t，因此 GRU 通过门控机制连接每一个单元可以捕获语音的时间序列信息。

3.4 自注意力机制

注意力机制(Attention Mechanism)源自于对人类视觉运动的广泛研究。当人类通过视觉感知并处理信息时，一般会有选择地将注意力集中在视觉空间的某些部分上，以便在需要的时间和地点获取信息，这种机制通常被称为注意力机制[28]。

深度学习中的注意力机制便源于人脑的注意力机制。人的大脑接受视觉或听觉等信息时，往往不会一次性处理和理解全部信息，而是有选择地将注意力集中在特征比较明显的局部信息上，从而利于滤除不重要的信息，提升信息的处理效率。与此相对应，深度学习的注意力机制通过对向量间的相似度计算，能够建模输入序列中的相关性，从而让网络的注意力集中在重要信息上，忽略无用信息。

自注意力机制是注意力机制的改进，有时称为内部注意力，是一种将特征序列的不同位置关联起来以计算序列表示的注意力机制[29]。矩阵点乘的相似度计算方式减少了对外部信息的依赖，能够更有效地捕捉数据或特征的内部相关性。

由于语音信号在时域上的前后帧具有高度的关联性，为此考虑使用自注意力机制[28-29]来关注音频的某一区域的纯净语音特征，从而获取清晰的目标语音，同时通过减弱对噪声的注意以降低干扰，然后随着训练过程调整注意力，最终达到增强输出语音帧而抑制噪声帧的效果。

本文所采用的自注意力网络层设计结构如图 7 所示。其输入为前一级网络输出的特征图 $G\in R^{T\times C}$。此处，T 为时间

维度，C 为通道数。自注意力机制层通过计算查询矩阵 Q、键阵 K 和值矩 V 之间的映射关系关注序列特征的全局相关性。Q，K 和 V 分别由各自对应的卷积网络对输入张量 G 进行卷积操作而计算得出，其计算过程如图 8 所示。

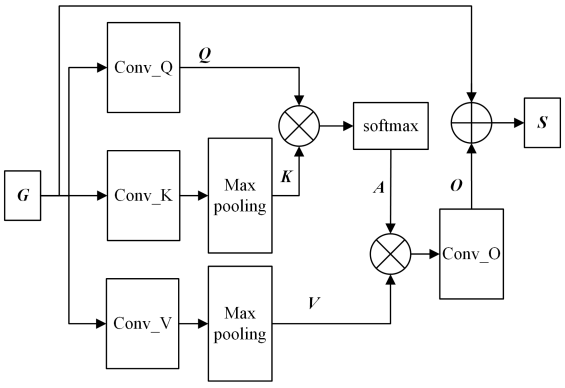


图 7　自注意力机制结构图

Fig. 7　Structure diagram of self-attention mechanism

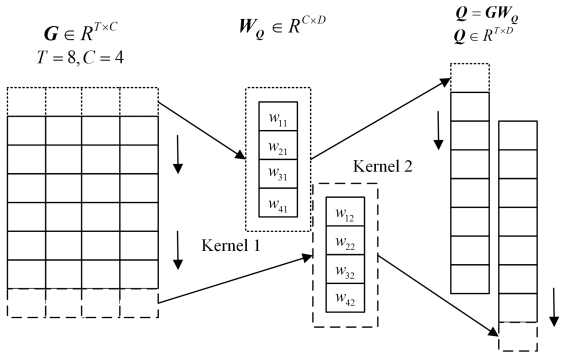


图 8　查询矩阵 Q 的计算过程示意图

Fig. 8　Schematic diagram of the calculation process of query matrix Q

以 Q 的计算为例，输入矩阵 G 由数据为 D、尺寸为 $C\times 1$ 的卷积核进行卷积，得到输出矩阵 Q。图 8 中为便于示例，假设 $T=8$，$C=4$，$D=2$。这种计算过程用公式可以表示为：

$$Q=GW_Q \tag{5}$$

$$K=GW_K \tag{6}$$

$$V=GW_V \tag{7}$$

其中，$W_Q\in R^{T\times D}$，$W_K\in R^{T\times D}$，$W_V\in R^{T\times D}$ 分别表示计算 Q，K 和 V 的卷积核向量组成的矩阵。$D=C/k$ 表示卷积之后输出张量的维度，即输出通道数目。k 为整数，表示经过卷积之后特征的通道数减少为 $1/k$。这种降维运算的目的是减少内存占用量。

此外，为了进一步降低内存占用量，通过两个最大池化层分别将 K 和 V 的时间维度降低为 $1/p$。最终可以得到：

$$Q\in R^{T\times \frac{C}{k}},K\in R^{\frac{T}{p}\times \frac{C}{k}},V\in R^{\frac{T}{p}\times \frac{C}{k}} \tag{8}$$

在本文中，令 $k=8$，$p=4$。

查询矩阵 Q 和键矩阵 K 转置的点积运算进行相似度计算，得到注意力权重，注意力权重与经 softmax(·)函数归一化的权重值相乘，便得到了含有注意力信息的特征矩阵 A。注意力映射 A 与值矩阵 V 的点积送入一个有 C 个卷积核的 1×1 卷积层，得到注意力输出 O。

$$A=softmax(QK^{\mathrm{T}}),A\in R^{T\times \frac{T}{p}} \tag{9}$$

$$O=(AV)W_O,W_O\in R^{\frac{C}{k}\times C} \tag{10}$$

其中，每个元素 $a_{ij}\in A$ 表示模型产生 O 的第 i 个输出 o_i 时对 V 的第 j 列 v_j 的关注程度。此外，注意力输出 O 之前的卷积操作用于将 O 的形状恢复为原来的形状 $T\times C$。

最后，为了防止特征丢失，将输入和注意力输出融合，最终的输出为：

$$S=\beta O+G \tag{11}$$

其中，β 是一种可习得的参数。

3.5　模型训练

本文所提出的融合 GRU 和 self-attention 的语音增强 GAN 网络的训练图如图 9 所示。模型的训练目标是将一个服从先验分布 $P_z(z)$的随机噪声样本 z 映射为服从纯净语音数据 $P_{\mathrm{data}}(x)$分布的增强语音样本 $\hat{x}$。GAN 网络训练过程中，生成器通过学习噪声语音 x_c 中的纯净语音特征，将随机噪声样本 z 映射为增强语音样本 $\hat{x}$。鉴别器通过对输入的语音样本提取特征，决定输入语音样本是真实数据还是虚假数据。鉴别器的输入分为两种，分别为纯净语音与带噪语音的拼接，以及增强语音与带噪语音的拼接。在对抗过程中，当鉴别器鉴别精度低下时，鉴别器通过反向传播机制调整其参数来提高其分类能力；当增强语音无法迷惑鉴别器时，鉴别器参数锁定，生成器根据反向传播机制调整其参数，使生成器的生成样本更加真实。生成器和鉴别器在对抗训练过程中不断地更新它们的参数，以使生成对抗模型能够更好地表示噪声语音与纯净语音之间的映射关系。

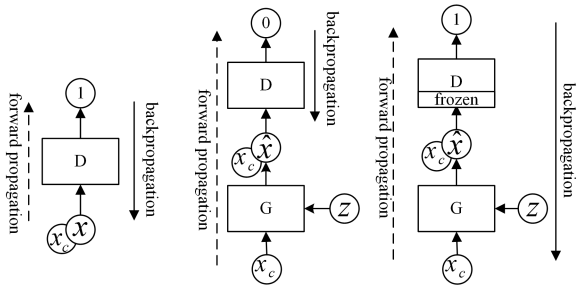


图 9　生成对抗网络的训练流程图

Fig. 9　Training flowchart of generative adversarial network

对于生成对抗网络的训练问题，研究者们提出了各种各样的损失来改善对抗训练。在这里，继续沿用基线算法 SA-SEGAN 使用的最小二乘损失。D 和 G 的最小二乘目标函数如下所示：

$$L(D)=\frac{1}{2}\mathbb{E}_{x,x_c\sim p_{\mathrm{data}}(x,x_c)}[(D(x,x_c)-1)^2]+\frac{1}{2}\mathbb{E}_{z\sim p_z(z),x_c\sim p_{\mathrm{data}}(x_c)}[D(G(z,x_c),x_c)^2] \tag{12}$$

$$L(G)=\frac{1}{2}\mathbb{E}_{z\sim p_z(z),x_c\sim p_{\mathrm{data}}(x_c)}[(D(G(z,x_c),x_c)-1)^2]+\lambda\|G(z,x_c)-x\|_1 \tag{13}$$

其中，D 为鉴别器，负责鉴别输入数据的真假；G 为生成器，负责生成新的数据；$\mathbb{E}[\cdot]$为期望运算，x 为纯净语音信号，x_c 为带噪语音信号，x 和 x_c 均采样于真实数据分布 $P_{\mathrm{data}}(\cdot)$；z 为随机噪声向量，采样于先验分布 $P_z(\cdot)$(如高斯噪声分布)；$G(z,x_c)$为生成器生成的增强语音，$\|\cdot\|_1$ 为 L1 范数，1 和 0 分别表示真实样本和虚假样本的标签。选择 L1 范数

来度量本算法增强语音与纯净语音之间的差距，因为它在多个深度学习领域已经被证明是成效显著的。通过这种方式，可以让损失函数产生更真实的结果。L1 范数的大小由一个超参数 λ 控制，本模型中 λ 固定为 100。

3.6 实验设置

除训练阶段，在数据预处理阶段，模型通过 50%重叠的滑动窗口从语音数据中提取语音样本(每段长度为 16 384×1，约为 1 s)，然后使用系数为 0.95 的预加重滤波器进行预加重处理。在测试阶段，将每个测试语音样本输入到训练后的神经网络，进行增强和去加重处理，最后连接生成增强的语音。

该模型实现基于 Tensorflow 框架，使用 R-MSprop 优化算法训练网络。实验迭代 100 次，最小批处理大小为 150，生成器和鉴别器学习速率均设为 0.000 2。

4 实验结果

4.1 基线设置

本研究共选用 6 种方法作为基线方法：噪声(Noisy，即未处理的语音信号)，维纳滤波(Wiener)[30]语音增强方法，SEGAN[20]语音增强算法，ISEGAN[21]语音增强算法，DSEGAN[21]语音增强算法，SASEGAN[22]语音增强算法。

4.2 数据集

实验所采用的数据集是 Valentini 2016 语音数据集[31]。该数据集中的纯净语音信号由 30 名讲话人录制而成，训练集和测试集分别包含 11 572 和 824 个干净的语音对。数据集中的含噪音频由说话人录制的纯净语音与 DEMAND 数据集中的 8 种真实的场景噪声(cafeteria，car，kitchen，meeting，metro，restaurant，station，traffic)和 2 种生成噪声合成得到。训练集中合成音频的信噪比分为 0 dB，5 dB，10 dB，15 dB 这 4 种情况，测试集中合成音频的信噪比按为 2.5 dB，7.5 dB，12.5 dB，17.5 dB 这 4 种情况，与训练集不同的信噪比设置让测试集可以有效地检验模型的泛化能力。

4.3 实验指标

采用 6 种客观质量评价指标对实验结果进行评估。

(1)语音质量的听觉估计(PESQ)[33]，评测分数范围为−0.5～4.5；

(2)语音信号失真平均主观意见分(CSIG)[34]，评测分数范围为 1～5；

(3)背景噪声入侵性平均主观分(CBAK)[34]，评测分数范围为 1～5；

(4)总体效果平均意见得分(COVL)[34]，评测分数范围为 1～5；

(5)语音信号分段信噪比(SSNR)，评测分数范围为 1～∞；

(6)短时客观可懂度(STOI)[35]，评测分数范围为 0～1。

所有指标以在整个测试数据集的平均值作为最终评测结果，6 种评测指标数值越大表示效果越好。

4.4 实验结果

SEGAN 与 SASEGAN 等基线方法的增强语音信号的效果如表 1 所列。需要注意的是，在表中 DSEGAN-BEST 为文献[21]中 DSEGAN 方法在本数据集上实验的最好效果。本文表示第 L 层(反)卷积层带有 self-attention 的 SASEGAN 为 SASEGAN-L，$4 \leqslant L \leqslant 11$。

表 1 Valentini 2016 测试集上基线方法的客观评估结果

Table 1 Objective evaluation results of baseline methods on test set Valentini 2016

method	*PESQ*	*CSIG*	*CBAK*	*COVL*	*SSNR*	*STOI*
Noisy	1.97	3.35	2.44	2.63	1.68	0.92
Wiener	2.22	3.23	2.68	2.67	5.07	0.91
SEGAN	2.19	3.39	2.90	2.76	7.36	0.93
DSEGAN-BEST	2.39	3.46	3.11	2.90	8.72	0.93
SASEGAN-4	2.36	3.57	3.08	2.95	8.38	0.93
SASEGAN-5	2.31	3.46	2.94	2.85	7.20	0.93
SASEGAN-6	2.38	3.46	3.12	2.90	8.86	0.93
SASEGAN-7	2.30	3.52	2.98	2.89	7.34	0.93
SASEGAN-8	2.34	3.55	3.03	2.92	8.03	0.93
SASEGAN-9	2.29	3.45	3.05	2.85	8.48	0.93
SASEGAN-10	2.41	3.62	3.06	2.99	7.87	0.93
SASEGAN-11	2.35	3.57	3.03	2.94	7.76	0.93
Average	2.34	3.52	3.04	2.91	8.05	0.93

GSA-SEGAN 语音增强方法在 Valentini 2016 测试集上增强语音信号的效果如表 2 所列。第 L 层(反)卷积层带有时间建模模块 1 或时间建模模块 2 的 GSASEGAN 表示为 GSA-SEGAN-1-L 和 GSA-SEGAN-2-L，$4 \leqslant L \leqslant 1$。根据表 1 和表 2，在 PESQ，CSIG，CBAK，COVL，SSNR 和 STOI 这 6 种指标的平均得分下进行对比。结果显示：对比所有指标的平均得分，本文算法得分明显优于 Wiener，SEGAN，ISEGAN，DSEGAN 和 SASEGAN 基线算法。

表 2 Valentini 2016 测试集上 GSA-SEGAN 方法的客观评估结果

Table 2 Objective evaluation results of GSA-SEGAN method on test set Valentini 2016

method	*PESQ*	*CSIG*	*CBAK*	*COVL*	*SSNR*	*STOI*
GSA-SEGAN-1-4	2.43	3.68	3.11	3.03	8.46	0.93
GSA-SEGAN-1-5	2.51	3.68	3.19	3.08	8.90	0.93
GSA-SEGAN-1-6	2.44	3.68	3.13	3.04	8.63	0.93
GSA-SEGAN-1-7	2.46	3.71	3.05	3.06	7.70	0.93
GSA-SEGAN-1-8	2.39	3.54	3.07	2.93	8.88	0.93
GSA-SEGAN-1-9	2.45	3.59	3.11	2.99	8.62	0.93
GSA-SEGAN-1-10	2.49	3.66	3.13	3.06	8.19	0.93
GSA-SEGAN-1-11	2.38	3.40	3.02	2.84	8.52	0.93
Average	2.44	3.62	3.10	3.00	8.49	0.93
GSA-SEGAN-2-4	2.42	3.50	3.03	2.92	8.14	0.93
GSA-SEGAN-2-5	2.41	3.56	3.04	2.95	8.11	0.93
GSA-SEGAN-2-6	2.40	3.60	3.12	2.98	8.93	0.94
GSA-SEGAN-2-7	2.48	3.62	3.10	3.02	8.45	0.93
GSA-SEGAN-2-8	2.47	3.69	3.17	3.06	8.93	0.94
GSA-SEGAN-2-9	2.50	3.58	3.14	3.02	8.48	0.93
GSA-SEGAN-2-10	2.41	3.66	3.10	3.02	8.25	0.93
GSA-SEGAN-2-11	2.42	3.59	3.14	2.99	8.92	0.93
Average	2.44	3.60	3.11	3.00	8.53	0.93

从本文算法与基线 SASEGAN 算法在所有设置情况下的平均指标值对比可以看出，虽然它们在 STOI 指标上相差无几，但是 GSA-SEGAN-1 在 *PESQ*，*CSIG*，*CBAK*，*COVL* 和 *SSNR* 分别获得 0.1，0.1，0.06，0.09，0.44 的绝对增益效果；GSA-SEGAN-2 则在 *PESQ*，*CSIG*，*CBAK*，*COVL* 和 *SSNR* 分别获得 0.1，0.08，0.07，0.09，0.48 的增益效果。GSA-SEGAN-1 与 GSA-SEGAN-2 最好的增强效果 PESQ 均能够达到 2.50 及以上，而 SASEGAN 最好的增强效果 PESQ 仅有 2.41。在测试集中，随机选择了一个被噪声严重侵袭的语音样本进行基线 SASEGAN 方法与两种 GSA-SEGAN 方法的

效果对比,包含语音波形的对比和语谱图的对比情况,如图10—图14所示(左栏为语音波形图,右栏为语谱图)。图13和图14中GSA-SEGAN-1方法、GSA-SEGAN-2方法和SA-SEGAN方法处于相同的网络设置情况下(生成对抗网络的第10层(反)卷积层带有时间建模模块或者自注意力机制层),所提方法的增强效果明显好于基线方法。

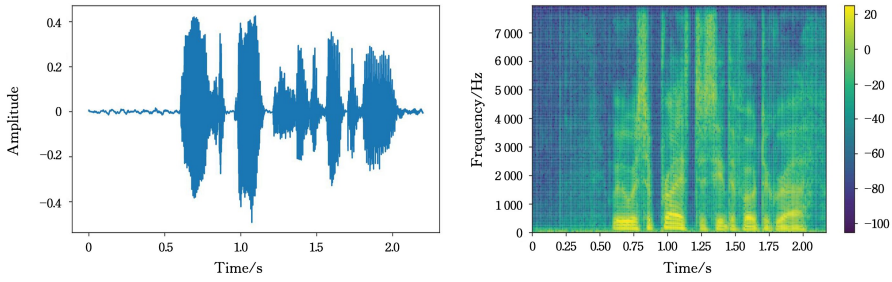


图10 纯净语音

Fig. 10 Clean speech

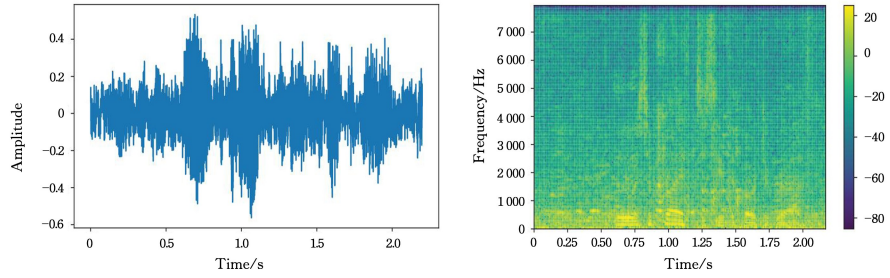


图11 带噪语音

Fig. 11 Noisy speech

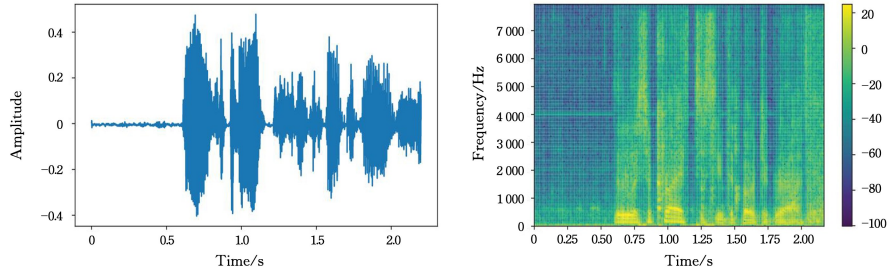


图12 SASEGAN增强语音

Fig. 12 SASEGAN enhanced speech

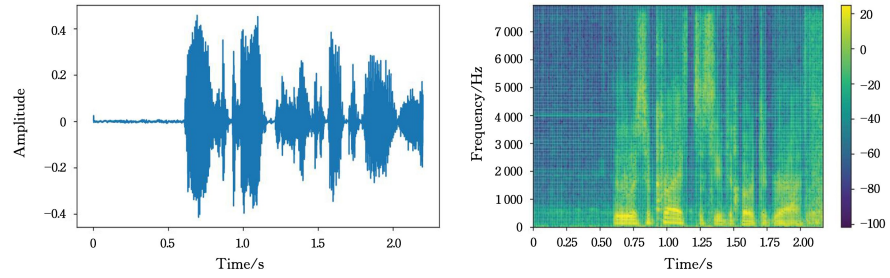


图13 GSA-SEGAN-1增强语音

Fig. 13 GSA-SEGAN-1 enhanced speech

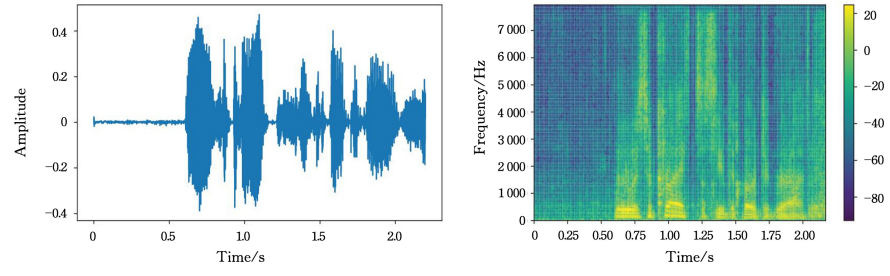


图14 GSA-SEGAN-2增强语音

Fig. 14 GSA-SEGAN-2 enhanced speech

在相同实验条件下，本文还测试了第 L 层(反)卷积层带有 GRU 的 SEGAN 结构的语音增强效果(表示为 G-SEGAN)，如表 3 所列。与 G-SEGAN 相比，GSA-SEGAN-1 和 GSA-SEGAN-2 在全部平均指标得分上明显领先。

表 3 Valentini 2016 测试集上 G-SEGAN 方法的客观评估结果

Table 3 Objective evaluation results of G-SEGAN method on test set Valentini 2016

method	*PESQ*	*CSIG*	*CBAK*	*COVL*	*SSNR*	*STOI*
G-SEGAN-4	2.41	3.63	3.12	3.00	8.51	0.93
G-SEGAN-5	2.37	3.63	3.05	2.98	8.13	0.93
G-SEGAN-6	2.39	3.56	3.03	2.95	7.86	0.93
G-SEGAN-7	2.40	3.50	2.94	2.91	6.97	0.93
G-SEGAN-8	2.30	3.48	3.01	2.85	8.73	0.93
G-SEGAN-9	2.35	3.63	3.04	2.97	7.93	0.93
G-SEGAN-10	2.45	3.66	3.09	3.03	8.14	0.93
G-SEGAN-11	2.40	3.58	3.13	2.97	9.00	0.93
Average	2.38	3.58	3.05	2.96	8.16	0.93

本文实验中，7 种方法使用的数据均相同。根据表 1、表 2 和表 3 的对比，两种 GSA-SEGAN 算法语音增强效果比基线 SASEGAN 算法和 G-SEGAN 算法要更加显著，这也说明 GRU 与 self-attention 联合提取特征的时间建模结构比单独使用 self-attention 或者单独使用 GRU 的网络结构能够更好地对语音时域特征的时间依赖性进行建模。需要注意的是，由于 GPU 内存的限制，本文没有验证网络模型前 3 层(反)卷积层加入自注意力机制或者时间建模模块的增强效果，这3层的特征谱的时间维度分别为 8192，4096 和 2048。

4.5 计算开销

在本文实验中，使用训练时每个 epoch 的平均处理时间测评了 GSA-SEGAN 算法和基线 SASEGAN 算法的计算开销，训练所使用的实验环境如表 4 所列。

表 4 实验环境

Table 4 Experimental environment

实验环境	环境配置
主机	DESKTOP-4ICAR7I
操作系统	Windows 10
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz 2.39GHz
内存	16.0GB
GPU	NVIDIA GeForce RTX 2080Ti
Python 版本	Python 3.5
深度学习框架	TensorFlow 1.90

表 5 展示了两种 GSA-SEGAN 方法与基线方法在模型第 L 层(反)卷积层带有时间建模模块或者自注意力模块的情况下，每个 epoch 需要运行的时间。运行时间越长，则计算开销越大。

表 5 相同配置下 GSA-SEGAN 方法和 SASEGAN 方法的计算开销对比

Table 5 Comparison of computation cost between GSA-SEGAN method and SASEGAN method in the same configuration

(单位：min)

method	SASEGAN-L	GSA-SEGAN-1-L	GSA-SEGAN-2-L
$L=4$	8.91	77.22	78.31
$L=5$	8.59	33.48	33.59
$L=6$	8.53	20.30	20.35
$L=7$	8.48	13.39	13.41
$L=8$	8.49	10.26	10.27
$L=9$	8.37	9.82	9.13
$L=10$	8.45	9.45	9.05
$L=11$	8.43	9.23	9.07

如表 5 所列，加入 GRU 的两种 GSA-SEGAN 方法明显比基线方法的计算开销要大，尤其是在第四层(反)卷积层带有时间建模模块的情况下，两种 GSA-SEGAN 方法的每个 epoch 运行时间分别达到了 77.22 min 和 78.31 min。但是在 $L=8,9,10,11$ 这 4 种情况下，由于模型中特征图尺寸的缩小，两种 GSA-SEGAN 方法的计算开销接近于 SASEGAN 方法。

结束语 本文结合 self-attention 层和 GRU 层提出了两种耦合在生成对抗网络中的时间建模模块，改进了以往端到端 GAN 语音增强方法对语音时域特征时间依赖性建模角度单一的问题。在设备处理内存足够的前提下，提出的时间建模模块可以用于 SEGAN 生成器和鉴别器的不同(反)卷积层，甚至所有的卷积层。实验结果显示，在所有的客观评价指标上，两种 GSA-SEGAN 方法的增强效果均优于 SASEGAN 基线方法。这也表明，GRU 与 self-attention 的联合使用比单独使用 self-atten-tion 或 GRU 进行时间依赖性建模更加有效。需要注意的是，这样的网络结构设置会随着特征的时间维度增大给算法带来巨大的计算开销，下一步将对此问题的改善进行研究。研究结果表明，GRU 和 self-attention 组成的时间建模模块可以从特征时间序列和特征全局充分地进行时间依赖性建模，并且能够显著提升语音增强的效果。此外，它可以很容易地应用到现有的网络结构，帮助模型充分提取多重特征。

参 考 文 献

[1] LAN T, PENG C, LI S, et al. Review of monophonic speech noise reduction and dereverberation research [J]. Computer Research and Development, 2020, 57(5): 26.

[2] XIANG Q, TANG Y. Research on Chinese Speech Enhancement Technology Based on Generative Adversarial Networks [J]. Computer Application Research, 2020(S02): 150-151.

[3] LOIZOU P C. Speech enhancement: theory and practice [M]. CRC Press, 2007.

[4] WANG H, LI J, ZHAO H M, et al. Speech enhancement algorithm based on sparse low-rank model and phase spectrum compensation [J]. Computer Engineering and Applications, 2018, 54(5): 6.

[5] BOLL S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[6] LIM J S, OPPENHEIM A V. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 1979, 67(12): 1586-1604.

[7] MCAULAY R, MALPASS M. Speech enhancement using a softdecision noise suppression filer[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(2): 137-145.

[8] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.

[9] TAHA T M F, ADEEL A, HUSSAIN A. A survey on techniques for enhancing speech[J]. International Journal of Computer Applications, 2018, 179(17): 1-14.

[10] WANG Y, WANG D L. Towards scaling up classification-based speech separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1381-1390.

[11] FU S W，TSAO Y，LU X. SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement[C]// Interspeech. 2016：3768-3772.

[12] TAN K，CHEN J，WANG D L. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2018，27(1)：189-198.

[13] HUANG P S，KIMM，HASEGAWA-JOHNSON M，et al. Joint optimization of masks and deep recurrent neural networks for monaural source separation[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2015，23(12)：2136-2147.

[14] XIAO C X，CHEN Y. Real-time speech enhancement algorithm based on recurrent neural network [J]. Computer Engineering and Design，2021，42(7)：6.

[15] WANG Z，ZHANG T，SHAO Y，et al. LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement[J]. Applied Acoustics，2021，172：107647.

[16] BAO C C，XIANG Y. Review of single-channel speech enhancement methods based on deep neural network [J]. Signal Processing，2019，35(12)：11.

[17] XU Y，DU J，DAI L R，et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2014，23(1)：7-19.

[18] GAO G，YIN W B，CHEN Y，et al. A Speech Enhancement Method Based on Generative Adversarial Networks in Time-Frequency Domain [J]. Computer Science，2022，49(6)：6.

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

[20] PASCUAL S，BONAFONTE A，SERRAJ. SEGAN：Speech enhancement generative adversarial network[J]. arXiv：1703.09452，2017.

[21] PHAN H，MCLOUGHLIN I V，PHAM L，et al. Improving GANs for speech enhancement[J]. IEEE Signal Processing Letters，2020，27：1700-1704.

[22] PHAN H，LE NGUYEN H，CHÉNO Y，et al. Self-attention generative adversarial network for speech enhancement[C]// ICASSP 2021 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2021：7103-7107.

[23] DONAHUE C，LI B，PRABHAVALKAR R. Exploring speech enhancement with generative adversarial networks for robust speech recognition[C]// 2018 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2018：5024-5028.

[24] LI P，JIANG Z，YIN S，et al. Pagan：A phase-adapted generative adversarial networks for speech enhancement[C]// 2020 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2020：6234-6238.

[25] HE K，ZHANG X，REN S，et al. Delving deep into rectifiers：Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：1026-1034.

[26] TONG T，LI G，LIU X，et al. Image super-resolution using dense skip connections[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：4799-4807.

[27] CHO K，VAN MERRIËNBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine trans-lation[J]. arXiv：1406. 1078，2014.

[28] MNIH V，HEESS N，Graves A. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems，2014，27.

[29] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Advances in Neural Information Processing Systems，2017，30.

[30] LIM J，OPPENHEIM A. All-pole modeling of degraded speech [J]. IEEE Transactions on Acoustics，Speech，and Signal Processing，1978，26(3)：197-210.

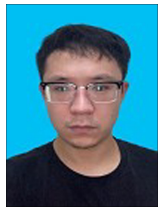
[31] VALENTINI-BOTINHAO C，WANG X，TAKAKI S，et al. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]// SSW. 2016：146-152.

[32] THIEMANN J，ITO N，VINCENT E. The diverse environments multichannel acoustic noise database (demand)：A database of multichannel environmental noise recordings[C]// Proceedings of Meetings on Acoustics ICA2013. Acoustical Society of America，2013.

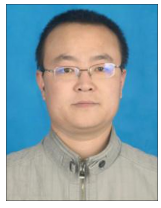
[33] UNION I T. Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs[J]. International Telecommunication Union，Recommendation P，2007，862.

[34] HU Y，LOIZOU P C. Evaluation of objective quality measures for speech enhancement [J]. IEEE Transactions on Audio，Speech，and Language Processing，2007，16(1)：229-238.

[35] TAAL C H，HENDRIKS R C，HEUSDENS R，et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]// 2010 IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2010：4214-4217.



ZHANG Dehui，born in 1997，master，is a student member of China Computer Federation. His main research interests include deep learning and speech enhancement.



DONG Anming，born in 1982，Ph.D，associate professor，postgraduate supervisor，is a member of China Computer Federation. His main research interests include Time series signal processing，wireless communication and artificial intelligence.