基于深度卷积生成对抗网络的语音生成技术

朱　纯[1]；王翰林[2]；魏天远[2]；王　伟[2]
（1. 东南大学 软件工程系，江苏 南京 211189；2. 南京医科大学 生物医学工程系，江苏 南京 211166）

摘要： 提出了一种基于深度卷积生成对抗网络（Deep Convolutional Generative Adversarial Networks，DCGAN）的语音生成技术，通过大量学习语音库，能够自主生成全新的语音。生成式对抗网络是一种近年来大热的深度学习模型，其由一个判别网络（Discriminator，D）和一个生成网络（Generator，G）组成。使用 Tensorflow 作为学习框架，利用 DCGAN 模型对大量语音进行训练。在基本训练过程中，语音生成网络 G 的目标就是尽量生成真实的、接近自然的语音去欺骗语音判别网络 D，而 D 的目标就是尽量把 G 生成的语音和真实的语音区分出来，语音生成网络努力生成的语音让判别网络认为是真实的语音，利用 G 和 D 构成动态“博弈过程”，最终生成接近原始学习内容的自然语音信号，实现语音的自动生成。

关键词： 深度学习；人工智能；生成对抗网络；语音生成
中图分类号： TP181　　**文献标志码：** A　　**文章编号：** 1006－2394（2018）02－0013－03

Speech Generation Based on Depth Convolution for Adversarial Networks

ZHU Chun[1], WANG Han-lin[2], WEI Tian-yuan[2], WANG Wei[2]
（1. Department of Software Engineering, Southeast University, Nanjing 211189, China;
2. Department of Biomedical Engineering, Nanjing Medical University, Nanjing 211166, China）

Abstract: This paper presents a speech generation technology based on Deep Convolutional Generative Adversarial Networks (DCGAN), which can generate new speech by learning a large number of voice libraries. Generative confrontation network is a kind of deep learning model in recent years, which consists of a discriminant network Discriminator (D) and a generating network Generator (G). In this paper, tensorflow is used as the learning framework, and the large number of speech is trained by DCGAN model. In the basic training process, the voice generation network G's goal is to generate voice which is as real as possible, close to the natural voice to deceive the voice to identify the network D. And D's goal is to try to distinguish the G-generated voice and the real voice. The voice generation network tries to generate the voice which is identified as the real voice by the network. Thus, G and D constitute a dynamic "game process", and ultimately generate the natural voice signal close to the original learning content to achieve automatic voice generation.

Key words: deep learning; artificial intelligence; generation of confrontation network; speech generation
DOI:10.19432/j.cnki.issn1006-2394.2018.02.004

0　引言

近年来，深度学习技术被广泛应用于各类数据处理任务中，例如图像、语音以及文本等领域[1-3]。生成对抗网络（Generative Adversarial Nerworks，GAN）是一种近年来大热的深度学习模型[4]，自 2014 年 Ian Goodfellow 创造性地提出了生成对抗网络之后，生成对抗网络就成为了学术界的一个热门的研究热点。深度卷积生成对抗网络（Deep Convolutional Generative Adversarial Nerworks，DCGAN）是 GAN 的一个衍生模型，其利用卷积神经网络的特征提取能力提高了生成网络的学习效果。

使智能设备具有“说话”的功能，这在真正的“面对面人机交流”中扮演着很重要的角色[5]。借助于语音生成系统，智能设备已经可以清晰、自然地说话，普通用户很容易听懂并接受。然而，现有能说话的智能设备往往只能根据固定的语音库来发声，模式单调缺乏变化，且不够自然。

本文提出了一种基于深度卷积生成对抗网络的语音生成技术，通过大量学习语音库，最终自主生成全新的语音，可以解决在人机面对面交流过程中智能设备过度依赖固定的语音库来发声和模式单调缺乏变化、

收稿日期： 2017－09
基金项目： 江苏省高等学校大学生创新创业训练计划（201610312053X）
作者简介： 朱纯（1995—），男，硕士研究生，研究方向为机器学习。

相似度低且不够自然的问题。

1 DCGAN 的原理和实现步骤

GAN 网络模型包含一个生成网络和一个判别网络[6]，GAN 的目标函数是关于生成网络与判别网络的一个零和游戏，也是一个最小值和最大值的问题。生成对抗网络训练一个生成器，从随机噪声或者潜在变量中生成逼真的样本，同时训练一个鉴别器来鉴别真实数据和生成数据，两者同时训练，直到达到一个纳什均衡[7]，此时生成网络生成的数据与真实的样本是一样的，判别网络也无法正确地区分生成网络生成的数据和真实的数据。GAN 的网络结构如图 1 所示[8]。

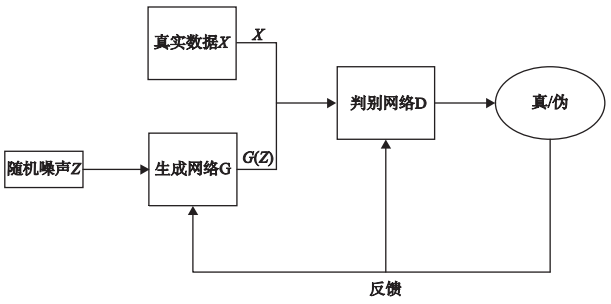

图 1 生成对抗网络的基本架构

生成对抗网络实现的具体步骤如下:

第 1 步，生成对抗网络存在两个网络，一个是生成网络 G，用于接收一个随机的噪声 z，一个是判别网络 D，判别生成的数据是不是“真实的”。

第 2 步，计算生成网络 G 的损失函数:

$$(1-y)\log(1-D(G(z)))$$

其中，z 表示生成网络接收一个随机的噪声，$G(z)$ 表示生成网络的输出，log 表示以 10 为底的对数操作，$D(G(z))$ 表示判别网络 D 判断生成网络 G 生成的数据是否为真实的概率。

第 3 步，计算判别网络 D 的损失函数:

$$-((1-y)\log(1-D(G(z)))+y\log D(x))$$

其中，x 表示输入参数，即真实样本数据，$D(x)$ 表示判别模型的输出，即输入 x 为真实数据的概率，$G(z)$ 表示生成模型的输出，log 表示以 10 为底的对数操作，$D(G(z))$ 表示判别网络 D 网络判断生成网络 G 生成的数据是否为真实的概率。

第 4 步，计算优化函数:

$$\min_G\max_D V(D,G)=IE_{X\sim P_{data}(x)}[\log D(x)]+IE_{Z\sim P_Z(z)}[\log(1-D(G(z)))]$$

其中，z 表示生成模型接收一个随机的噪声，x 表示输入参数，$G(z)$ 表示生成模型的输出，log 表示以 10 为底的对数操作，$D(G(z))$ 表示判别模型 D 判断生成模型 G 生成的数据是否为真实的概率，$P_z(z)$ 表示随机噪声 z 的概率密度，$P_{data}(x)$ 表示参数数据 x 的概率密度。

第 5 步，在训练过程当中，生成网络 G 尽可能生成真实的数据去欺骗判别网络 D，而判别网络 D 则尽量把生成网络 G 生成的数据和真实的数据区分开来，最终生成网络和判别网络形成了一个动态的“博弈过程”。

第 6 步，判断判别网络 D 是否能判别生成网络 G 所生成的数据为真实，若能，则得到训练好的数据，否则，需要调整输入参数 x 后继续执行第 2 步。

第 7 步，理想状态下，生成模型 G 生成足以“以假乱真”的 $G(z)$，对于判别模型 D 而言，无法判定 G 所生成的数据是否真实，即 $D(G(z))=0.5$。

DCGAN 是 GAN 网络的一大改进，其在训练方面更加稳定，并且能够生成更高质量的样例。DCGAN 的原理和 GAN 的原理是相似的，其只是将 GAN 中的生成网络和判别网络替换为了两个卷积神经网络。但也不是简单地直接换就行了，DCGAN 对卷积神经网络的结构做了一些变化，以用来提高样本数据的质量以及收敛的速度。DCGAN 在判别网络中用步幅卷积取代了所有的池化层，在生成网络中用微步幅卷积取代所有的池化层。除以上之外，DCGAN 在生成网络和判别网络中都使用批量归一化方法，去掉了全连接层，使网络变成了全卷积网络，在生成网络中，除了输出层采用了双曲正切 Tanh 作为激活函数，其他层都采用 ReLU 作为激活函数，判别网络中的所有层都采用了 LeakyReLU 作为激活函数。

2 利用 DCGAN 对语音数据进行训练

利用深度卷积生成对抗网络生成数据通常需要三步，第一步是收集原始数据，第二步是选择一个合适的学习框架来搭建训练环境，最后一步就是将收集到的原始数据导入 DCGAN 进行训练并生成想要得到的数据。

本文利用 audacity 软件采集了大量语音数据，语音信号是一种非平稳的并且是时变的信号，其携带着很多信息。为了便于后续处理，本文设计了一个低通巴特沃斯滤波器对采集好的语音数据进行了处理，以达到去噪的目的，并将处理好的语音数据转化成统一的格式。

在对语音数据进行预处理之后，将处理好的语音数据导入深度卷积生成对抗网络的数据文件夹当中。在利用深度卷积生成对抗网络开始训练之前，需要对相关参数进行设置，选取合适的参数，才能够得到正确的结果。

本文采用 Tensorflow 作为学习框架[9]，并基于 Tensorflow 学习框架搭建深度卷积生成对抗网络，网络搭建完成后，在 Ubuntu 系统的终端输入以下运行指令进行训练：

```
python main.py --dataset yuyin --input_height=200 --output_height=200 --c_dim=1 --is_train --epoch 1 000
```

参数含义：dataset yuyin 表示指定语音数据，input_height=200 表示 yuyin 文件夹中的语音数据块的大小是 200×200 的，output_height=200 是用来指定生成的语音数据块的大小为 200×200，epoch 1 000 表示跑 1 000个 epoch，1 个 epoch 等于使用训练集中的全部样本训练一次。

3　实验结果分析

原始语音波形和利用 DCGAN 生成的语音波形对比如下所示，其中图 2 是原始语音波形图，图 3 是利用 DCGAN 生成的语音波形图。

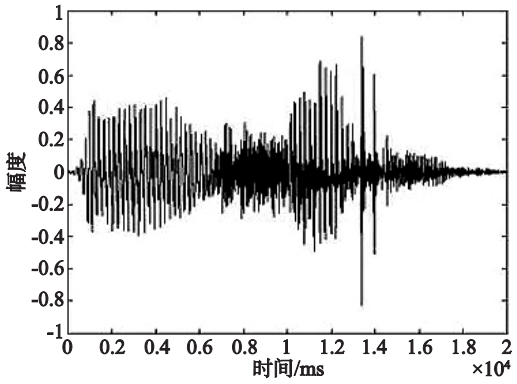

图 2　原始语音波形图

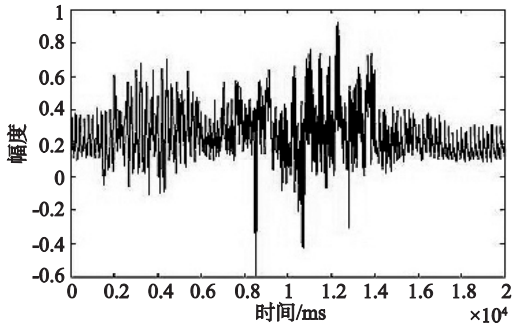

图 3　利用 DCGAN 生成的语音波形图

为了更好地将 DCGAN 生成的语音信号与原始语音信号进行对比，本文对原始语音信号与利用深度卷积生成对抗网络生成的语音信号进行了进一步的处理。

本文从频域的角度对原始语音信号与利用深度卷积生成对抗网络生成的语音信号的频谱进行对比分析，利用 matlab 软件中的 wavread 命令来读入 WAV 格式的语音信号，先将语音信号赋值给一向量，然后将该向量看作是一个普通的信号，对其先进行滤波处理，再进行快速傅里叶变换进行频谱分析，实验结果如图 4、图 5 所示。

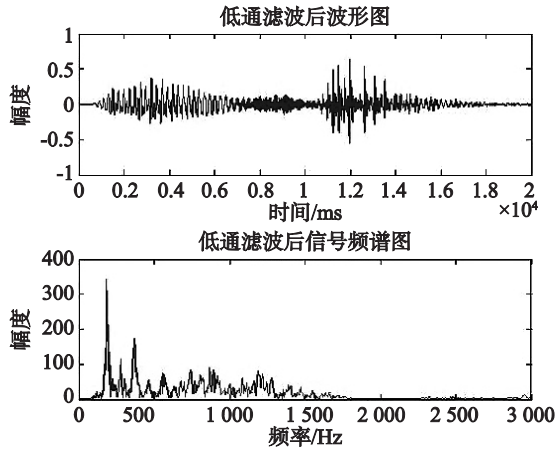

图 4　原始语音信号低通滤波后的波形以及低通滤波后的频谱图

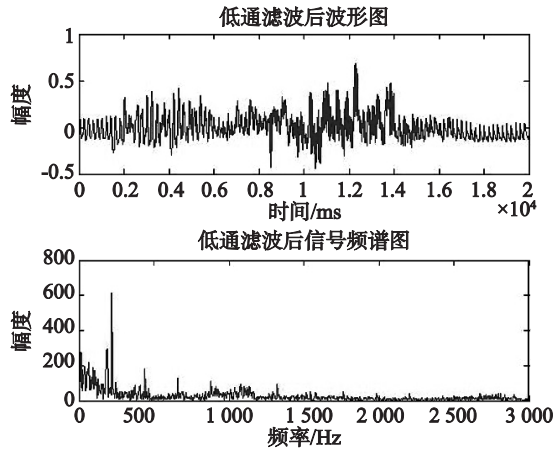

图 5　利用 DCGAN 生成的语音信号低通滤波后的波形以及低通滤波后的频谱图

由图 4 和图 5 可以看出原始语音信号的频谱与利用深度卷积生成对抗网络生成的语音信号的频谱的相同之处和不同之处。

两者在频率为 200 Hz 和 250 Hz 时幅值达到波峰，在频率为 500 Hz 和 1 000 Hz 时幅值达到波谷，在频率超过 1 500 Hz 后，幅值随频率变化比较平缓。但是在频率为 0～100 Hz 之间，原始语音信号的频谱幅值几乎为 0，而利用深度卷积生成对抗网络生成的语音信号的频谱幅值较大，这主要是因为本文使用的语音数据量较少（小于 10 000），导致利用 DCGAN 生成的语音信号存在低频干扰。

4　总结与展望

本文首先介绍了生成对抗网络的原理，然后将原始语音库导入深度卷积生成对抗网络进行训练，实现

（下转第 20 页）

测,确定零序电流行波到达母线处的初始时刻;

(3) 求取各个馈线零序暂态电流信号的 HMSE 以及系统 HMSEM;

(4) 根据 HMSEM,采用相应的方法对故障进行检测。

3.3 仿真验证

为了验证基于 HMSEM 的故障选线算法的正确性,利用 EMTP-ATP 在不同类型的故障线路下对图 1 各种故障工况时发生单项接地故障进行仿真验证。分别当电缆线路(l_1) 距离母线在 2 km、7 km、13 km 处发生单相接地故障,电缆混合线路(l_2) 距离母线在 2 km、12 km、22 km 发生单相接地故障,架空线路(l_5) 距离母线在 3 km、21 km、40 km 处发生单相接地故障,母线处发生单相接地故障,并且故障相位分别设为 0°、30°、60°、90°,接地电阻分别为 2 Ω、20 Ω、200 Ω、1 000 Ω、2 000 Ω,其选线结果都是正确的。

4 结论

针对小电流接地电网单相接地故障选线问题,将 HHT 应用于经过消弧线圈的单相接地故障检测中,首先通过选择一个提取故障信号暂态分量的时间窗口,提取馈线零序电流的暂态分量,使用 HHT 对暂态信号作时频分析;其次利用 EMTP-ATP 仿真不同情况下的单相接地故障,将 EEMD 方法引入信号突变检测,使用 EEMD 检测信号馈线零序电流信号到达母线的时刻;最后引入 HMSEM 的概念,分析各种情况下(主要是不同接地电阻和相位),各馈线零序暂态电流的 HMSE 的特点,应用所提出的基于 HMSEM 的选线法进行选线。仿真结果表明了基于 HMSEM 的选线方法不受线路结构的影响,有较高的选线的可靠性和准确率。

参考文献:

[1] N. E. Huang, Z. Shen, S. R. Long, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [J]. Proc R Soc Lond A, 1998 (454): 903 - 995.

[2] 张小丽. 基于希尔伯特-黄变换的输电线路故障行波定位与保护方法 [D]. 长沙: 长沙理工大学, 2008.

[3] 李丹丹. 基于 HHT 与遗传算法的配电网单相接地故障测距研究 [D]. 青岛: 中国石油大学, 2010.

[4] 蔡晋, 林榛, 高伟, 等. 基于 HHT 及信号注入的配电网谐振与单相接地故障识别 [J]. 电气技术, 2015(12): 31 - 35.

[5] 汤涛, 黄纯, 江亚群, 等. 基于高低频段暂态信号相关分析的谐振接地故障选线方法 [J]. 电力系统自动化, 2016, 40(16): 105 - 111.

[6] 董红生, 邱天爽, 张爱华, 等. 基于 HHT 边际谱熵和能量谱熵的心率变异信号的分析方法 [J]. 中国生物医学工程学报, 2010, 29(3): 336 - 344.

[7] 张林. 基于 HHT 的单相接地故障检测研究 [D]. 合肥: 合肥工业大学, 2013.

[8] 束洪春, 赵文渊, 彭仕欣. 配电网缆-线混合线路故障选线的 HHT 检测方法 [J]. 电力自动化设备, 2009, 29(5): 4 - 9.

[9] 张海申. 暂态信号小波分析系统开发及谐振接地系统故障选线研究 [D]. 成都: 西南交通大学, 2011.

(郁菁编发)

(上接第 15 页)

语音的生成,并将利用深度卷积生成对抗网络生成的语音信号与原始语音信号进行了对比。

如果可以获得更多的语音样本进行训练,效果应该会更好。随着 GAN 算法的不断改进,语音生成技术必将获得更大的发展。

参考文献:

[1] 郭丽丽,丁世飞. 深度学习研究进展 [J]. 计算机科学,2015(5): 28 - 33.

[2] 马冬梅. 基于深度学习的图像检索研究 [D]. 呼和浩特: 内蒙古大学,2014.

[3] 陈先昌. 基于卷积神经网络的深度学习算法与应用研究 [D]. 杭州: 浙江工商大学,2014.

[4] Goodfellow IJ, Pougetabadie J, Mirza M, et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems, 2014(3): 2672 - 2680.

[5] 白金刚. 语音生成中口腔鼻腔气流压力检测设备的设计与实现 [D]. 天津: 天津大学, 2014.

[6] 张喜升. 对抗样本和生成对抗网络——深度学习中的对抗方法综述 [D]. 天津: 南开大学, 2016.

[7] Sánchez-Gutiérrez ME, Albornoz EM, Martinez-Licona F, et al. Deep learning for emotional speech recognition [M]. Berlin: Springer International Publishing, 2014.

[8] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望 [J]. 自动化学报, 2017, 43(3): 321 - 332.

[9] 张俊, 李鑫. TensorFlow 平台下的手写字符识别 [J]. 电脑知识与技术, 2016, 12(16): 199 - 201.

(郁菁编发)