基于DC-CNN的电子伪装语音还原研究

王永全[1,2] 施正昱[1,2,3] 张 晓[4]

(华东政法大学刑事司法学院 上海 201620)[1] (华东政法大学信息科学与技术系 上海 201620)[2]

(复旦大学大数据学院 上海 200433)[3]

(公安部第三研究所信息网络安全公安部重点实验室 上海 200120)[4]

摘 要 针对电子伪装语音还原研究在还原模型的构建方面并无突破性进展的状况,提出了一种基于扩大的因果卷积神经网络(Dilated Casual-Convolution Neural Network,DC-CNN)的电子伪装语音还原模型。该还原模型以DC-CNN为框架,对电子伪装语音历史采样点的声学信息与还原因子进行卷积和非线性映射运算。同时模型的神经网络采用跃层连接技术以优化深层传递,再经过压扩转换后输出还原语音。该模型具有非线性映射性、扩展性、多适应性与条件性、并发性等明显特点。在实验分析中,以3个基本变声功能:音调(pitch)、节拍(tempo)和速度(rate)对钢琴曲和英文语音分别进行电子伪装变声处理,再经模型还原,将还原语音与原始语音进行声纹特征比对、LPC数据分析和语音同一性的人耳测听辨识,结果表明,还原语音与原始语音的声纹特征十分吻合,且实现了高质量的共振峰波形复原,钢琴曲和英文语音的共振峰参数总体还原拟合率分别达到79.03%和79.06%,远超电子伪装语音与原始语音35%的相似比例,这说明该模型能有效削减语音中的电子伪装特征,较好地实现了电子伪装的钢琴曲和英文语音的还原。

关键词 DC-CNN,电子伪装语音,还原语音,还原因子,门激活单元

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.030

Study on Restoration of Electronic Disguised Voice Based on DC-CNN

WANG Yong-quan[1,2] SHI Zheng-yu[1,2,3] ZHANG Xiao[4]

(School of Criminal Justice,East China University of Political Science and Law,Shanghai 201620,China)[1]

(Department of Information Science and Technology,East China University of Political Science and Law,Shanghai 201620,China)[2]

(School of Data Science,Fudan University,Shanghai 200433,China)[3]

(Key Laboratory of Information Network Security of Ministry of Public Security,The Third Research Institute of the Ministry of Public Security,Shanghai 200120,China)[4]

Abstract Aiming at the fact that there is no breakthrough in modeling for the electronic disguised voicer estoration, this paper proposed a new model based on Dilated Casual-Convolution Neural Network (DC-CNN) for restoring electronic disguised voice. DC-CNN is used as the framework of restoring model,and convolution and nonlinear mapping are performed on the historical sampling acoustic information and restoring factors of the electronic disguised voice. Meanwhile,the model's neural network adopts skip-connection for deep transmission and outputs the restoring voice after companding transformation. The model has obvious characteristics such as nonlinear mapping,expansibility,adaptability and conditionality,concurrency,etc. In the experiment,the original voice was processed by three basic disguised functions:pitch,tempo and rate. Then,voiceprint features comparison,LPC analysis and voice identity of human audiometry recognition were made between restoring voice and original voice. The voiceprint of the restoring voice fits that of the original voice perfectly,and high quality formant waveform restoration is achieved. The piano music's and English voice's

到稿日期:2018-10-05 返修日期:2018-12-15 本文受2014年国家社会科学基金重大项目(第二批)(14ZDB147),公安部科技强警基础工作专项项目(2017GABJC33),教育部2017年第二批"云数融合科教创新"基金课题(2017B06106),华东政法大学《人工智能导论》通识重点课程建设项目(A-0312-18-174794)资助。

王永全(1964—),男,博士,教授,博士生导师,主要研究方向为网络空间安全、大数据与人工智能,E-mail:wangyongquan@ecupl.edu.cn(通信作者);**施正昱**(1996—),女,硕士生,主要研究方向为大数据与人工智能;**张 晓**(1987—),女,硕士,助理研究员,主要研究方向为信息网络安全、电子数据与声像资料司法鉴定。

general restoring fitting rates of the formant's parameters are 79.03% and 79.06% respectively, which are much higher than the similarity of electronic disguised voice to original voice. The results turn out that this model can minify the electronic disguised characteristics effectively and it is efficient on the restoration of electronic disguised piano music and English voice.

Keywords DC-CNN, Electronic disguised voice, Restoring voice, Restoring factor, Gated activation units

1 引言

电子伪装语音是一种严重畸变的语音,它从根本上掩盖了声源的声学特征。变声器与变声软件等在网络、游戏、聊天工具等系统中的广泛应用使得不法分子有机可乘,而经其伪装后的语音较一般采用物理方式伪装的语音更难以辨识。实验表明,语音经过电子伪装后,由熟悉说话者的人进行测听并依次辨识说话者的性别,其准确率仅为 20%[1];而由不熟悉说话者的人进行测听,则有 90%的人表示无法辨识出说话者的性别。据研究统计[2],通过电子语音伪装系统对语音进行处理,当仅更改语音的基频时,利用电声学仪器提取的电子伪装语音的声波频谱与原始语音的频谱的相似比例仅为 35%。在司法鉴定实践中,语音经过电子伪装后将大大降低声纹鉴定的有效识别率[3],导致无法对变声处理后的说话人及时进行认定,这给留有电子伪装语音证据的犯罪活动侦破工作造成了很大的困难。目前电子伪装语音的研究主要集中于鉴别语音是否经过电子伪装[4-5]以及电子伪装的特征提取[1,6]等方面,而在电子伪装语音的还原模型构建方面并无突破性进展。因此,电子伪装语音的还原及其模型构建问题已经成为司法鉴定领域研究的热点之一,在这方面现有的研究多采用传统的物证技术,研究样本较小,声纹特征的抓取和还原算法的构建面临着极其繁杂的工作和极大的挑战,还原技术的研究困难重重。为此,本文提出了一种基于扩大的因果卷积神经网络(Dilated Casual-Convolution Neural Network, DC-CNN)的电子伪装语音还原模型,并在以钢琴曲为例的纯音乐音频和以英文为例的含语言音频还原中取得了良好的效果。实验表明,该模型所生成的还原语音与原始语音的频谱相似率远远高于电子伪装语音与原始语音的频谱相似率,且钢琴曲与英文语音的共振峰参数总体还原拟合率均达到约 80%,在原始语音与还原语音同一性的人耳测听辨识中也取得了良好的效果,这在司法鉴定实践中对电子伪装语音进行还原具有重要的理论和实践意义。

2 电子伪装语音还原

电子伪装语音(Electronic Disguised Voice, EDV)是指采用电子设备或音频处理软件对原始语音通过变声伪装处理后所产生的畸变语音[4]。电子伪装语音对声源身份的伪装程度极高,它改变了原始语音中的声学特征,不仅给人耳测听辨识带来极大困难,而且通过电声学仪器检测也难以判别。

电子伪装语音还原是指通过一定的算法模型来弱化或消除语音中的电子伪装特征,生成更为接近原始音频的语音。电子伪装语音一般基于某种算法来实现自身声学特征的改变,原始语音转换为电子伪装语音的过程存在一定的变化规律。而同一声源的音频又具有短时平稳的特性,因此可以通过统计对比原始语音与电子伪装语音之间的声纹偏差特征,为电子伪装语音的还原提供依据。

3 基于 DC-CNN 的电子伪装语音还原模型

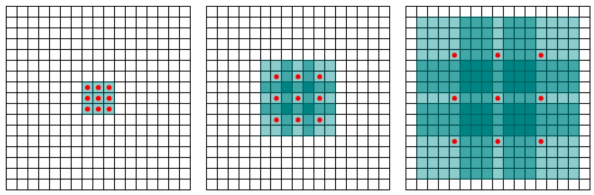
本文提出的还原模型采用了与 PixelCNN[7] 类似的不含池化层(Pool Layer)的多层卷积堆叠与门激活函数模型。该模型以扩大的因果卷积神经网络(DC-CNN)为基础,通过控制神经网络中每个神经元的门激活单元引入还原特征 h,从而实现电子伪装语音的还原。模型中所采用的 DC-CNN 存在因果卷积与扩大卷积,在语音合成模型 WaveNet[8-9] 中,DC-CNN 取得了良好的成效。

3.1 扩大的因果卷积神经网络(DC-CNN)

一般地,卷积神经网络(Convolution Neural Network, CNN)的每个神经元由负责提取上一神经元局部特征的特征提取层和在该神经元计算过程中所需的多个特征映射平面共同组成的特征映射层构成[10-11]。

因果卷积(Casual Convolution)[12]多用于具有一定排列顺序的数据,处理较长的序列化数据时有着良好的建模效果。因果卷积的序列化特点使其十分适合处理语音这种时序性极强的数据信息。但如果仅使用因果卷积,则需要极深的神经网络或极大的卷积核才能获得较好的训练结果。而过深的神经网络与过大的卷积核不仅会大大降低运算效率,而且容易造成模型训练难以收敛或者退化的现象。

为了弥补这一缺陷,在模型的卷积神经网络中引入了扩大卷积。扩大卷积[13-14]是一种稀疏化的卷积核,它通过忽略部分输入数据来增加感受野的范围,即按照一定规则在原始的卷积核中增加零来生成"扩张"卷积核。如图 1 所示,阴影部分表示在这一扩展系数下的感受野,圆点表示实际的卷积核。



(a) Dilated=1 (b) Dilated=1 (c) Dilated=1

图 1 扩展系数与感受野的关系示例[13]

Fig. 1 Illustration of dilated coefficient and receptive field[13]

由图 1 可知，随着扩展系数的增加，感受野以指数形式扩大。

如图 2 所示，将扩大卷积与因果卷积相结合，形成扩大的因果卷积神经网络(DC-CNN)。该网络既可以控制语音数据按时间顺序向后有序传输，又能在不增加神经网络层数与卷积核大小的情况下扩大感受野，使其在处理语音信号数据时有着优良的性能。

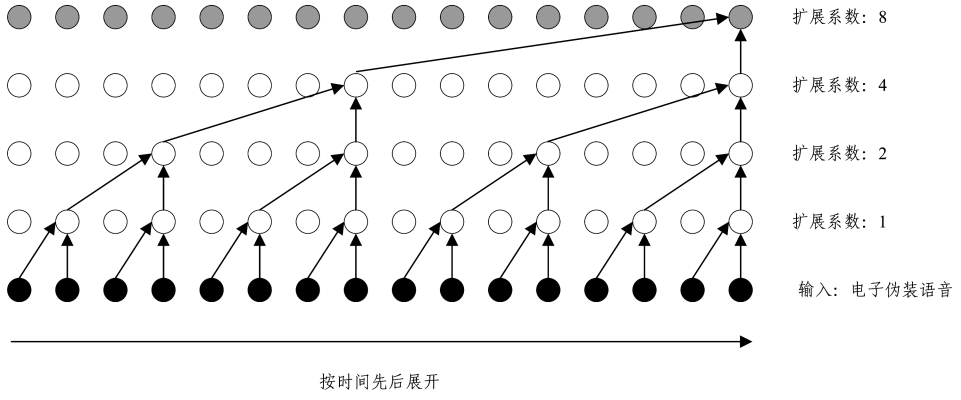


图 2　扩大的因果卷积神经网络示意图[8]

Fig. 2　Visualization of DC-CNN[8]

3.2　基于 DC-CNN 的电子伪装语音还原模型的结构

语音信号 x_T 是由 T 时刻之前的输入语音信号与还原因子 h 来预测还原的。一段时间下的语音信号序列 $X=(x_1, x_2, \cdots, x_T)$ 的多维联合变量分布可表示为：

$$P(X)=\prod_{t=1}^{T} P(x_t \mid x_1, x_2, \cdots, x_{t-1}, h) \tag{1}$$

如图 3 所示，为使还原语音序列按上述条件概率生成，在基于 DC-CNN 的电子伪装语音还原模型的神经网络主体中，采用多层的扩大因果卷积块堆叠建模，并通过引入门激活函数实现非线性映射。

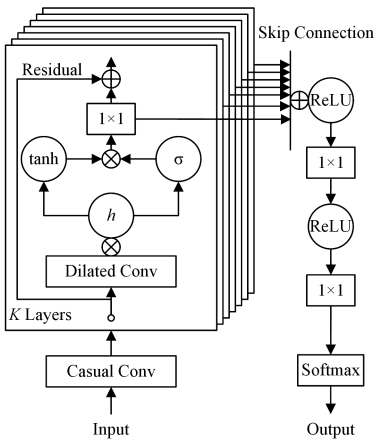


图 3　基于 DC-CNN 的电子伪装语音还原模型的结构

Fig. 3　Structure of electronic disguised voice restoration model based on DC-CNN

3.3　门激活与还原因子

在基于 DC-CNN 的电子伪装语音还原模型中，采用门激活单元(Gated Activation Units，GAU)[15]使神经网络具备分层的非线性映射学习能力。

Furui 等[6]主张声道的共振特性与频谱包络较基音频率对语音的贡献更大。因此，在上述还原模型中，选取语音中共振峰的中心频率、带宽和强度为主要参数，将原始语音与实验中选取的不同伪装程度下的电子伪装语音转换成采样频率为 8000 Hz、16 位、单声道的统一格式后，导入 FIAS 智能声纹鉴定工作站。首先对音频中的多声源进行声源分离，而后分别采集各个声源的宽带语谱图，并通过长时平均的 LPC 数据分析方法生成每个音节的发音部分共振峰的中心频率、带宽和声强[16]。经共振峰折损清洗、合并优化、序列调整等一系列预处理之后，分别计算原始语音与各相应电子伪装语音所对应的各条共振峰主要参数的回归方程，这些回归方程分别表征了各相应电子伪装语音与原始语音的偏差。记各项回归系数为对应的还原因子 h，h 与卷积后的电子伪装语音信息按式(2)整合：

$$Z=\tanh(W_{f,k} * X \cdot h) \cdot \sigma(W_{g,k} * X \cdot h) \tag{2}$$

其中，$*$ 表示卷积运算，W 指可学习的卷积滤波器(Learnable Convolution Filter，LCF)，k 表示层数，f 和 g 分别代表滤波(Filter)与门控(Gate)。

3.4　深层传递优化

模型中使用了 Residual Learning[17]的框架来促进模型收敛并使梯度传递至更深层次，以缓解因神经网络加深而导致的性能下降。如图 4 所示，在一个残差块结构中，将某层神经网络的输入 x 通过跃层连接，恒等映射至更深的神经网络层并叠加于该层网络卷积所得的残差值之上，经优化计算即可得到期望输出。

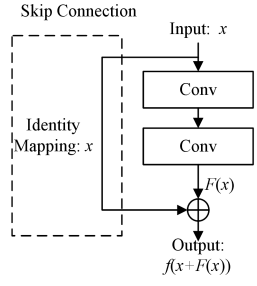


图 4　残差块结构

Fig. 4　Structure of residual block

3.5 还原语音的输出

输出层采用 Softmax 函数对合成运算后的数据进行离散化分类。Softmax 函数的输入是一个 N 维的实数向量，设为 x，其函数表达式为：

$$\zeta(x)_i=\frac{e^{x_i}}{\sum_{n=1}^{N}e^{x_i}},i=1,2,\cdots,N \tag{3}$$

就其本质而言，Softmax 函数能将一个 N 维的任意实数向量映射为一个各个元素的取值都在(0,1)范围内的 N 维向量[18]，实现向量的归一化。

为降低模型系统的运算量，通过 μ-law 压扩转换使输出的数据量降至 2^8，即 $\mu=255$，以提高模型的预测效率。

$$f(x_t)=sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)},|x_t|<1 \tag{4}$$

4 还原模型的实验分析

采用变声软件 SoundTouch 的 3 个基本变声功能即音调(pitch)、节拍(tempo)和速度(rate)对钢琴曲和英文的原始语音进行电子伪装变声处理，总计获得 53 份不同伪装程度的电子伪装语音包。将每个电子伪装语音包及其所对应的还原因子 h 放入还原模型中分别进行训练。

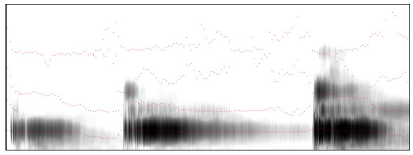
训练实验采用了一台处理器为 Xeon(R)E5-2620，计算单元为 NVIDIA Quadro M4000 高性能的服务器，以训练具有 20 层卷积神经网络的还原模型。这 20 层卷积神经网络被划分为 2 个卷积块，单个卷积块中的扩展系数依次为(2^0，2^1，2^2，…，2^9)。该还原模型中感受野的大小为 128 ms，跃层连接的通道为 256 条，初始学习率设定为 10^{-3}。训练集选择了 869 段音频，测试集分别包括钢琴曲 2 段、英文语音 503 段，全部音频的采样率均为 16 kHz 且以 16 bits 量化。

经统计，钢琴曲平均每次迭代的训练时长为 5.031 6 s，英文语音平均每次迭代的训练时长为 3.727 3 s，训练后每生成 1 s 语音平均用时约 36.15 s。

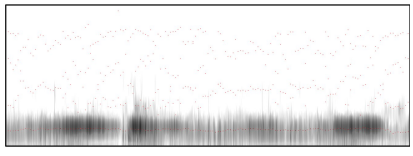
4.1 还原语音的声纹特征

提取还原后的钢琴曲和英文语音的宽带语谱图，观察其语谱包络结构，并进行声纹特征分析。如图 5 所示，还原生成的钢琴曲噪音较弱，音频的连续性较好，在低频部分的还原率高。此外还可以看出，还原语音的高频段弱化，部分高频信息缺失。相较于钢琴曲电子伪装语音还原，英文语音的还原所需的总时长更长，还原难度更大。还原后的英文语音与其对应的原始语音的声纹极为相似，电子伪装语音还原模型在共振峰低频部分表现良好，但有较为明显的噪音。声纹虽呈现出细小间断，但对声纹鉴定的干扰较小。

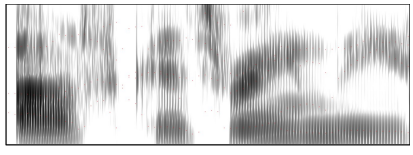
通过分别比对钢琴曲与英文语音的宽带语谱图声纹特征可知，模型还原出的语音的宽带语谱图清晰、波形明显，且还原率较高。



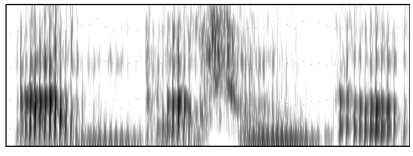
(a)钢琴曲的原始语音的宽带语谱图



(b)钢琴曲的还原语音的宽带语谱图



(c)英文语音的原始语音的宽带语谱图



(d)英文语音的还原语音的宽带语谱图

图 5 还原语音与原始语音的宽带语谱图(部分)

Fig. 5 Voice broad band spectrum of restoring voice and original voice(part)

4.2 还原语音的 LPC 数据分析

线性预测编码(Linear Predictive Coding，LPC)于 1967 年首次由板仓等应用于语音分析合成中[19]，之后便被广泛应用于语音信号处理技术之中。

如图 6 和图 7 所示，实验对还原生成的钢琴曲与英文语音分别进行 LPC 数据分析，还原语音与原始语音的共振峰图形走势一致，各共振峰峰值位置大体吻合，仅在声音的强度上有所偏差。图 6、图 7 中，黑色实线和灰色实线分别表示还原语音和原始语音的 LPC 数据。

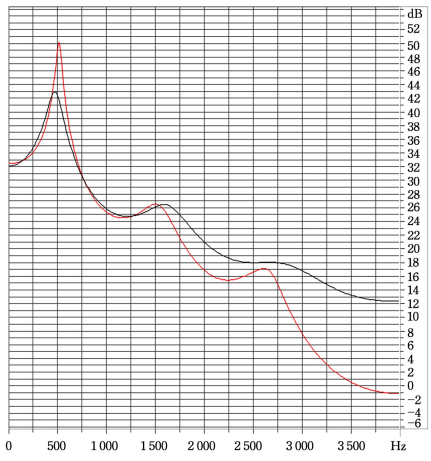


图 6 钢琴曲的还原语音与原始语音的 LPC 数据分析图(部分)

Fig. 6 LPC diagram of restoring voice and original voice of pianolude (part)

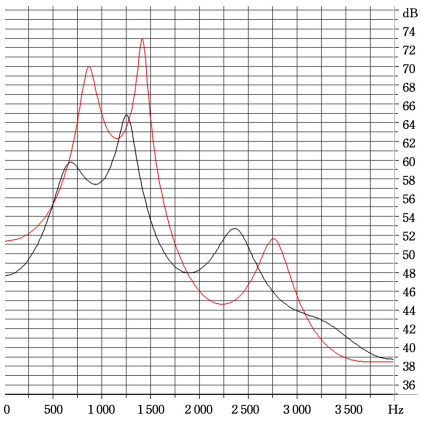


图 7　英文语音的还原语音与原始语音的 LPC 数据分析图(部分)

Fig. 7　LPC diagram of restoring voice and original voice of English vocal(part)

经 LPC 数据分析得到钢琴曲和英文语音的还原语音共振峰中心频率、带宽和强度，对这些主要参数分别计算其还原语音算术平均和原始语音算术平均之间的偏差。从表 1 可以看出，钢琴曲与英文语音的还原语音在中心频率上与其相应的原始语音十分接近，两者的绝对平均偏差分别为 3.79%与 0.97%；声音强度次之，各伪装方式下的绝对偏差比例均在 13%以内；仅带宽有着一定程度的偏移。

表 1　还原语音与原始语音的主要参数偏差

Table 1　Main parameters deviation of restoring voice and original voice

(单位：%)

音频	参数	Pitch	Rate	Tempo
钢琴曲	中心频率	6.31	4.47	0.60
	带宽	25.93	67.21	63.52
	强度	−9.98	0.55	−10.16
	总体绝对偏差=20.97 总体还原拟合率=79.03			
英文语音	中心频率	0.47	−2.09	−0.35
	带宽	49.22	71.82	29.46
	强度	−11.33	−10.85	12.91
	总体绝对偏差=20.94 总体还原拟合率=79.06			

注：总体还原拟合率 =100%−总体绝对偏差

分析结果证明，本文所提还原模型可以实现高质量的共振峰波形复原：钢琴曲与英文语音共振峰参数的总体还原拟合率分别达到了 79.03%和 79.06%，较电子伪装语音与原始语音 35%[2]的相似比例分别提升了 44.03%和 44.06%。

4.3　还原语音的同一性人耳测听辨识

除电声学仪器检测外，实验还邀请了 15 名志愿者，分别对钢琴曲和英文语音各自的电子伪装语音、还原语音与其对应的原始语音是否具有同一性进行人耳测听辨识。

在表 2 所列的统计结果中，钢琴曲与英文语音的还原语音与其对应的原始语音的同一性比例较其分别对应的电子伪装语音与其原始语音的同一性比例有大幅提升，最高提升了 46.67%，最低也提升了 26.66%，这表明该还原模型能有效削减语音中的电子伪装特征，使得还原语音在人耳主观测听辨识方面更接近原始语音。

表 2　语音同一性人耳测听辨识

Table 2　Human audiometry recognition of voice identity

(单位：%)

音频	同一性辨识	Pitch	Rate	Tempo
钢琴曲	电子伪装语音	33.33	20.00	46.67
	还原语音	73.33	60.00	80.00
	提升比例	40.00	40.00	33.33
英文语音	电子伪装语音	26.67	20.00	26.67
	还原语音	53.33	66.67	73.33
	提升比例	26.66	46.67	46.66

由于受到噪音影响，还原语音的人耳测听辨识结果较声纹特征与 LPC 数据分析存在一定差异，因此对于噪声强的还原语音，志愿者在人耳测听辨识时判断其与原始语音为同一声源的比例会偏低。此外，由于还原语音生成时进行了 μ-law 压扩转换，还原语音的质量受到了一定影响，听觉效果欠佳，使得部分音频在人耳测听辨识实验中表现力不足。

4.4　还原模型的敏感度

上述的实验结果表明，该还原模型能适用于多种不同电子伪装方式的语音还原。进一步分析可以发现，除还原因子 h 之外，还原模型对于不同的电子伪装方式的敏感度同时也受到模型训练时的超越参数的影响。譬如通过调节模型训练过程中的初始迭代率，会使还原结果在不同电子伪装的条件下有着不同的表现。还原结果表明，把对于 tempo 方式所伪装的语音进行还原时能取得良好效果的迭代率应用于 rate 与 pitch 方式所伪装的语音还原，取得的还原效果不同：对于 rate 而言，会导致还原训练的过程中引入较多的噪声，从而影响还原效果；对 pitch 而言，还原的结果比 tempo 和 rate 两种伪装方式的还原结果更易产生过拟合现象。因此，通过调整迭代率，可使还原模型更好地适用于不同电子伪装方式的还原处理。

5　还原模型的特征

由还原模型的结构及其应用结果可知，基于 DC-CNN 的电子伪装语音还原模型具有以下 4 个明显的特征。

(1)非线性映射性：在语音信号的建模中，非线性模型的效果比一般的线性模型的效果更好，为了使还原模型能更好地解决语音生成问题，本文所提出的模型在各神经网络层采用门激活函数进行非线性运算，使得模型具有非线性因素，以此提升其还原性能。

(2)扩展性：由于基于 DC-CNN 的电子伪装语音还原模型中含有扩大卷积结构，因此在相同的神经网络层数下模型具有更宽的感受野，单位时间内可接收的音频数据量得到了扩展。

(3)多适应性与条件性：实验验证表明，对于 tempo，rate，pitch 3 种电子伪装方式转换的 53 种不同的电子伪装语音，该还原模型均表现出很好的适应性。这表明通过修改还原因子 h 来改变还原条件，可使该还原模型适用于多种不同方式形成的电子伪装语音的还原。

(4)并发性：对于较长的电子伪装语音，可对其进行切割

并按时序编号后输入该还原模型进行并发处理，这将大大缩短还原过程中的训练时长，使得还原过程更为高效。

结束语 本文提出的基于 DC-CNN 的电子伪装语音还原模型实现了对钢琴曲和英文语音在不同伪装方式下的电子伪装语音的还原，从宽带语谱图、LPC 数据分析和人耳测听辨识 3 个方面对还原语音进行测评，取得了较好的还原效果。这不仅是电子伪装语音还原模型研究方面的一个突破性进展，对我们深入开展相关研究也具有重要的理论指导意义和实践价值。特别是在此基础上，针对中文语音发音结构较为复杂、音节种类和变化情况较多的情况，进行中文电子伪装语音的还原研究将是下一步的重要工作之一。

参 考 文 献

[1] 张翠玲，赵晓波. 电声伪装语音的声学研究[C]// 第七届中国语音学学术会议暨语音学前沿问题国际论坛. 北京，2006.

[2] ZHANG C L，TAN T J，LIU S. Study on Automatic Speaker Recognition of Disguised Voices [J]. Forensic Science and Technology，2007(2)：18-21. (in Chinese)
张翠玲，谭铁军，刘昇. 伪装语音的自动话者识别研究[J]. 刑事技术，2007(2)：18-21.

[3] GONZALEZ R，KANERVISTO A，HAUTAMÄKI V，et al. Perceptual Evaluation of the Effectiveness of Voice Disguise by Age Modification[J]. arXiv：1804. 08910，2018.

[4] TAO D Y. Study on Speaker Recognition Under Electronic Disguised Voices[D]. Nanjing：Nanjing University of Posts and Telecommunications，2016. (in Chinese)
陶定元. 电子伪装语音下的说话人识别方法研究[D]. 南京：南京邮电大学，2016.

[5] LI Y P，TAO D Y，LIN L. Study on Electronic Disguised Voice Speaker Recognition Based on DTW Model Compensation [J]. Computer Technology and Development，2017(1)：93-96. (in Chinese)
李燕萍，陶定元，林乐. 基于 DTW 模型补偿的伪装语音说话人识别研究[J]. 计算机技术与发展，2017(1)：93-96.

[6] ZHANG G Q，JIN Y Z，LIU H W，et al. Study on Changing Rules of Electronic Disguised Voice [J]. Evidence Science，2010，18(4)：503-509. (in Chinese)
张桂清，金怡珠，刘红伟，等. 电子伪装语音的变声规律研究[J]. 证据科学，2010，18(4)：503-509.

[7] OORD A，KALCHBRENNER N，VINYALS O，et al. Conditional Image Generation with PixelCNNDecoders[J]. arXiv：1606. 05328，2016.

[8] OORD A，DIELEMAN S，ZEN H，et al. WaveNet：A Generative Model for Raw Audio[J]. arXiv：1609. 03499，2016.

[9] CHEN K，ZHANG W，DUBNOV S，et al. The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation[J]. arXiv：1811. 08380，2018.

[10] YIN W，KANN K，YU M，et al. Comparative Study of CNN and RNN for Natural Language Processing[J]. arXiv：1702. 01923，2017.

[11] FU W B，SUN T，LIANG J，et al. Review of Principle and Application of Deep Learning[J]. COMPUTER SCIENCE，2018，45(s1)：24-28，53. (in Chinese)
付文博，孙涛，梁藉，等. 深度学习原理及应用综述[J]. 计算机科学，2018，45(s1)：24-28，53.

[12] 伍宏，传顾宇，凌震华. 基于深度卷积神经网络的语音参数合成器[C]// 第十四届全国人机语音通讯学术会议. 江苏，2017.

[13] YU F，KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions [C]// International Conference on Learning Representations. 2016.

[14] WANG Z，JI S. Smoothed Dilated Convolutions for Improved Dense Prediction [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. London，2018.

[15] TANAKA M. Weighted Sigmoid Gate Unit for an Activation Function of Deep Neural Network[J]. arXiv：1810. 01829，2018.

[16] 王永全. 声像资料司法鉴定实务[M]. 北京：法律出版社，2013.

[17] MCCANE B，SZYMANSKI L. Some Approximation Bounds for Deep Networks[J]. arXiv：1803. 02956，2018.

[18] LIU G，XU C，CHEN S Y，et al. Image Classification with Stacked Restricted Boltzmann Machines and Hybrid Neural Network [J]. Journal of Chinese Computer Systems，2017，38(9)：2146-2151. (in Chinese)
刘罡，徐超，陈思义，等. 结合深度置信网络与混合神经网络的图像分类方法[J]. 小型微型计算机系统，2017，38(9)：2146-2151.

[19] 赵力. 语音信号处理[M]. 北京：机械工业出版社，2009：72.