计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500148

# 基于深度学习的语音合成与转换技术综述

潘孝勤　芦天亮　杜彦辉　仝　鑫
中国人民公安大学信息网络安全学院　北京 100038
(m18811328909@163.com)

**摘　要**　语音信息处理技术在深度学习的推动下发展迅速,其中语音合成和转换技术相结合能实现实时高保真的指定对象、内容的语音输出,在人机交互、泛娱乐等领域具有广泛的应用前景。文中旨在对基于深度学习的语音合成与转换技术进行综述。首先,简要回顾了语音合成和转换技术的发展历程;接着,列举了在语音合成、转换领域的常见公开数据集以便研究者开展相关探索;然后,讨论了从文本到语音模型,包括在风格、韵律、速度等方面进行改进的经典和前沿的模型、算法,并分别对比评述了其效果与发展潜力;进一步针对语音转换进行综述,归纳总结了转换方法与优化思路;最后,总结了语音合成与转换的应用与挑战,并根据其在模型、应用和规范方面所面临的问题,展望了未来在模型压缩、少样本学习和伪造检测方面的发展方向。

**关键词**:语音信息处理;语音合成;语音转换;深度学习;生成对抗网络
**中图法分类号**　TP301;TP18

## Overview of Speech Synthesis and Voice Conversion Technology Based on Deep Learning

PAN Xiao-qin,LU Tian-liang,DU Yan-hui and TONG Xin
College of Informationand Cyber Security,People's Public Security University of China,Beijing 100038,China

**Abstract**　Voice information processing technology is developing rapidly under the impetus of deep learning. The combination of speech synthesis and voice conversion technology can achieve real-time high-fidelity voice output of designated objects and content,and has broad application prospects in man-machine interaction,pan-entertainment and other fields. This paper aims to provide an overview of speech synthesis and voice conversion technology based on deep learning. First,this paper briefly reviews the development of speech synthesis and voice conversion technology. Next,it enumerates the common public datasets in these fields so that it is convenient for researchers to carry out related explorations. Then,it discusses the TTS models,including the classic and cutting-edge models and algorithms in terms of style,rhythm,speed,and compares their effects and development potentials respectively. Then,it reviews voice conversion by summarizing the voice conversion methods and optimization methods. Finally,it summarizes the applications and challenges of speech synthesis and voice conversion,and looks forward to their future development direction in model compression,few-shot learning and forgery detection,based on the problems faced by them in terms of model,application and regulation.

**Keywords**　Voice information processing,Speech synthesis,Voice conversion,Deep learning,Generative adversarial networks

## 1　引言

人工智能(Artificial Intelligence,AI)技术带来的产业变革正逐步深入各行各业,面对层出不穷的智能化产品,人机交互的用户体验很大程度地影响着产品效果。其中,从文本到语音(Text To Speech,TTS)和语音转换(Voice Conversion,VC)两种语音信息处理技术相结合,可在实现让机器"开口说话"的基础上,产生用户想要的真人声音。近年来,谷歌DeepMind、蒙特利尔学习算法研究所和百度硅谷人工智能实验室等团队开展的TTS和VC相关研究成绩斐然,部分成果已形成产品落地应用,如RealTalk[1],Melnet[2]等软件可以仅通过输入文本就生成逼近真人的声音。

纵观发展历程,早在1997年,Mbius等[3]就设计出基于统计方法的语音合成管道系统,后续学者的改进大都延续这种模块化的研究思路,如使用树模型、隐马尔可夫模型[4-5]和高斯混合模型[6-7]实现频谱建模。随着深度学习的诞生,神经网络开始逐步替代上述统计模型,并削弱了声学知识在研究过程中的重要性,其中代表性的方法Deep Voice[8],Tacotron[9]等进一步提高了合成效果。而生成对抗网络(Generative Adversarial Networks,GAN)[10]等前沿深度学习技术的

到稿日期:2020-05-27　返修日期:2020-12-03　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2017YFB0802804);中国人民公安大学基本科研业务费重大项目(2020JKF101)
This work was supported by the National Key R&D Program of China(2017YFB0802804) and Fundamental Research Funds for the Central Universities of PPSUC(2020JKF101).
通信作者:芦天亮(lutianliang@ppsuc.edu.cn)

发展和语音领域公开数据集的不断增多,让语音合成、转换系统对语音的刻画描述能力持续增强。

本文将目光放在阐述语音合成与转换的技术原理和相关发展上,简要概括语音信息处理早期使用的传统机器学习方法的不足,重点梳理近年来在深度学习推动下提出的新方法,从 TTS 和 VC 两方面着手,归纳整理技术沿革,分析各类模型、算法的特点优势与不足,并对该领域未来的发展方向进行探索和展望。

## 2 语音合成和转换的常用数据集

表 1 列出了部分公开且常用的数据集,其中 LJSpeech,LibriSpeech,VCTK 等数据集中的语音片段主要来自书籍、报纸朗读,配合相关的文本数据集可用于训练从文本到语音的模型。Voxceleb2,TED-LIUM 3 等数据集中的语音片段则来自采访视频、演讲等视听片段。ASVspoof 2019 数据库与上述通用数据集略有不同,该库主要提供检验基于 TTS 和 VC 的语音信息处理效果的数据集。

表 1 语音合成和转换数据集

Table 1 Speech synthesis and voice conversion data sets

| 数据集 | 时间 | 语言 | 发声人数 | 片段数/时长 |
|---|---|---|---|---|
| ASVspoof 2019 | 2020 年 | 英语 | 107 | 121 407 个 |
| M-AILABS Speech | 2018 年 | 英语 | — | 102 小时 |
| LJSpeech | 2018 年 | 英语 | 1 | 13 100 个 |
| Voxceleb2 | 2018 年 | 多语种 | 9 000+ | 1 128 246 个 |
| TED-LIUM 3 | 2018 年 | 多语种 | 2 351 | 452 小时 |
| Aishell | 2017 年 | 中文 | 400 | 178 小时 |
| SurfingTech | 2017 年 | 中文 | 855 | 102 600 个 |
| VCTK | 2017 年 | 英语 | 109 | 44 070 个 |
| LibriSpeech | 2015 年 | 英语 | 8 | 1 000 小时 |

这些公开数据集的不断丰富为神经网络逐渐代替原有的统计学习方法,以及完全端到端的、大规模的深度学习系统的出现提供了强有力的数据基础,并推动语音合成和转换技术不断取得突破性进展,最终生成了高度逼真的目标语音。

## 3 从文本到语音

从文本到语音(TTS)是指利用机器学习模型将给定文本转换成语音并加以输出的一种技术,也是语音合成技术的核心一环。具体来说,TTS 重点解决的问题是从字素到音素的转换,可细分为文本预处理、创建语音数据及分析确定韵律特征 3 个阶段[11],具体包括分隔单词,扩展符号、数字、缩写,辨析同形词发音消除歧义,确定语调、重音和持续时间等诸多方面,其转换难度往往因语种而异。以基于隐马尔可夫模型(Hidden Markov Models,HMM)的语音合成为代表的早期 TTS,在使用统计参数的同时,会不可避免地破坏自然语音频谱的精细结构,并且其受限于算力,往往只能考虑相邻 1 个或 2 个音素的影响,导致前文潜在的有意义信息被丢弃,造成了信息损失。为了多维度地保证生成语音的质量,传统模块化管道系统中的部分统计学习模型被神经网络模型替代,通过深度学习训练组件提升合成语音质量;随着一系列语音生成模型的提出,利用神经网络实现端到端的方案出现,简化了系统构建的复杂性,语音合成技术得到进一步升级。

### 3.1 非端到端的深度学习 TTS

随着深度学习技术在语音合成领域逐渐受到关注并得以应用,管道系统中部分使用传统机器学习方法实现的模块被神经网络取代,合成音频质量得到提升。文献[12-13]利用前馈神经网络将从输入文本中导出的语言表示直接映射到声学特征,相比传统的手工特征工程方法,其能够发现更多的隐含特征,在一定程度上提升了下游模型的效果上限,并且通过梯度下降方法进行优化,在基于大规模数据进行训练时能够得到优于传统统计学习的效果。但该方法存在的问题在于利用前馈网络提取特征时很难进行控制,可能会输出冗余特征信息。为了精简特征、降低噪声干扰,Wu 等[14]提出一种深度降噪自动编码器技术,以非线性、数据驱动、无监督的方式自动从高维谱特征中提取低维特征,在为下游模型提供更有意义的信息的同时,也通过特征降维降低了模型训练的难度,使之更易收敛。此外,文献[15-16]使用深度信念网络(deep belief network),替代 HMM 直接建模语言和声学表示之间的关系,输出语音轨迹的概率密度。

尽管前馈网络有效地改进了传统语音合成方法的不足并极大地提升了生成音频样本的质量,但仍存在以下缺陷:前馈网络将序列化的声波流数据进行一次性输入和处理,忽略了声音数据时间维度上的先后信息,可能会导致信息丢失。为了改善这一问题,基于循环神经网络(Recurrent Neural Network,RNN)和一维卷积的方法被相继提出。例如,为提升语音的自然性和清晰度,使用堆叠前馈网络和双向长短期记忆网络(Long Short-Term Memory,LSTM)直接预测汉字韵律边界标签,利用从原始文本中学到的嵌入功能提高模型的预测性能[17-19]。

### 3.2 端到端的深度学习 TTS

上述为传统方法引入部分神经网络来代替原有组件的方法已经证明了深度学习在语音信息处理领域的惊人效果,但其中的神经网络和其他组件不是有机统一的关系,而是主要通过贪心算法分别训练每个独立的组件以提升整体模型的效果。由于各个组件的最优解结合后并不一定是全局的最优解,因此深度学习的效果并没有被完全发挥出来。

近两年,随着大规模数据集的出现和更科学高效的神经网络的提出,构建“端到端”的深度学习模型并使用梯度下降法求解整体的最优解逐渐成为可能。其中,WaveNet 及基于该方案的改进设计,由于可直接处理原始音频并输出语音波形而被视为端到端的语音合成模型;DeepVoice 家族与 Tacotron 系列中的初代模型仍需要音频合成组件的辅助,但在后续发展过程中逐步融合了较为成熟的声码器,就系统整体而言可实现端到端。此外,针对模型韵律结构、合成速度、稳定性等性能的改进研究也不断涌现。

#### 3.2.1 WaveNet 及改进方法

作为一维卷积的一种典型应用,基于 PixelCNN 的自回归显示生成模型 WaveNet[20]可直接处理原始音频波形,利用带洞因果卷积(dialated causal convolutions)高效地预测新样本的分布,如图 1 所示。尽管抽样过程、数据离散算法与之类似,但不同于基于卷积层的 WaveNet,SampleRNN[21]在底层循环层逐样本生成音频,并能在长时间跨度内捕获时间序列

中潜在的变化源，同时提高训练的存储效率。为解决WaveNet因自回归特性而产生的合成速度慢、难以应用于实际场景中的问题，WaveRNN[22]结合多种加速技巧，实现了模型简化、稀疏化，并利用高并行度生成语音算法高效运行。

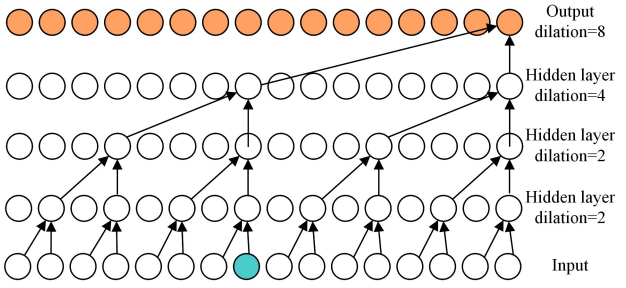


图1 WaveNet带洞因果卷积[20]

Fig. 1 WaveNet dialated causal convolutions[20]

该类方法主要面临的挑战在于：LSTM，RNN等模型虽然能够挖掘声音时序特征，但受模型本身结构和网络容量的影响，仍然存在“长期依赖”问题，下一步的改进方向或许可考虑添加注意力机制以促使模型关注更有意义的信息而屏蔽无效信息，并可尝试引入自然语言处理领域的BERT，GPT等超大规模神经网络，以提升网络容量和增强信息表达能力。

此外，一些方法尝试借助神经网络的不同结构设计，达到在保证一定的合成语音质量的同时，大幅提升计算速度的目的。Parallel WaveNet[23]利用概率密度蒸馏(probability density distillation)从WaveNet中训练出了一个并行前馈网络，该思路与GAN略有相似之处，即利用学生网络模拟教师网络生成的概率分布，在保持教师网络不变的同时不断训练学生网络，从而使之均得到归一化分布。尽管相比WaveNet，该模型可实现20倍的实时高保真语音合成，但教师模型的训练难度仍不可小觑。在重新设计基于流的WaveGlow[24]模型架构的基础之上，轻型声码器SqueezeWave[25]重新排列音频张量，并采用深度可分离卷积进行模型优化，大幅减少了因并行信息处理而产生的计算量，使模型所需算力缩小为WaveGlow的1/214。

上述研究进一步促进了深度学习应用于语音信息处理领域的发展，并且这些模型在下文介绍的典型模型中常被用作声码器，如表2所列。

表2 声码器对比

Table 2 Vocoder comparison

| 名称 | 时间 | 特点 |
|---|---|---|
| SqueezeWave[25] | 2020年 | 优化WaveGlow的结构和计算方法，大幅提高模型计算效率 |
| WaveGlow[24] | 2019年 | 无需自回归，使用似然度作为损失函数进行训练 |
| WaveRNN[22] | 2018年 | 利用一个单层RNN架构，拆分长序列为若干短序列并同时生成 |
| SampleRNN[21] | 2017年 | 基于循环层，在长时间跨度内捕获时间序列中潜在的变化源 |
| Parallel WaveNet[23] | 2017年 | 使用概率密度蒸馏法，训练新的并行前馈网络 |
| WaveNet[20] | 2016年 | 结合因果卷积和扩大卷积，高效预测样本分布 |

### 3.2.2 Deep Voice，Tacotron及其衍生方法

端到端的TTS模型中最具代表性的是Deep Voice[8,26-27]家族和Tacotron系列算法[9,28]，实际应用中常常利用该模型将文本转换为频谱，然后结合WaveNet等波形生成模型或Griffin-Lim算法，将频谱转换为原始波形并输出。

Deep Voice最早由百度AI研发并于2017年提出，由此开辟了一个TTS技术分支。该方法的主要特点是将参数合成步骤中的各个模块用神经网络替代，从而形成一个完整的TTS解决方案，详细来说：Deep Voice[8]和Deep Voice 2[26]保留了传统的TTS管道，通过神经网络实现了单独的音素边界分割模型、字素到音素转换模型、音素持续时间预测模型、基本频率预测模型和音频合成模型，二代在一代的基础上又添加了说话人向量，可实现多个说话人声音的合成。Deep Voice 3[27]则抛弃了上述模块化思路，引入基于注意力机制的seq2seq模型，利用注意力机制的带洞卷积，将学习表示解码为梅尔频谱，最后借助解码器的隐藏状态预测声码器参数，使体系结构更为紧凑。不断地改进和完善使得Deep Voice系列技术的效果一次次提升，并被广泛应用于实际的TTS系统。

Tacotron及其衍生算法是端到端TTS技术的另一大算法分支。在该系列算法诞生以前出现的Char2wav[29]，首先实现了从了端到端的语音合成，即利用读取器与单独的波形合成器实现文本到音频的转换。但Char2wav在在使用声码器之前仍需预测声码器参数，而文献[9]提出的端到端的Tacotron方案(见图2)利用给定的〈text，audio〉对，随机初始化从头开始训练模型，可直接预测原始谱图，最后通过Griffin-Lim实现了声谱到声波的合成。该方案同样由encoder和decoder组成，但为了提升模型的泛化能力，Tacotron在encoder部分所采用的方法与Char2wav不同，Tacotron在双向RNN前通过卷积层和残差连接组合的方式，利用该瓶颈层对输入进行预处理，可减少网络参数，提升收敛速度。

其后续版本Tacotron2[28]在此基础上提出用普通的LSTM和卷积层替换高层次特征提取模块CBHG，结合改进的WaveNet声码器，获得了媲美人声的合成效果。受到Tacotron算法的启发，文献[30-31]等进一步探索了注意力机制和seq2seq模型在语音合成领域的应用，采用CTC识别器、多个输入编码器进一步丰富了模型的特征挖掘能力，有效地降低了发声错误率，从而提高了生成能力。上述方法主要利用词嵌入、互信息等手段进行特征提取和特征选择，增加上下文覆盖范围来提取敏感特征，避免出现局部信息偏好，但互信息本身易受词边缘概率的影响，同时训练大量文本语料数据对模型的训练速度提出了新的要求。相比Deep Voice家族，Tacotron系列采用基于编码器-解码器的结构进一步提升了序列模型的网络表示能力，因此也能生成效果更逼真的合成语音。

表3对比了Deep Voice家族和Tacotron系列模型的各项性能指标，其中平均主观意见分(Mean Opinion Scores，MOS)带95%置信区间，是行业内一致认可的合成效果批判标准，但由于评判具有主观性且受不同数据集的影响，合成效果有所波动，故表中展示的MOS数值为现有测试下该模型表现出的最佳状态。可以看出，尽管Tacotron系列和Deep Voice家族的性能都在不断提升，但是Tacotron2的提升是以扩大网络规模和提升计算复杂度为代价的，而Deep

Voice 3 采用了全卷积架构，利用 GPU 的并行性缓解了计算复杂度激增的难题。

表 3 典型模型效果对比

Table 3 Comparison of the performance softy picalmodels

| 模型名称 | MOS | 每次迭代时间/ms | 迭代次数 |
|---|---|---|---|
| Tacotron(Griffin-Lim)[28] | 4.001±0.087 | — | — |
| Tacotron(WaveNet)[26] | 4.17±0.18 | 590 | $2\times10^6$ |
| Tacotron2(Mel+WaveNet)[28] | 4.526±0.016 | — | — |
| Deep Voice 2(WaveNet)[26] | 3.53±0.12 | 780 | $500\times10^5$ |
| Deep Voice 3(WaveNet)[27] | 3.78±0.30 | 60 | $500\times10^5$ |

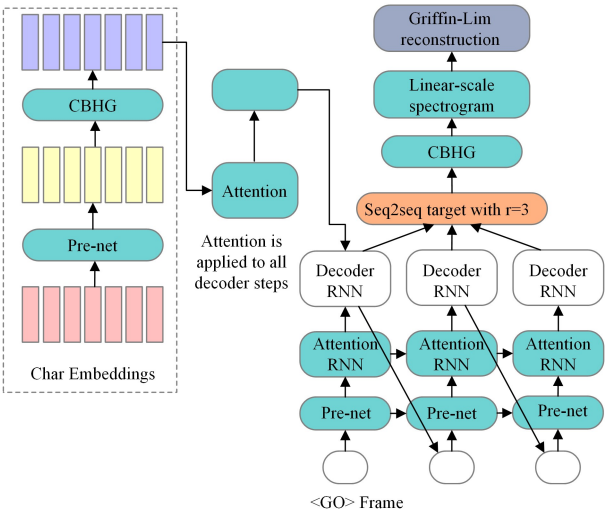


图 2 Tacotron 结构[9]

Fig. 2 Tacotron structure[9]

### 3.2.3 风格韵律改进方法

风格韵律改进是指通过韵律建模或语音风格控制，改变语音的重读、音调、节奏感、情感等因素，使之自然度更高且更具表现力。

Tacotron 等经典模型由于并未清晰地实现韵律建模，合成的语音相比真实人声显得生硬呆板。常见的改进措施包括：为 Tacotron 引入更丰富的特征以及增加能够控制全局或局部粒度特征的机制；利用 VAE 获得演绎形式的多种不同分布，并迁移至目标对象；使用带注意力的模型以关注核心特征等。

改进风格韵律的代表性方法如在 Tacotron 中引入全局样式标记(Global Style Token，GST)，用基于注意力的参考编码器传输全局样式属性，借助 token 不同权重的组合实现风格变化，在噪声条件下 MOS 值可达 4.0[32]。尽管 GST 可提供多种风格选择，但其韵律控制的效果却并不理想。因此，不同于全局嵌入的方式，文献[33]在嵌入网络中引入时间结构以控制局部特征，可分别在编码器端和解码器端进行韵律嵌入，实现了对合成语音细粒度的局部控制。

另一类利用 VAE 获得的解纠缠的隐变量表征语音特征，通过调整特征编码控制风格，简化了基于似然性的的自回归模型的概率学习过程。如文献[34]将 GST 模型作为基线模型，通过 VAE 学习隐变量作为语音风格的表示，输入 TTS 网络从而进行特定风格的语音合成；文献[35]相较前者应用正则化流 Householderer Flow 丰富了 VAE 的后验分布，同时显著降低 KL 散度，仅使用大约 1s 的目标风格表达就能合成高质量的语音。

考虑到现有的无监督模型在从真实样本中提取风格信息时会不可避免地产生内容信息泄漏，文献[36]创新性地提出了一种无监督的内容和风格分离方法，通过估计互信息并最小化重构损失，可减少内容泄漏，降低字错率(word error rate)，相比 GST 模型效果亦有提升。

由于上述模型本身无监督的特点，基于这些模型提取的风格特征可解释性较差，为此在 Tacotron2 的基础上，文献[37]通过信息驱动的条件变分自编码器(Conditional Variational Auto-Encoder，CVAE)提出了一种细粒度且可解释的韵律隐变量模型。

此外，参考图像风格迁移领域的相关技术，基于 GAN 的模型利用其对抗训练机制匹配联合分布，并在训练过程中引入额外的重建损失和样式损失，实现了内容与风格的可分离和可控性[38]。尽管上述方法生成样本的自然度有较大提升，但高质量的合成需要满足文本和音频对齐的要求，且受限于图像和语音数据的差异，实际应用时效果仍略有瑕疵。Tacotron 模型与相关改进方法汇总如表 4 所列。

总体来说，在实现了可控风格、韵律变换的基础上，未来的模型发展将着眼于平衡样本标注的代价与无监督学习模式带来的负作用，以及根据不同的文本内容与背景环境，自动选择匹配合适的风格、韵律，以加强其落地应用。

表 4 Tacotron 及改进方法汇总

Table 4 Summary of Tacotron and improvement methods

| 模型名称/方法名称 | 时间 | 特点 |
|---|---|---|
| Tacotron[9] | 2017 年 | 给定的〈text，audio〉对，直接输出梅尔频谱，通过 Griffin-Lim 生成波形 |
| Tacotron2[28] | 2018 年 | 简化特征提取，使用一个 5 层卷积神经网络精调梅尔频谱，改变波形生成方式 |
| GSTs[32] | 2018 年 | 在 Tacotron 中引入 GST，用无监督方式控制和迁移风格 |
| VAE[34] | 2018 年 | 利用 VAE 学习语音风格的潜在表现，并结合 Tacotron2 |
| Auxiliary CTC Recognizer[30] | 2019 年 | 改进 Tacotron1 和 Tacotron2，利用辅助 CTC 识别器最大化文本和预测的声学特征之间的互信息，解决单词丢失或重复 |
| Multiple Input Encoder[31] | 2019 年 | 利用多个输入编码器获取音素序列、预训练的词嵌入和语法结构，增强特征 |
| CVAE[37] | 2020 年 | 改进 Tacotron2，使用自回归的 CVAE 对潜在维度施加分层条件，提出一种细粒度的韵律隐变量模型 |

#### 3.2.4 快速生成模型

受限于 Tacotron 等语音合成方法中 RNN 的运算速度以及自回归模型中误差累积带来的鲁棒性问题，为提高模型的运算速度，增强模型的生成能力，优化模型的实现效果，许多新方法不断涌现。

文献[39]提出利用深度卷积 TTS 替代 RNN，并结合可快速训练的注意力机制，可在 15 h 内实现 20 万次迭代。上述模型延续从文本到梅尔频谱图，再由梅尔频谱图转换到音频的策略，与之不同的是文献[40-41]提出并增强的全卷积文本到波形神经结构 ClariNet，可直接将文本转换为原始的音频波形。该方法通过最大似然估计训练 WaveNet，对比 Parallel WaveNet，其优势在于概率密度蒸馏采用闭式计算，并借助最小化峰值输出正则 KL 散度，增强了训练过程的稳定性，模型整体的训练速度比基于 RNN 的模型提升了 10 倍以上。

为弥补 WaveNet，ClariNet 因自回归或递归分量产生的运算速度方面的不足，FastSpeech 并行生成了梅尔谱图，相比自回归 Transformer 模型提速 269.4 倍，并通过长度调节器灵活控制语音速度、韵律，使合成的语音兼具鲁棒性与可控性[42]；GAN-TTS 利用一个由卷积神经网络（Convolutional Neural Networks，CNN）构成的高效前馈生成器生成语音，并结合多个鉴别器不断提升语音的自然度[43]。

随着研究的深入，人们发现对端到端模型的原始输入数据进行一定的特征预处理操作更有利于提升模型的性能，因此很多研究者尝试在原始数据流入模型前，新增一些特征处理模块，尽管看起来这种思路与传统的非端到端的方法有些相似，但是这些特征处理模块的规则往往更简单，也更有利于后续自动化处理的进行，因此仍然可被视为端到端的发展分支。如 BOFFIN TTS[44]利用字素到音素模块对输入的文本进行预处理，之后不再手工选取超参数，而是利用贝叶斯优化方法对现有 TTS 模型微调和对超参数进行搜索，实现了小样本学习，并且可利用不到 10 min 的语音样本合成新的语音。

## 4 语音转换

通过 TTS 得到的语音虽然很大程度上接近于自然人声，但为了丰富合成语音的可变性，还需要进一步提取并转换特征，然后重新合成语音，将自然的人声转变为预设目标对象的声音。

语音转换研究早期常利用高斯混合模型（Gaussian Mixture Model，GMM），通过谐波加噪声模型（harmonic plus noise model）、STRAIGHT 等方法提取特征，调整频谱包络与基频等关键特征，从而实现变换。但基于统计的变换方式在实现过程中易出现过度平滑与过拟合问题，为此文献[45-46]利用条件受限玻尔兹曼机（conditional restricted boltzmann machine）学习原始语音和目标语音向量之间详细的线性和非线性关系，比主流的联合密度高斯混合模型（joint density gaussian mixture model）表现更佳。结合时间分解（temporal decomposition）的局部线性变换（local linear transformation）提出了一种新的转换方法，通过寻找合成语音和原始语音之间的映射关系，提高基于 HMM 合成的语音质量[47]。但上述转换方法受限于模型能力，对于语音中突然的急促、音调提高等不规则现象无法进行动态处理。

除去模型本身的能力不足，这类方法也有共性的问题存在，即训练模型需要事先指定源说话人和目标说话人的身份，语音转换灵活性有所欠缺；此外，训练依赖的语料数据需要满足帧级对齐，而相关数据库建设缓慢，收集语料难度大，帧级对齐要求高，每更换一个目标都需要重新训练，这使得该监督学习方法的可行性不高。为解决上述问题，基于非并行数据的非监督学习方法显示出其优越性，根据实现手段的不同其可大致分为 3 类：1）利用 GAN 通过不断地生成、判别，实现高质量语音转换；2）借助自编码器将语音内容与表征说话人身份特征的音色信息分开；3）基于与说话人无关的音素识别器获得语音后验图（Phonetic PosteriorGram，PPG），再通过非线性映射函数进行说话人声学特征转换。

### 4.1 基于生成对抗网络的 VC

GAN 模型已经被证明在图像生成、风格转换和图像编辑领域具有优异的性能，除了可以用于前文所述的语音合成，还可用于生成类似目标人物音色的声音。

Kaneko 等[48]尝试利用 GAN 模型来处理语音数据并提出 CycleGAN-VC 模型，在 GAN 网络中加入带有门控单元的 CNN 单元，其原理如式(1)所示：

$$H_{l+1}=(H_l * W_l+b_l)\otimes\sigma(H_l * V_l+c_l) \tag{1}$$

其中，$H_l$ 是第 $l$ 层的输出结果，$H_{l+1}$ 是第 $l+1$ 层的输出，$W_l$，$b_l$，$V_l$，$c_l$ 是神经网络参数，$\otimes$ 是“按元素相乘”，$\sigma$ 表示常用的 sigmoid 激活函数。在此基础上，Kaneko 等还采用了对抗损失、循环一致性损失和身份映射损失以减少转换语音的过平滑，使得模型能够更好地提取语音数据的顺序和分层特征，最终取得了显著的成效。其后的升级版本 CycleGAN-VC2[49]在上一代的基础上添加了两步对抗性损失，其原理如图 3 所示。

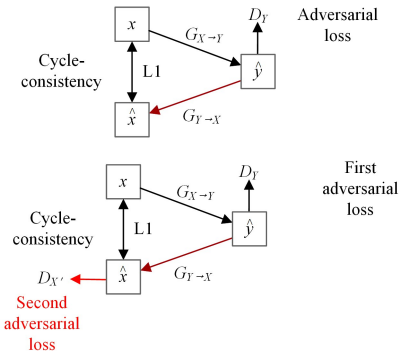


图 3 一步对抗性损失与两步对抗性损失的计算流程[49]

Fig. 3 Flow charts of one-step adversarial loss and two-step adversarial losses[49]

最终整体的损失函数形式如式(2)所示：

$$Loss_{\text{full}}=Loss_{\text{adv}}(G_{X\to Y},D_Y)+Loss_{\text{adv}}(G_{Y\to X},D_X)+\lambda_{\text{cyc}}Loss_{\text{cyc}}(G_{X\to Y},G_{Y\to X})+\lambda_{\text{id}}Loss_{\text{id}}(G_{X\to Y},G_{Y\to X}) \tag{2}$$

其中，$\lambda_{\text{cyc}}$ 和 $\lambda_{\text{id}}$ 用于平衡损失权重。

Isola 等[50]对生成器模型结构做了微调并采用 Patch GAN 作为判别器，使用全卷积网络代替原理的普通卷积层，能够捕获更多的音频特征，进而可以满足生成更高清晰度、高

逼真细节的语音转换任务要求，进一步提升了模型的性能，使得模型在不依赖于后滤波器或声码器的情况下能出色地完成语音合成和性别内语音转换任务。

StarGAN-VC 和 StarGAN-VC 2 是 Kameoka 等[51-52]在 StarGAN 的基础上引入新的损失函数和网络结构而提出的面向语音转换的 GAN 模型。他们认为传统的对抗损失虽然有助于提升模型判别器的训练效果，但是无法保证转换后的语音能否保留输入音频中的必要信息，因此通过加入一致性损失来改善这一问题。损失函数如式(3)所示：

$$Loss_{cyc}=E_{(x,c)\sim P(x,c),c'\sim P(c')}[\|x-G(G(x,c'),c)\|1] \quad (3)$$

这一创新也启发了上述 CycleGAN-VC 模型的改进。此外，StarGAN-VC 2 还加入了 GLU 激活层等新特性以实现语音转换的一种语音转换模型，不仅在图像风格迁移任务中表现优异，同时也能够在语音领域实现非平行数据集下的“多对多”音色转换。相比 CycleGAN-VC，它在预测时无需任何关于输入语音属性的信息，且具备更快的推理速度，但总体来说 GAN 模型的使用难题在于训练不稳定，导致生成的音频存在失真，包含噪声等异常特征；此外，尽管其突破了早期模型依赖平行语料数据的限制，但其语音转换目标仅限于训练样本中的人。

表 5 VC 中 GAN 的应用汇总

Table 5 Summary of GAN application in VC

| 名称 | 时间 | 特点 |
|---|---|---|
| CycleGAN-VC[48] | 2017 年 | 在 GAN 中加入带有门控单元的 CNN 结构，结合使用多种损失以减少过平滑 |
| CycleGAN-VC 2[49] | 2019 年 | 新增两步对抗性损失，微调生成器模型结构，采用 Patch GAN 作为判别器 |
| StarGAN-VC[51] | 2018 年 | 在 StarGAN 基础上，新增一致性损失，独立分类器，在生成器和判别中都连接说话人特征向量 |
| StarGAN-VC 2[52] | 2019 年 | 新增源和目标的条件对抗损失，引入基于调制的条件方法以特定域的方式转换声学特征的调制 |

### 4.2 基于自编码器的 VC

为进一步解决目标说话人身份受限问题，自编码器的特殊结构为语音转换提供了新思路，如何在编码器将数据转变为隐变量后，将其中的内容与身份两部分信息分隔开成为研究的焦点。事实上，在 TTS 的语音风格转换部分同样涉及对隐变量的处理，但不同的是，TTS 部分的重点在于精细化的隐变量表达与控制，而 VC 部分更强调隐变量解纠缠中身份信息的提取与转换。

文献[53]基于 VAE 的语音转换模型显式地使用一个表示说话人身份信息的属性向量作为解码器部分的输入，并通过该属性向量对语音的特征进行编辑以实现语音转换。相比 GAN 模型，VAE 的隐变量具有易操作的优势。尽管上述方案也实现了解纠缠，但采用 one-hot 编码方式而不通过目标函数去促使编码器舍弃身份信息，使模型不具备提取说话人身份特征嵌入的能力，也无法实现未知身份说话人的语音转换。

与前者不同，文献[54]的方法则更清晰地进行了解纠缠，如图 4 所示，其借助两个不同的编码器 $E_s$ 和 $E_c$ 分别提取说话人的音色信息与音频的内容信息，在提取内容信息的过程中使用实例正则化(Instance Normalization，IN)去除音色信息，随后的解码模块中再通过自适应实例正则化(Adaptive Instance Normalization，AdaIN)重新添加新的音色信息。这种归一化算法借鉴了视觉任务中的风格迁移，在加速模型收敛的同时能实现灵活多样的风格转变。

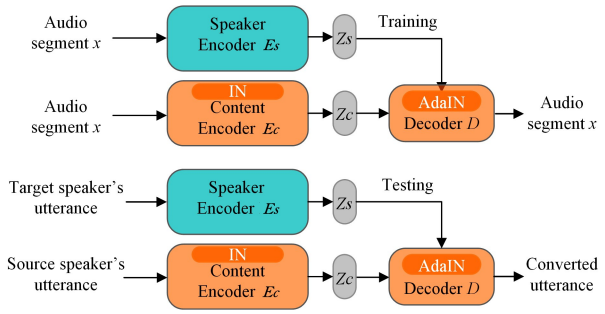


图 4 单样本语音转换模型的结构[54]

Fig. 4 Architecture of one-shot voice conversion model[54]

AutoVC[55]通过数学方式证明了选择适当的自编码器中隐变量的维数，使之刚好容纳内容信息而容不下身份信息，可实现解纠缠，且适合的维数可通过实践比较得出。但这种通过信息约束瓶颈实现解纠缠的方式易导致大量韵律信息 F0 泄漏且无法控制转换后的韵律。因此，文献[56]利用从源说话人音频中提取的基频 F0 调整解码器，使内容、F0 和说话人身份同时分离，避免泄漏的同时增强了可控性，显著提高了语音质量。该方案的出现在一定程度上证明了传统声学特征的研究并非如一些人所说的那般过时，而是能够通过与深度学习方法相结合发挥出更大的作用，为语音转换提供一条新的途径。

### 4.3 基于语音后验图的 VC

同样作为不依赖于并行语料的非监督方案，基于语音后验图的语音转换另辟蹊径，基线架构首先训练出一个与音色等说话人身份特征无关的音素识别器，然后将得到的 PPG 通过映射嵌入目标说话人的身份信息从而完成转换，如图 5 所示。

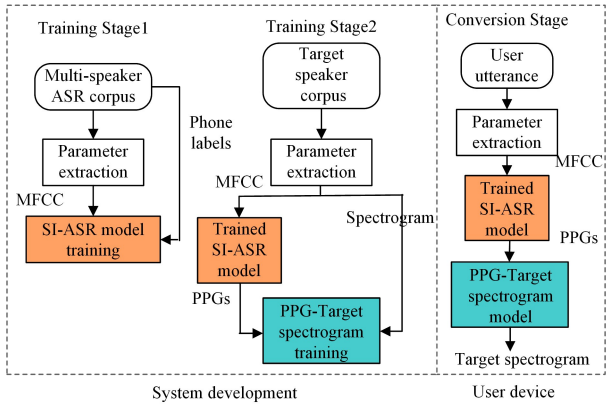


图 5 基线非并行语音转换体系的结构[57]

Fig. 5 Architecture of baseline non-parallel voice conversion[57]

然而，这种架构的语音转换模型在训练阶段与转换阶段都需要同时运行两个级联网络，使得模型的实时性较差。为缩小模型、提升效率，文献[57]提出一种单一网络的从原始语音到目标语音的方案，通过该方案完成了多说话人梅尔频率

倒谱系数(mel-frequency cepstrum coefficient)到目标线性谱图之间的映射,即在保持训练阶段传统方式的基础上,转换阶段不再使用 PPG,最终转换时间与网络参数数量分别减少了 44.5%和 41.9%。

不同于基线架构,以 TTS Skins[58]为代表的方法利用自动语音识别模型(Automatic Speech Recognition, ASR)可将音频转换为音素的特点,在编码器 Wav2Letter 中加入预先训练的 ASR 网络以去除音频文件中的音色信息,得到 PPG,然后用新增的音素信息通过 WaveNet 解码器合成目标对象的音频。上述方案中 WaveNet 的使用使得模型速度堪忧,因此在设计方案中,Rebryk 等[59]使用参数量仅为前者 1/20 的轻量级的编码器 QuartzNet-5x5 与非自回归的 WaveGlow 作为解码器,在缩小模型、提高计算效率的同时,可进一步将其推广到不可见的说话人。但这种语音转换思路的实现效果严重依赖于 ASR 的准确率。因此尽管 PPG 能体现出面对噪声的鲁棒性,但综合基于该思路的模型表现来看,仍有较大的提升空间。

## 5 应用与风险

语音信息处理技术在深度学习的推动下发展增速,其中较完备的技术成果已在有声读物、智能硬件、泛娱乐等领域落地应用。具体而言,基于 TTS 的语音合成结合语音转换技术实现的 AI 变声,能在进行自动化文本朗读的同时选择不同音色的声音,使朗读摆脱生硬机械而更显自然亲切。这种媲美人声的听觉体验,可替代传统有声内容真人朗读录制的创作模式,减少人力、物力、资源的消耗,从而有效降低创作门槛。此外,将该技术应用于诸如助力盲人“阅读”的公益事业中也能发挥良好作用。考虑到合成音频的情感表达与真实人声相比仍有距离,在情感共鸣需求较高的场景下该技术带来的听众体验稍显不足,但对于儿童教育、导航软件、旅游解说等场景,语音合成与转换产生的声音已经能够较好地满足应用要求。

作为一种依靠数据支持的技术,语音合成与转换本身具有可定制特性。因此,针对不同服务对象所提出的不同声音场景、音质、类别需求等,进行个性化的模型训练,在人机交互愈发频繁、语音交互需求日益提升的今天,其应用前景显得尤为广阔。语音表征学习、迁移,结合个性变量植入,使智能硬件、机器人在个性化发展道路上产生了长足的进步,定制的使用体验也增强了使用的便捷度。此外,融合全息人物展示等虚拟技术,语音合成与转换也能灵活运用于泛娱乐场景,在贯通游戏、动漫、文学等文化创意领域的互动娱乐新生态中发挥其独有的价值。

技术落地应用的背后不仅仅是商业化的成功,也为社会带来了普惠的价值理念,越来越多的应用场景的出现为技术的价值提供了确定性并带来了更大的想象空间。尽管技术本身无不良属性,但在其更新迭代带来升级体验的同时,也容易被不当利用,如伪造声音实施诈骗、传播舆论等,故而在发展过程中也需要警惕潜藏的伦理道德风险和安全隐患问题。

当前利用深度伪造技术进行视频换脸已经达到了以假乱真的效果,产生的公共安全风险也逐步暴露,而利用语音信息处理技术能实现一定程度上的语音伪造[60],二者结合将使风险再度升级。此外,以音频的方式对文字作品进行演绎,如何把控版权和品质将成为音频上载平台需明确的问题。与日新月异的技术发展相伴而生的是潜在的技术漏洞与技术滥用带来的诸多不利,因此综合研究分析语音合成和语音转换的技术手段,不仅对满足产业增长需求、促进性能优化升级具有重要意义,也为应对这些风险与挑战提供了有力的支撑。

**结束语** 本文从 TTS 和 VC 两步出发,对语音信息处理中关键技术的研究进展进行了归纳总结,介绍了各种语音合成模型、转化模型以及相关优化方法。语音信息处理技术发展至今,尽管在模型简化方面取得了一定进步,但是在降低训练成本、提升训练速度、优化生成质量方面,仍有许多问题亟待完善和解决,这也是语音合成及转换领域的下一步关注焦点。

(1)少样本学习(few-shot learning):现有语音信息处理技术高度依赖于大规模的音频数据库,而现实中获取的目标音频往往只有短短数句,这为定向生成目标声音的工作带来困难,如何能够在仅获少量目标语音的前提下生成高质量的语音样本仍将是充满挑战挑的研究。

(2)模型压缩:大量使用 GAN 和 VAE 等复杂网络,使得在训练模型时不仅需要耗费更多的算力、更长的时间,也导致训练好的模型无法移植到手机、个人电脑等低算力终端设备上使用,通用性不强。如何在保证模型效果的前提下,减小模型规模、删除冗余参数是语音信息处理技术走向实用必须克服的问题。

(3)伪造检测:语音信息处理技术生成的样本不仅能够欺骗人类的听觉,甚至也能绕过机器学习检测系统的识别。对该技术的恶意利用如模仿领导人讲话、攻击语音身份验证系统等行为,对人工智能安全和公共安全造成了威胁。而相关的检测技术却仍处于起步阶段,不仅准确率难以达到要求,而且往往只能针对一种或寥寥数种伪造方法进行识别,无论是在通用性还是可靠性方面都还有很长的路要走。

(4)数据集:当前语音数据集音源多来自新闻、演讲、电子读物等,缺乏日常生活语音,数据集多样性有所欠缺,且针对监督学习方案要求的并行数据,当前语音数据集在数量与质量上都无法满足需求,因此模型训练效果受限。扩充语音数据集类型、保证数据集质量,是提升合成语音自然度、增强检测模型泛化能力的必然要求。

(5)法律规范:针对语音合成和语音转换等语音信息处理技术,相应的规范管理条例还未出台,利用该技术发起的攻击将会对现有法律造成冲击。针对此类技术的合理使用范围和违法违规认定条件,需要法律学界在不影响正常技术进步的前提下,加以思考。

## 参 考 文 献

[1] RealTalk[OL]. https://medium.com/dessa-news/real-talk-speechsynthesis-5dd0897eef7f.

[2] MelNet[OL]. https://sjvasquez.github.io/blog/melnet/.

[3] MBIUS B, SPROAT R, SANTEN J, et al. The bell labs German text-to-speech system: an overview[C]// Fifth European Confe-

rence on Speech Communication and Technology. 1997:22-25.

[4] WU Y J,WANG R H. Minimum Generation Error Training for HMM-Based Speech Synthesis[C]// International Conference on Acoustics,Speech,and Signal Processing. IEEE,2006:89-92.

[5] ZEN H,BRAUNSCHWEILER N. Context-dependent additive log F0 model for HMM-based speech synthesis[C]// Conference of the International Speech Communication Association. 2009:2091-2094.

[6] TODA T,SARUWATARI H,SHIKANO K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum[C]// International Conference on Acoustics,Speech,and Signal Processing. IEEE,2001:841-844.

[7] AIHARA R,TAKASHIMA R,TAKIGUCHI T,et al. GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features[J]. American Journal of Signal Processing,2012,2(5):134-138.

[8] ARIK S O,CHRZANOWSKI M,COATES A,et al. Deep Voice:Real-time Neural Text-to-Speech[J]. arXiv:1702.07825,2017.

[9] WANG Y,SKERRYRYAN R J,STANTON D,et al. Tacotron:Towards End-to-End Speech Synthesis[C]// Conference of the International Speech Communication Association. 2017:4006-4010.

[10] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems,2014,3:2672-2680.

[11] LEMMETTY S. Review of Speech Synthesis Technology[D]. Helsinki University of Technology,1999.

[12] ZE H,SENIOR A W,SCHUSTER M,et al. Statistical parametric speech synthesis using deep neural networks[C]// International Conference on Acoustics,Speech,and Signal Processing. IEEE,2013:7962-7966.

[13] LU H,SIMON K,OLIVER W. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis[C]// The 8th ISCA Speech Synthesis Workshop. 2013:261-265.

[14] WU Z,TAKAKI S,YAMAGISHI J. Deep Denoising Auto-encoder for Statistical Speech Synthesis[J]. arXiv:1506.05268,2015.

[15] KANG S,QIAN X,MENG H. Multi-distribution deep belief network for speech synthesis[C]// International Conference on Acoustics,Speech and Signal Processing. IEEE,2013:8012-8016.

[16] YIN X,LING Z H,HU Y J. Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis[J]. IEEE Transactions on Audio,Speech,and Language Processing,2013,21(10):2129-2139.

[17] FERNANDEZ R,RENDEL A,RAMABHADRAN B,et al. Prosody Contour Prediction with Long Short-Term Memory,Bi-Directional,Deep Recurrent Neural Networks[C]// Conference of the International Speech Communication Association. 2014:2268-2272.

[18] FAN Y,QIAN Y,XIE F,et al. TTS synthesis with bidirectional LSTM based recurrent neural networks[C]// Conference of the International Speech Communication Association. 2014:1964-1968.

[19] DING C,XIE L,YAN J,et al. Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features[C]// 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE,2015:98-102.

[20] OORD A V D,DIELEMAN S,ZEN H,et al. WaveNet:A Generative Model for Raw Audio[J]. arXiv:1609.03499,2016.

[21] MEHRI S,KUMAR K,GULRAJANI I,et al. SampleRNN:An Unconditional End-to-End Neural Audio Generation Model[J]. arXiv:1612.07837,2016.

[22] KALCHBRENNER N,ELSEN E,SIMONYAN K,et al. Efficient neural audio synthesis[J]. arXiv:1802.08435,2018.

[23] OORD A V D,LI Y,BABUSCHKIN I,et al. Parallel WaveNet:Fast High-Fidelity Speech Synthesis[J]. arXiv:1711.10433,2017.

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

[25] ZHAI B,GAO T,XUE F,et al. SqueezeWave:Extremely Lightweight Vocoders for On-device Speech Synthesis[J]. arXiv:2001.05685,2020.

[26] ARIK S O,DIAMOS G,GIBIANSKY A,et al. Deep Voice 2:Multi-Speaker Neural Text-to-Speech[C]// Advances in Neural Information Processing Systems. Curran Associates,2017:2962-2970.

[27] PING W,PENG K,GIBIANSKY A,et al. Deep Voice 3:Scaling Text-to-Speech with Convolutional Sequence Learning[J]. arXiv:1710.07654,2017.

[28] SHEN J,PANG R,WEISS R,et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions[C]// International Conference on Acoustics,Speech,and Signal Processing. IEEE,2018:4779-4783.

[29] SOTELO J,MEHRI S,KUMAR K,et al. Char2Wav:End-to-End Speech Synthesis. ICLR 2017 Workshop Submission[EB/OL]. (2017-04-16)[2020-05-26]. https://openreview.net/forum?id=B1VWyySKx.

[30] LIU P,WU X,KANG S,et al. Maximizing Mutual Information for Tacotron[J]. arXiv:1909.01145,2019.

[31] MING H,HE L,GUO H,et al. Feature reinforcement with word embedding and parsing information in neural TTS[J]. arXiv:1901.00707,2019.

[32] WANG Y,STANTON D,ZHANG Y,et al. Style Tokens:Unsupervised Style Modeling,Control and Transfer in End-to-End Speech Synthesis[C]// Proceedings of the 35th International Conference on Machine Learning. PMLR,2018:5180-5189.

[33] LEE Y,KIM T. Robust and Fine-grained Prosody Control of End-to-end Speech Synthesis[C]// International Conference on Acoustics,Speech and Signal Processing. IEEE,2019:5911-5915.

[34] ZHANG Y,PAN S,HE L,et al. Learning Latent Representa-

tions for Style Control and Transfer in End-to-end Speech Synthesis[C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2019: 6945-6949.
[35] AGGARWAL V, COTESCU M, PRATEEK N, et al. Using VAEs and Normalizing Flows for One-shot Text-To-Speech Synthesis of Expressive Speech[J]. arXiv: 1911.12760, 2019.
[36] HU T Y, SHRIVASTAVA A, TUZEL O, et al. Unsupervised Style and Content Separation by Minimizing Mutual Information for Speech Synthesis[C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020: 3267-3271.
[37] SUN G, ZHANG Y, WEISS R J, et al. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis [C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020: 6264-6268.
[38] MA S, MCDUFF D, SONG Y, et al. Neural TTS Stylization with Adversarial and Collaborative Games. ICLR 2019 Conference Blind Submission[EB/OL]. (2019-02-23)[2020-05-26]. https://openreview.net/pdf?id=ByzcS3AcYX.
[39] TACHIBANA H, UENOYAMA K, AIHARA S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention[C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2018: 4784-4788.
[40] PING W, PENG K, CHEN J, et al. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech[J]. arXiv: 1807.07281, 2018.
[41] PARK J, ZHAO K, PENG K, et al. Multi-Speaker End-to-End Speech Synthesis[J]. arXiv: 1907.04462, 2019.
[42] REN Y, RUAN Y, TAN X, et al. FastSpeech: Fast, Robust and Controllable Text to Speech[C]// Advances in Neural Information Processing Systems. 2019: 3171-3180.
[43] BINKOWSKI M, DONAHUE J, DIELEMAN S, et al. High Fidelity Speech Synthesis with Adversarial Networks[J]. arXiv: 1909.11646, 2019.
[44] MOSS H B, AGGARWAL V, PRATEEK N, et al. BOFFIN TTS: Few-Shot Speaker Adaptation by Bayesian Optimization [C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020: 7639-7643.
[45] WU Z, CHNG E S, LI H, et al. Conditional restricted Boltzmann machine for voice conversion[C]// International Conference on Signal and Information Processing. IEEE, 2013: 104-108.
[46] NAKASHIKA T, TAKIGUCHI T, ARIKI Y. Voice conversion in time-invariant speaker-independent space[C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014: 7889-7893.
[47] JIAO Y, XIE X, NA X, et al. Improving voice quality of HMM-based speech synthesis using voice conversion method[C]// International Conference on Acoustics Speech and Signal Processing. IEEE, 2014: 7914-7918.
[48] KANEKO T, KAMEOKA H. Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks[J]. arXiv: 1711.11293, 2017.
[49] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion [C]// International Conference on Acoustics Speech and Signal Processing. IEEE, 2019: 6820-6824.
[50] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1125-1134.
[51] KAMEOKA H, KANEKO T, TANAKA K, et al. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks[C]// Spoken Language Technology Workshop. IEEE, 2018: 266-273.
[52] KANEKO T, KAMEOKA H, TANAKA K, et al. StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion[C]// Conference of the International Speech Communication Association. 2019: 679-683.
[53] HSU C C, HWANG H T, WU Y C, et al. Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder [C]// Asia Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, 2016: 1-6.
[54] CHOU J C, LEE H Y. One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization[C]// Conference of the International Speech Communication Association. 2019: 664-668.
[55] QIAN K, ZHANG Y, CHANG S, et al. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss[C]// Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 5210-5219.
[56] QIAN K, JIN Z, HASEGAWA-JOHNSON M, et al. F0-consistent Many-to-many Non-parallel Voice Conversion via Conditional Autoencoder[C]// International Conference on Acoustics, Speech and Signal Processing. IEEE, 2020: 6284-6288.
[57] JUNG S, SUH Y, CHOI Y, et al. Non-parallel Voice Conversion Based on Source-to-target Direct Mapping[J]. arXiv: 2006.06937, 2020.
[58] POLYAK A, WOLF L, TAIGMAN Y. TTS Skins: Speaker Conversion via ASR[J]. arXiv: 1904.08983, 2019.
[59] REBRYK Y, BELIAEV S. ConVoice: Real-Time Zero-Shot Voice Style Transfer with Convolutional Network[J]. arXiv: 2005.07815, 2020.
[60] TAO J H, FU R B, YI J Y, et al. Development and Challenge of Speech Forgery and Detection[J]. Journal of Cyber Security, 2020, 5(2): 28-38.

**PAN Xiao-qin**, born in 1997, postgraduate. Her main research interests include cyber security and artificial intelligence.

**LU Tian-liang**, born in 1985, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include cyber security and artificial intelligence.