设计研究与应用

语音合成技术综述及研究现状

魏伟华
（江西财经大学网络信息管理中心，江西南昌 330013）

摘　要：近年来，AI 技术发展越来越迅速，人机之间的交互也越来越频繁。语音合成技术就是人机交互的重要一环。语音合成技术又被称为文语转换 (TTS) 技术，就是可以将文字信息转化为流畅标准的语音。语音合成技术可以改善人机交互困难的情景，尤其是对有身体障碍，只能通过语音来交流的特殊人群，它可以让人类和计算机的交流更加方便快捷。本文首先介绍了语音合成技术发展历史和系统原理，然后叙述了语音合成技术的关键技术及最近主流模型，接着对 CNKI 和 ACM 数据库中近一年的最新论文进行了分析介绍，最后指明了当前语音合成技术未来大的发展方向。

关键词：语音合成；TTS；综述
中图分类号：TN912.33　　**文献标识码：**A　　**DOI：**10.3969/j.issn.1003-6970.2020.12.048
本文著录格式：魏伟华 . 语音合成技术综述及研究现状 [J]. 软件 ,2020,41(12):214-217

Overview and Research Status of Speech Synthesis Technology

WEI Weihua
(Network & Information Administration Center, JiangXi University of Finance and Economics, Nanchang Jiangxi 330013)

【Abstract】：In recent years, AI technology has developed more and more rapidly, and the interaction between humans and machines has become more and more frequent. Speech synthesis technology is an important part of human-computer interaction. Speech synthesis technology is also called text-to-speech (TTS) technology, which can convert text information into smooth and standard speech. Speech synthesis technology can improve scenarios where human-computer interaction is difficult, especially for special groups of people who have physical disabilities and can only communicate through voice. It can make communication between humans and computers more convenient and faster. This article first introduces the development history and system principles of speech synthesis technology, then narrates the key technologies of speech synthesis technology and recent mainstream models, and then analyzes and introduces the latest papers in CNKI and ACM databases for nearly a year, and finally points out the current speech The future direction of synthesis technology.

【Key words】：speech synthesis;TTS;review

0 引言

语音是人类最常用的沟通方式。语音技术包括二个分支，一是语音合成，另一是语音识别，一个是将文本信息转换为对应的语音，而另一个则将语音转换成人们能够理解的文本、情感等信息载体。从技术实现来分析，语音识别前期入门较快，但后期由于语音识别受口音、语种、环境等多种因素的影响，研究过程难度大大增加。语音合成前期对语音库的要求则相对较高，需要对采集标准的语音，然后标注，需要耗费大量人力物力，对语言语音学的基础理论知识也提出较高要求。本文主要针对语音合成技术，使读者可以深入理解语音合成，掌握 TTS 发展的脉络，为后期选题工作的正确开展提供一个方向性的指引。

语音合成技术大概可以归纳为 6 个阶段：一是起源阶段，时间可以追溯到 18 到 19 世纪，人们使用机械装置来模拟合成人类发声；二是电子合成器阶段，从 20 世纪初期开始，最具代表性的是贝尔实验室的 Dudley 发明的“VODER”的电子发声器；三是共振峰合成器

作者简介：魏伟华（1977—），男，工程师，研究方向：语音合成与识别、人工智能。

阶段，在集成电路技术的发展的推动下，20 世纪 80 年代 KLATT 在 1980 年发布的串 / 并联混合共振峰合成器；四是单元选择拼接合成阶段，时间为 20 世纪 80、90 年代；五是基于 HMM 的参数合成阶段，出现于 20 世纪末期；六是基于深度学习的语音合成阶段，随着人工智能技术快速发展，DNN/CNN/RNN、Tacotron 等各种神经网络模型用来语音合成的训练，模拟人声也更加自然成熟。

1 语音合成系统的原理

语音合成功能模块包括文本分析、韵律建模和语音合成三个部分。最后一个合成模块是 TTS 系统中最基本、最重要，前面二个模块可以看成语音预处理过程。语音合成的主要功能归纳为三个步骤，第一个步骤是根据韵律建模的结果，从原始语音库中取出对应的语音基元；第二步，利用特定的语音合成技术对语音基元进行韵律调整和修改；最后一步是合成出期望的语音。

语音合成方法按设计的主要思想，可以划分为规则驱动方法和数据驱动两类，前者的主要思想是根据人类发音物理过程制定规则来模拟重现发音过程，后者则是利用语音库中的数据，通过统计建模的方法来完成语音合成，所以对语音语料库的质量、最小单元和规模有更多的依赖。语音合成系统中的关键技术包括：共振峰合成、波形拼接、谐波加噪声模型和神经网络及深度神经网络模型。它们在合成过程中各有优缺点，需要有机地结合，取长补短，用一种技术的优点克服另一种技术不足之处。

2 语音合成系统的关键技术

2.1 共振峰合成

共振峰合成技术的基本原理如下：

不同人音色各异，其语音具有不同的共振峰模式，可以抽取每个共振峰频率及其带宽作为参数，这些参数可以构成共振峰滤波器。再通过若干个共振峰滤波器组合来模拟声道的传输特性，即频率响应；随后对激励源发出的信号进行调制，再经过辐射模型，最后合成语音。

基于共振峰的理论包含三种实用模型，分别是级联型共振峰模型、并联型共振峰模型和混合型共振峰模型。级联型共振峰模型主要用于绝大部分元音的合成，并联型共振峰模型主要针对鼻化元音等非一般元音以及大部分辅音，混合型共振峰模型是前二者的结合。

2.2 波形拼接

基于波形拼接的合成技术前期准备一般分为二个步骤，一是将语音单元切分成适合的合成单元，二是利用这些切分好的合成单元构建一个语音库。在合成阶段分为三个步骤，先选出合成单元，然后从语音库中提取出相应合成单元，最后将提取的单元按照韵律要求进行时长、基频的变换，最后采用重叠相加的方法重新输出合成语音。基于波形拼接的语音合成技术不需要从原始的语音中提取语音参数，而是将原始的语音信号直接存储，从而存储单元的要求要高于共振峰合成，在韵律的调节方面也要差一点。但是由于所采用的合成单元为原始语音文件，合成出来的语音清晰度要优于共振峰合成。基于波形拼接的语音合成方法包括 TD—PS0LA、FD-PSOLA、LP—PSOLA 等。

2.3 谐波加噪声模型

谐波加噪声模型 (Harmonic Plus Noise Model,HNM) 是由 Stylianous 首次提出。它将信号分成二个部分，分别是谐波成分和噪声成分。谐波成分针对信号中的低频部分，可以使用基频、幅度、相位三个参数来表示；噪声成分针对信号的高频部分，可通过高通滤波器得到高斯白噪声来表示。通过对信号的进行高低频率的分解使得合成出的语音信号更加自然，贴近人的真实语音。

2.4 神经网络及深度神经网络模型

2006 年，Hinton 提出 DBN（Deep Belief Network）模型，成为神经网络发展的新契机。深度神经网络 DNN (Deep Neural Networks) 是多层神经网络 MLP(Multiple Layer Perception)，二个模型在结构上大致相似，不同之处是深度学习网络先做非监督学习，学习到的权值做为监督学习的初值进行训练。

3 语音合成的最新发展

3.1 WaveNet: 原始音频生成模型

WaveNet 是生成原始音频波形的深层生成模型该技术已经被 Google DeepMind 引入，用于教授如何与计算机对话。引入结果令人满意，在网上可以找到合成声音的例子，如电脑学习如何用名人的声音与人们谈话。

WaveNet 属于卷积网络，其卷积层具有各种扩张因子，使得感受野随深度呈指数增长，有效地覆盖了数千个音频时间步长。它也是一种序列生成模型，可以用于语音生成建模。在语音合成的声学模型建模中，可以直接学习到采样值序列的映射，因此具有很好的合成效果。目前 wavenet 在语音合成声学模型建模、声码器方面都有应用，在语音合成领域有很大的潜力。

3.2 Deep Voice

Deep Voice 是百度 AI 研发的一个完全由深度神经网络构建的高质量语音转文本系统。Deep Vioce 与传统的 TTS 采用了同样的架构，但使用神经网络取代了所有组件，且使用了更简单的语音特征，从而使得系统

更适用于新数据集、没有标注的语音或其他需要特征调配的领域。

3.3 利用小样本的神经网络语音克隆

语音克隆是个性化语音接口一项备受期待的能力，而利用小样本的神经网络语音克隆模型引入了一个神经语音克隆系统，它可以通过学习，从少量音频样本合成人的语音。系统使用两种方法：说话人自适应和说话人编码。说话人自适应基于在一个多说话人生成模型上，使用少量克隆样本进行微调。说话人编码基于训练一个单独的模型以直接由克隆音频推断新的说话人，这在一个多说话人生成模型中被使用。前者可以实现更好的自然度和相似度，后者克隆时间或所需存储明显更少且有利于低资源部署。

3.4 Tacotron: 端到端的语音合成

Tacotron 是基于序列到序列的注意力机制的端到端语音生成模型。这个模型呢输入的是字符，输出的是原始频谱图，同时采用多种方法提高 seq2seq 模型的能力。对于给定的 < 文本 , 语音 > 对，Tacotron 能够通过随机初始化，完整地从头开始训练。它不需要音素级对齐和标注，可以轻松地扩展为大量的带有文本的声学数据。然后通过简单的波形合成，生成特别好的美式英语语音，就自然性而言优于参数模型。

3.5 Tacotron2: 端到端的语音合成

该系统包括两个部分，一是循环序列到序列的特征预测网络，将特征叠加到梅尔光谱图上，再通过一个修正过的 Wavenet 作为声码器，梅尔光谱图作为输入，经过声码器，合成为时域波形，最后得到的合成语音比参照语音的 MOS 仅仅低了 0.05。Wavenet 和 Tacotron 在 TTS 领域的应用都取得了非常优秀的成果，但也都存在着自身的缺陷。Tacotron2 将这两种方法的优点联合起来，应用于其中。

4 近一年研究情况

对 2019 年 7 月 1 日至 2020 年 6 月 30 日期间发表的最近发表的研究论文进行检索。中文论文选用的是中国知网 (CNKI) 数据库，分别进行期刊检索与会议检索，检索条件设置为主题与关键词与的关系，检索词为语音合成、合成语音和 TTS，来源类别选择 SCI 来源期刊、EI 来源期刊、核心期刊、CSSCI 和 CSCD。英文论文选用的是 ACM 数据库，检索采用标题与作者关键词且的关系，检索词为 Speech synthesis、Synthesized speech 和 TTS，出版日期为 2019-07-01 至 2020-06-30，内容类型为研究文章。

经过学科精炼、逐篇浏览等步骤，最终筛选出 18 篇文献作为样本，其中中文 9 篇，英文 9 篇。中文文献中针对少数民族语言的有 2 篇，帕丽旦 · 木合塔尔等针对维吾尔语语音合成输入文本进行了情感分类任务，提出了基于注意力机制的 BiRNN 情感分类模型 [1]。都格草等为了提高语音合成的自然度和清晰度，通过分析藏文字结构与拼读规则，融合 Seq2Seq 模型和注意力机制研究基于神经网络的藏语语音合成技术 [3]。端到端语音合成 2 篇，包括邱泽宇等用基于注意力机制的 Seq2Seq 模型训练一个特征预测网络，然后获取待合成语音的梅尔声谱图，利用 WaveNet 架构恢复损失的相位信息来实现语音合成 [4]。王国梁等验证了一个基于 Tacotron2 的中文 CNN 语音合成方案，在语料有限的情况下，可以实现端到端的较高质量中文语音合成 [5]。张鹏远等在预训练语言表示模型 BERT 的基础上构建了韵律结构预测模型，将多级韵律边界的预测视为相关的任务，通过多任务学习的框架捕捉各层级间的关系，实现了对它们的同时预测 [6]。何家勇等进行基于噪音条件下合成语音的可理解度实验，研究对我国的英语语音教学有重要启示 [7]。李燕萍等提出一种基于变分自编码器和 ACGAN 的语音转换框架，可以进一步提升非平行文本条件下多对多语音转换的性能 [8]。吴彭龙等提出采用两种改进型截幅修复方法对截幅语音进行修复，提升低速语音编码性能上的优越性 [9]。

9 篇英文论文，其中 5 篇是关于模型或算法改进的，Liu R 等针对 tacotron2 长时间语音合成的不足，提出一种基于 Tacotron2 的无监督生成对抗网络模型 [10]；Zheng Y 等针对 Tacotron 之类的神经网络端到端 TTS 在某些具有挑战性的测试集上性能不太令人满意的情况，提出两种新颖的方法用于规范端到端 TTS 的前向后解码序列 [11]；Zhou X 等提出一种使用深度中性神经网络（DNN）学习和建模单元嵌入的方法，用于基于单元选择的普通话语音合成 [12]；Naoto Umezaki 等提出了一种新的分类器训练概念，该概念结合了代表分类器参数语音合成能力的正则化项 [13]；Fuchun Liu 等提出一种新的算法可以提供可靠的二维声音定位信息，从而确保系统的实时性能并对基于固定波束形成的语音进行增强 [14]。2 篇关于感知分析方面，Noé Tits 等提出了一种方法来分析控制 TTS 系统参数对生成句子的感知的影响 [15]；Takuya Asakura 等提出一种情感语音转换方法，该方法可以使用周期一致的生成对抗网络（CycleGAN）从中性语音中产生情感发声 [16]。Wang Y 等介绍了一种新颖的方法来合成漫画的逼真的语音 [17]。Kristen M Scott 等研究了语音合成技术中有关语音捐

赠的道德问题 [18]。

5 结论

本文系统回顾语音合成领域的发展历程、研究概况及研究热点和关键技术，理清研究语音合成发展的脉络。研究发现，语音合成的理论与实践应用已经相对成熟，神经网络模型成为语音合成领域里的主力军。经过上述工作，希望为语音合成领域工作者的研究提供进一步的参考和帮助。

在 TTS 研究方向，有很多种方法，关于 AI 语音的未来之路，将来可以从神经网络模型改进，也可以与从具体领域的结合着手，如机器翻译领域，未来语音进一步渗透，每个人可能都能用自己的声音通译世界；还比如声乐领域，融合歌唱合成技术将为未来虚拟 IP 打造提供强大助力；还有语音识别在智能家居领域的应用 [19]。

参考文献

[1] 帕丽旦·木合塔尔,买买提阿依甫,杨文忠,吾守尔·斯拉木.基于BiRNN的维吾尔语情感韵律短语注意力模型[J].电子科技大学学报,2019,48(01):88-95.

[2] 刘梦媛,杨鉴.基于HMM的缅甸语语音合成系统设计与实现[J].云南大学学报(自然科学版),2020,42(01):19-27.

[3] 都格草,才让卓玛,南措吉,算太本.基于神经网络的藏语语音合成[J].中文信息学报,2019,33(02):75-80.

[4] 邱泽宇,屈丹,张连海.基于WaveNet的端到端语音合成方法[J].计算机应用,2019,39(05):1325-1329.

[5] 王国梁,陈梦楠,陈蕾.一种基于Tacotron 2的端到端中文语音合成方案[J].华东师范大学学报(自然科学版),2019(04):111-119.

[6] 张鹏远,卢春晖,王睿敏.基于预训练语言表示模型的汉语韵律结构预测[J].天津大学学报(自然科学与工程技术版),2020,53(03):265-271.

[7] 何家勇,周阳,刘伊梅.音段与韵律对中国学习者英语可理解度的贡献——基于噪音条件下合成语音的可理解度实验[J].外语学刊,2019(06):71-78.

[8] 李燕萍,曹盼,石杨,张燕,钱博.非平行文本下基于变分自编码器和辅助分类器生成对抗网络的语音转换[J].复旦学报(自然科学版),2020,59(03):322-329.

[9] 吴彭龙,邹霞,孙蒙,张星昱.截幅失真对低速语音编码的影响分析及改进[J].信号处理,2020,36(03):426-438.

[10] Liu R,Yang J,Liu M.A New End-to-End Long-Time Speech Synthesis System Based on Tacotron2[C]//the 2019 International Symposium.2019:46-50.

[11] Zheng Y,Tao J,Wen Z,et al.Forward–Backward Decoding Sequence for Regularizing End-to-End TTS[J].IEEE/ACM Transactions on Audio,Speech,and Language Processing,2019:1.

[12] Zhou X,Ling Z,Dai L.Learning and Modeling Unit Embeddings Using Deep Neural Networks for Unit-Selection-Based Mandarin Speech Synthesis[J].2020(PP)1-14.

[13] Naoto Umezaki,Takumi Okubo,Hideyuki Watanabe,Shigeru Katagiri,Miho Ohsaki.Minimum Classification Error Training with Speech Synthesis-Based Regularization for Speech Recognition[C]//SPML'19: Proceedings of the 2019 2nd International Conference on Signal Processing and Machine LearningNovember 2019:62-72.

[14] Fuchun Liu,Yang Yang,Qiguang Lin.Sound Source Localization and Speech Enhancement Algorithm Based on Fixed Beamforming[C]//CACRE2019:Proceedings of the 2019 4th International Conference on Automation,Control and Robotics EngineeringJuly 2019:1-7.

[15] Noé Tits,Kevin El Haddad,Thierry Dutoit.Neural Speech Synthesis with Style Intensity Interpolation:A Perceptual Analysis[C]//HRI '20:Companion of the 2020 ACM/IEEE International Conference on Human-Robot InteractionMarch 2020:485-487

[16] Takuya Asakura,Shunsuke Akama,Eri Shimokawara,Toru Yamaguchi,Shoji Yamamoto.Emotional Speech Generator by using Generative Adversarial Networks[C]//SoICT 2019:Proceedings of the Tenth International Symposium on Information and Communication Technology December 2019:9-14.

[17] Wang Y,Wang W,Liang W,et al.Comic-guided speech synthesis[J].ACM Transactions on Graphics,2019,38(6):1-14.

[18] Kristen M Scott,Simone R Ashby,David A Braude,Matthew P Aylett.Who owns your voice?:ethically sourced voices for non-commercial tts applications[C]//CUI '19:Proceedings of the 1st International Conference on Conversational User InterfacesAugust 2019:1-3.

[19] 王爱芸.语音识别技术在智能家居中的应用[J].软件,2015,36(7):104-107.