2016 年 1 月 第 1 期
Vol. 37 No. 1 2016

语音合成方法和发展综述

张 斌[1],全昌勤[2],任福继[1,3]

[1](合肥工业大学 计算机与信息学院 情感计算研究所,合肥 230009)
[2](神户大学 大学院系统情报学研究科,日本 神户 657-8501)
[3](德岛大学 科学与技术研究所,日本 德岛 770-8506)
E-mail: ZhangBin@ mail. hfut. edu. cn

摘 要: 人机交互中,最自然、最理想的交流莫过于通过人的声音进行交流. 这其中主要涉及到了语音合成,即文本转换为语音的技术. 本文的目的在于提供一个对语音合成发展简明深刻的介绍,并提出发展中的问题和解决方案,使刚进入这个领域的研究人员能够站在巨人的肩膀上,对语音合成有个清晰深刻的认识,并对自己即将展开的工作有个正确的判断. 文章首先总体上介绍语音合成已存在的,主流有效的方法,阐述方法的主要思想和优劣势,在此基础上启迪新想法新思路. 后分别说明近几年国内外研究人员在语音合成工作上所做的努力,客观评判与分析合成技术改进中的得与失,划出合成技术近年的发展趋势. 最后,得到语音合成技术的展望,指出发展的瓶颈并给出解决的方向.

关 键 词: 语音合成; HTS; 文本分析; 人机交互
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2016)01-0186-07

Overview of Speech Synthesis in Development and Methods

ZHANG Bin[1], QUAN Chang-qin[2], REN Fu-ji[1,3]

[1](Affective Computing Research Institute, Hefei University of Technology, Hefei 230009, China)
[2](Graduate School of System Informatics, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan)
[3](Institute of Technology and Science, The University of Tokushima 2-1, Tokushima 770-8506, Japan)

Abstract: In human-computer interaction, the most natural and the best way to exchange is the communication by human voice. Which is mainly related to speech synthesis, that is, the technology of converting text to speech. This paper provides a concise but deep introduction to the development of speech synthesis and propose the problems and solutions in the development of speech synthesis. The researchers who are just entering the field of speech synthesis can stand on the shoulders of giants, and have a clear and deep understanding of voice synthesis, and start working with a correct judgment. In this paper first the effective methods which are already exist and mainstream in speech synthesis are overall introduced and, the main idea of these methods as well as their advantages and disadvantages are described. On this basis, we inspire new ideas. After that, this paper illustrates the efforts respectively at home and abroad in recent years that researchers have done in the field of speech synthesis. Then objective evaluate and analyze gains and losses in synthesis technology improvements and draw the trend of synthesis technology in recent years. Finally, get the prospect in speech synthesis, pointing the bottleneck in development and trying to give direction to solve them.

Key words: speech synthesis; HTS; text analysis; HCI

1 引 言

语音,作为沟通最自然的方式. 语音合成和语音识别作为语音技术的两个分支,一个将文本信息转换为语音信号,而另一个则负责将语音转换成人们能够理解的形式,如文本,情感信息等. 从实现上来看,语音识别入门较快,原因在于语音识别对语音语料库的要求(标准等)没有语音合成用库那么严格,但由于语音识别受环境、语种、发音方式、吐字清晰度等因素的影响,后期研究过程中难度较大. 而语音合成,对语音库的要求则相对较高,需要对语料库进行系统、一致性的标注,不仅需要大量人力物力,对语言和语音的基础知识掌握也有较高的要求. 起点高,是语音合成的一大特点. 因此,目前急需一份简明深刻对语音合成主要思想、基本方法、以及发展的总结. 本文的目的就是为后来的研究者阐述语音合成的主要思想、关键技术和发展. 与其他综述[1,2]不同的是,本文更重视合成技术的发展. 使读者深入理解语音合成掌握发展的脉络因此能够快速但切中要害的掌握该研究领域并为工作的选题等做出正确的分析和判断.

2 语音合成的方法

语音合成的发展经历了机械式语音合成、电子式语音合

收稿日期: 2015-01-06 收修改稿日期: 2015-04-20 基金项目: 国家自然科学基金项目(61432004)资助; 国家自然科学基金面上项目(61472117)资助; 国家自然科学基金青年项目(61203312)资助; 国家"八六三"高技术研究发展计划项目(2012AA011103)资助; 教育部留学回国人员基金资助. 作者简介: 张 斌,女,1990 年生,硕士研究生,研究方向为语音信号处理、语音合成、语音识别; 全昌勤,女,1978 年生,博士,博士生导师,研究方向为自然语言处理、机器学习和情感计算; 任福继,男,1959 年生,博士,博士生导师,研究方向为人工智能、语言理解和交流、情感计算.

成和基于计算机的语音合成发展阶段. 基于计算机的合成方法由于侧重点不同,语音合成方法的分类也有差异. 但主流的、获得多数认同的分类则是将语音合成方法按照设计的主要思想分为规则驱动(rule-based)方法和数据驱动(data-based)方法[2]. 前者的主要思想是根据人类发音物理过程从

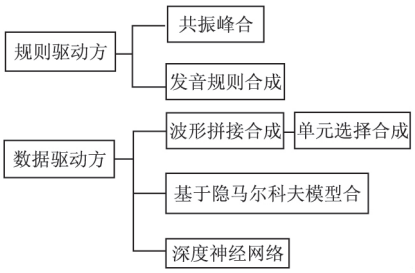

图 1　语音合成方法

Fig. 1　Methods in speech synthesis

而制定一系列规则来模拟这一过程,后者则是在语音库中的数据上利用统计方法如建模来实现合成的方法,因而数据驱动方法更多的依赖语音语料库的质量、规模和最小单元等. 语音合成的具体分类如图 1 所示,各个方法也不是完全独立的,近些年来研究人员取长补短地将它们整合到一起[3-5].

2.1　共振峰合成

共振峰是指声道的共振频率,共振峰合成是指用共振峰来加权叠加生成语音. 从滤波器的观点来看,语音的产生是一个 声源的激励加时变滤波的过程,如图2所示. 脉冲发生器模

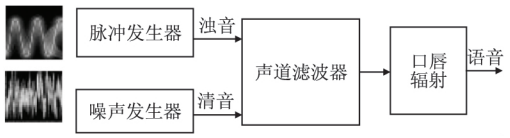

图 2　发声的源-滤波器模型

Fig. 2　Source-filter model in pronunciation

拟产生浊音的声带振动激励,清音是由声带中气息的湍流噪声造成的,用一个噪声发生器来模拟. 所有的语音都是这两类声源通过频率响应不同的滤波器处理后得到,用一个多通道的时变滤波器来模拟,使得其输出具有目标语音的频谱特性. 经过放大器(口唇辐射)输出,就可以听到合成语音. 最初,共振峰合成出的语音自然度很低,有些学者提出是因为共振峰建模时忽略了谱的变化. 经过在共振峰合成中加入或改进谱建模[6,7]. 共振峰合成的语音的自然度被提升了,由于合成时 F0-F5 可以控制变化,所以也常用来生成特色的语音,例如[5]中就用它来合成朝鲜族喜爱的乐器奚琴的乐音,[7]中则用共振峰合成来合成带有情感的语音. 这都源于它的灵活可调节性.

2.2　发音过程合成

发音过程模拟合成是直接模拟人的发音这一物理过程,通常制定一系列规则来操控模型发声. 由于得到真实发音的物理过程难度大,这一方法也较难实现. 但它的优点在于,一旦一个精细较为准确的规则建立,就使得这个系统有很大的可塑性和灵活性. 规则驱动方法的另一不足在于对超音段的控制不足[8],自然度受损,以至于有人们难以接受的机器声音. 为了在高复杂度和高自然度之间做一个平衡,研究人员采用预先录制的语音库,通过拼凑语音库单元来快速生成较高质量的语音.

2.3　波形拼接

波形拼接方法(concatenative synthesis)通过连接小的、事先录好的语音单元,如音素,双音素,三音素等并经过韵律修饰(prosodic modification)来拼接整合成完整的语音.

波形拼接技术是一种通过波形处理,使得言语的超音段特征发证改变,而音段特征(谱包络)保持不变的时间维处理技术. 这种技术最大限度的保留了原始发音人的音质,自然度和清晰度都很高,达到人们能够接受的水平. 但这样直接拼接的方法导致语音听起来人工、生硬,韵律修饰导致边界处明显不连续. 拼接处容易产生意想不到的错误,合成效果不稳定,音库容量大,构建周期长,可扩展性太差,不适宜作为嵌入式应用. 但如果要合成的语句中的大部分单元都在语音库里存在,那么合成出的语音的自然度要比规则拼接高得多,以至于当寻求高自然度时如商用,这类方法成为主流方法. 但它的代价则是设计精细、科学,占用内存大,人力物力耗费巨大的语音语料库.

2.4　单元选择

单元选择(unit selection)是一种波形拼接方法,但是它在事先录好的库中存储了每个拼接单元的大量不同韵律实例,这样就避免了传统波形拼接中的韵律修饰,也就解决了传统波形拼接方法中语音单元边界不连续的问题. 一般来说,单元选择方法合成的语音音质好,稳定,自然度较高. 但单元选择方法也像其他波形拼接方法一样存在拼接时选择了错误单元的情况.

2.5　谐波加噪声模型

为了解决单元选择中的误拼情况,研究人员又提出了谐波加噪声模型(harmonic plus noise model,HNM),该模型将语音信号看成是各种分量谐波和噪声的加权和. 对信号的这种分解使得合成出的信号更加自然.

2.6　HMM 模型和 STRAIGHT 合成技术

如前所述,波形拼接方法需要的语音语料库非常占用资源而且要求设计精细,因为它所有的拼接单元全都来自于库,而且训练模型的时间通常很长. 隐马尔科夫模型(HMM)结合谐波加噪声模型一起,解决了这个问题. 这种方法也被看作是最有用的统计建模方法. 它的流程如下: 首先,选择合适的特征表征语音库中的语音,训练模型; 然后,利用模型将文本生成序列状态的特征向量; 最后送入一个滤波器,将特征向量转换成语音. 基于 HMM 模型建模方法,灵活度高,库小,并且构建时间也少. 非常适合移动嵌入式平台.

20 世纪出现了数据驱动向规则驱动的倾向,其重要标注就是新的[9-12]语音处理技术和 HMM 统计模型,使得参数合成出现了新局面. 目前,这种方法已经相对成熟,它的训练流程在网上也是开源的,并被不定时更新,现在最新的训练流程版本为 HTS 2.2 [13].

2.7　神经网络及深度神经网络模型

深度神经网络 DNN(Deep Neural Networks)属于多层神经网络[14] MLP(Multiple Layer Perception),二者在结构上大致相似. 不同的是深度学习网络在做有监督学习的时候先做非监督学习,然后将非监督学习到的权值当作有监督学习的

初值进行训练.

深度学习网络首先被应用于语音识别领域,在 Google 语音搜索上最开始被应用,识别率提升 10% 以上,大大吸引了研究人员的关注. 后来研究人员慢慢发掘它在语音信号增强[15]、机器翻译等语音相关方向的应用. 深度学习的实质,是通过构建具有很多隐层的机器学习模型和大数据训练来学习更有用的特征,免去人工选取特征的过程,从而最终提升分类或者预测的手段,尤其适合语音、图像这种特征不明显的问题. 语音合成方面[17],DNN 可以用来给输入文本和对应的声学参数之间的关系建模[18]. DNN的应用解决了传统方法中

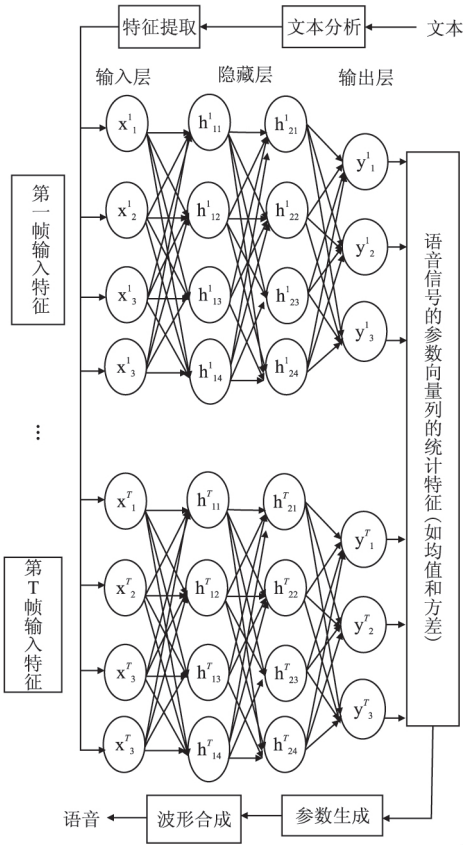

图 3 基于 DNN 的语音合成系统框架

Fig. 3 A speech synthesis framework based on DNN

上下文建模的低效率、上下文空间和输入空间分开聚类而导致的训练数据分裂、过拟合和音质受损的问题. 从图 3 中可以看出,DNN 用来给输入文本特征和声学特征的关系建立模型,输出的 Y 向量包含了谱特征和激励参数以及它们的时序导数. 在基于 HMM 模型的统计语音合成方法中,DNN 在清浊音分类和准周期预测都比传统的单纯基于 HMM 的方法要好,但在基频预测方面则不如原来的系统[19]. 针对这种情况,[20,21] 做出了各自的改进. DNN 的实现技术也开始被应用于合成方法的改进,如用 RBM 代替高斯分布来对 HMM 每个状态谱参数的分布建模[22] 来提高谱包络的建模精度.

3 实现算法

3.1 波形拼接-PSOLA 算法

用 PSOLA 算法实现语音合成时主要有三个步骤. 分别为基音同步分析、基音同步修改和基音同步合成.

同步分析主要是对合成单元进行同步标记设置. 同步标记是与合成单元浊音段的基音保持同步的一系列位置点,用它们来准确反映各基音周期的起始位置. 对于浊音段有基音周期,而清音段信号则属于白噪声,所以这两种类型需要区别对待. 同步修改通过对合成单元同步标记的插入、删除来改变合成语音的时长; 通过对合成单元标记间隔的增加、减小来改变合成语音的基频等. 基音同步合成是利用短时合成信号进行叠加合成. 如果合成信号仅仅在时长上有变化,则增加或减少相应的短时合成信号; 如果是基频上有变化,则首先将短时合成信号变换成符合要求的短时合成信号再进行合成.

由于韵律修改所针对的侧面不同,PSOLA 算法的实现目前有 3 种方式. 分别为: 时域基音同步叠加 TD-PSOLA,线性预测基音同步叠加 LPC-PSOLA,频域基音同步叠加 FD-PSOLA. 其中 TD-PSOLA 算法计算效率较高,已被广泛应用,是一种经典算法,TD-PSOLA 算法原理可以参见[8].

基音同步叠接相加法,优点: 良好的韵律调整能力的; 缺点: 基音频率修改过大时可能出现严重的谱包络失真.

PSOLA 技术在法语、德语、和日语的文语转换系统中获得成功. 直至今日,Praat 软件中,音高和时长的调节还是采用 PSOLA 技术.

3.2 STRAIGHT 算法

STRAIGHT 算法采用了以 Dudley(1939) 的 VOCODER 为原型[23] 的源-滤波器的思想来表征语音信号. 但这种方法合成出的音质并不好,调整能力也不强. STRAIGHT 算法对这些方面进行了改进,一方面通过采用一些基于听觉感知的方法对语音信号的合成端进行改进,以提高合成语音的音质; 另一方面,通过消除谱参数中的周期性来提高谱估计的准确性,通过重拾“听觉场景分析”模型来实现源与滤波器的完全剥离,提高参数调整时的灵活度. 改进的 STRAIGHT 算法在合成音质和调整能力上都比经典的 PSOLA 算法具有优势.

传统提取基频: 求 F0 转换为求 T0,即寻找除延时量为 0 的下一个自相关函数的极大值,对应 T′,T0 = T. 这样做存在的问题: 自然语言信号既非周期也非平稳. 所以不存在周期的概念. 解决[9]: 提取基波分量的瞬时频率的方法——STRAIGHT,即: 引入基频成分(fundamental component) 概念,它与基音的先验知识关联较少,且在高频处峰陡,低频处峰缓. 根据它对噪音、基频、谐波不同的位置关系值的不同(在基频处极大值),所以提取基频等价于寻求基音指数(fundamentalness index) M_C 极大,从而得到基频. 运用 STRAIGHT 不仅可以提取基频,谱包络,而且可以按照源-滤波器的模型,合成语音.

语图上的周期性扰动是窗函数带来的,并不是语音本身的特性. 如果我们在这些有周期性扰动的动态谱图上进行调制,再恢复到语音时必然会引入噪声. STRAIGHT 在分析过程中可以消除语图中频率维和时间维周期性扰动. 除了采用基音周期自适应的同步分析外,还采取了消除空洞的处理,即在设计时间窗的同时,设计了一个补偿窗,以补偿空洞的不良影响. 除了在分析阶段,消除周期性扰动之外,STRAITHT 还可以在分析调制过程中提取基频、合成重建语音. 由于其灵活性,针对人耳感知更好的调整谱参数,在个性化语音、情感语

音的生成中,常常使用 STRAIGHT.

STRAIGHT 的原理可以参见[9-12],从 STRAIGHT 的提出到改进.

3.3 Trainable TTS

Trainable TTS 主要包括训练和合成两部分[24,26,28].

在训练过程中,利用 HMM 训练对基频、时长、谱参数进行建模.

合成过程中,对输入文本进行属性分析,并利用训练好的模型进行参数预测,最后通过参数合成器合成出最后的语音. 其中,训练流程如图 4 上半部分,合成部分如图 4 下半部分. 具体流程参见[28]. 训练流程中有待研究的是更好的上下文属性集的设计和聚类方法.

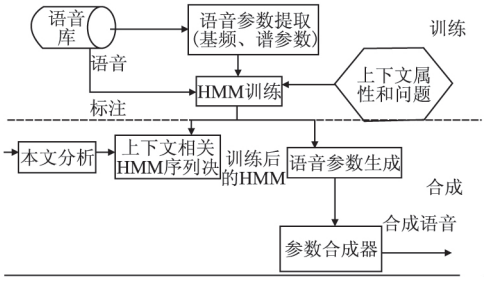

图 4　Trainable TTS 系统框架

Fig. 4　A framework of Trainable TTS

训练流程均是自动进行的,是一种快速建系统的方法.

STRAIGHT 方法提取基频 F0 和谱包络,对参数进行调制之后,产生新的声源和时变滤波器,按照源滤波器模型,采用特定的公式合成语音. 这个公式能够确保 F0、幅度、频率、和时间维都能得到调制.

除了上述算法之外,还有在 HMM 更新参数时采用 Viterbi 算法和 Baum-Welch 算法; 以及实现的工具 HTK,Praat 软件等. Praat 拥有强大的语音分析功能,详见[49]. 指出一点: 编写 Praat 的脚本程序可以将 Praat 做入系统中.

3.4 限制玻尔兹曼机和深信度网络

2006 年,深度学习泰斗 Geoffrey Hinton [16] 等提出的深度网络模型训练,被分成简单的两步:

1) 预训练: 用未标注数据无监督学习初始化模型参数;

2) 有监督学习: 利用传统的神经网络算法学习模型参数.

3.4.1 限制玻尔兹曼机

RBM 是一种二部图,包含可见层和隐含层的双层图模型[25],如图 5 所示. 可视层(v) 和隐含层(h),如果所有的节点都是随机二值变量节点(取值只能是 0 或 1),同时全概率分布 p(v,h) 满足 Boltzmannn 分布,这个模型就称为受限玻尔兹曼机. 在可见层 v 和隐含层 h 都已知的情况下,给定模型参数 $\theta=\{w_{ij},b_i,a_j,i=1,\cdots,M;j=1,\cdots,N\}$ 能量分布函数 E(v,h;θ):

$$E(v,h;\theta)=-\sum_{i=1}^{M}\sum_{j=1}^{N}v_i w_{ij}h_j+\frac{1}{2}\sum_{i=1}^{M}(v_i-b_i)^2-\sum_{j=1}^{N}h_j a_j \tag{1}$$

RBM 的联合概率分布为

$$p(v,h;\theta)=\frac{\exp(-E(v,h;\theta))}{Z} \tag{2}$$

联合概率对隐含层节点进行求和,就得到关于可见层节点状态的边缘概率 p(v;θ). RBM 训练的目标函数为

$$\hat{\theta}=\arg\max \log P(v;\theta) \tag{3}$$

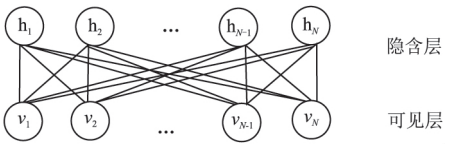

图 5　RBM 模型

Fig. 5　RBM model

参数更新公式则可以由求目标函数(3) 的偏导数得到. 如果将隐含层层数增加,就可以得到 Deep Boltzmannn Machine(DBM); 如果将最靠近可见层的那一层隐含层使用贝叶斯信念网络,离可视层最远的隐含层中使用 RBM,此时我们就得到 Deep Belief Net(DBN).

3.4.2 DBN-DNN

通过逐层构建 RBM 的方式可以构建一个 DBN 模型[27]. 前一个 RBM 的输出作为下一个 RBM 训练的输入,通过这种自底向上的层级构建最终形成 DBN. 如果在 DBN 的顶层加入一个 Softmax 分类器输出层,就形成具有初始化网络参数的 DNN,如图 6 所示. Softmax 输出层对应 DNN 输出向量,在语音合成中是语音参数向量序列的统计特征,如均值方差等.

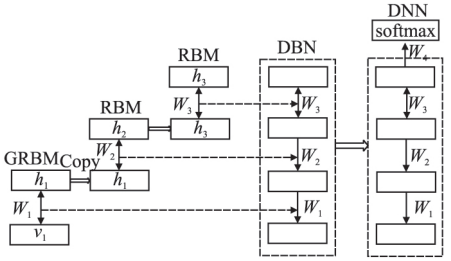

图 6　DBN-DNN 训练流程图

Fig. 6　Training flow chart of DBN-DNN

4 方法的改进

4.1 典型的针对性改进

针对性的改进需要对方法细节有足够深刻的理解. 它的流程中每一步的作用,它的某一个不足对合成出的语音产生什么样隐患都要了然于胸. 这种改进往往出现在要应用的这种方法总体符合合成语音的要求,但存在小的不足的时候. Udochukwu Ogbureke,Joao Cabral[29] 等人提出了用多层感知分类器(MLP) 来感知清浊音,因而改变了时长的隐马尔科夫模型(HMM) 提高了音质. 当 Hyes Rebai 和 Yassine BenAyed[30] 针对阿拉伯语中的边音符,提出在合成器前加一个变音符分析系统,其作用是分析得到传统的不带变音符的文本,然后采用多层感知人工神经网络(MLP) 训练,最后将数据送入梅尔对数谱近似滤波器来合成阿拉伯语. 同样的还有 Liping Kui,Jian Yang,Bing He[31] 等人用 HTS2.0 流程实现越南话的合成; Sara Bahaadini,Hossein Sameti[5] 利用谐波加噪声模型(HNM) 和 STRAIGHT(Speech Transformation and Representation using Adaptive Interpolation of weiGHTed Spectrum) 合成器来实现波斯语的合成. 这些改进的思路,都是采

用 HTS 流程的高度灵活性，资源占用小，快速构建等特点或是单元选择方法合成的音质好，或是应用一些更适合的模型，如人工神经网络、SVM、CRF[4] 和 DNN 等.

4.2 方法的融合

除了单独改进某种方法，研究者常常选择方法的融合，扬长避短，“各取所优”. 传统的基于 HMM 建模的语音合成中存在过平滑问题(导致语音波形起伏不如自然语音大，声音听起来有点“闷”)，Shinnosuke Takamichi，Tomoki Toda[5] 等人将基于 HMM 的合成方法和单元选择方法杂合，音质大幅度提高，但失去了基于 HMM 的灵活性. 同样地，Yannis Styliano[32] 也提出了谐波加噪声模型 HNM 和单元选择方法的融合，并与 TD-PSOLA 比较，波形拼接不连续现象被平滑掉，合成的语音在自然度、灵活性和愉悦度上都比 TD-PSOLA 方法要好. 但两种方法的融合，并不是单纯的拼凑到一起，为了发扬各自的优点，需在融合的时候扬长避短.

关于合成方法的融合问题，有一种趋势是将基于 HMM 合成方法等和单元选择方法融合. 在国际语音大赛 Blizzard Challenge[33] 等上，这种应用也较为常见. 单元选择法的原理可以参见 [34]. 优化单元选择 [35] 常常会被应用，对语音的连续性很关键.

5 语音合成的新发展和挑战

5.1 新要求新应用

当合成的语音自然度、灵活度等基本满足人们要求时，研究人员发现合成出的语音和自然人的语音还是有区别. 关于情感语音的合成，研究者考虑的是改变基频建模，使之能够调整基频来合成不同情感语音[36]. 为了增强语音的表现力，而不是一成不变的语音. 基于这个思想，Yan-You Chen[37] 等人在基于 HMM 的语音合成系统中，在训练隐马尔科夫模型时利用全协方差矩阵而不是原来的对角协方差矩阵，来保证语音的变化. 全协方差矩阵比对角矩阵携带了更多的信息用于训练. 不规律语音更能体现语音特色. 但不同于 [36] 中卷积一个随机向量，[38] 中在 HTS 流程中采用基音同步残差编码来给不规则语音建模，建模结果很理想，合成语句和原句很相似.

语音合成技术常常被用在人工智能(AI)领域，与语音识别，机器翻译等一起，旨在更好的和人沟通，更容易被人接受. 通过分析语音合成在机器翻译[39](speech to speech)中的作用，发现要转换句子的流利度直接影响合成效果. 近几年出现的自适应说话人技术，也是语音合成的一个应用. 它通过大量的多说话人语音来训练平均声学模型，在适应阶段，加入某个特定人的少量数据，来生成这个特定人的语音[40,41]. 语音合成在移动终端[42] 和嵌入式中[43] 也有应用. 移动设备上的合成，要求占用资源小，因而不可能采用波形拼接方法，通常是采用基于 HMM 的合成方法，或是采用云的方式访问占用内存大的资源. 因而在 [43] 中通过压缩数据库来在移动设备上采用参数法合成斯洛伐克语. 最后，深度学习网络 DNN 和已有其他合成方法的结合，如 2.7 节所述，常能够大幅度提高合成中建模的局部精度.

5.2 新挑战新方法

2006 年以来，DNN 方法在语音识别上大幅度提高识别率，让研究人员开始重视它. 在语音合成中，它也越来越多的被用来改进传统方法. 尤其在开源工具 kaldi[44] 的出现，使得 DNN 方法走下神坛，但其仍是较难入门的方法之一. 在 DNN 方法中，还有大量工作需要研究. 其中一个就是降低 DNN 的数据量，对数据进行降维. 对方法本身，也有较多的核心问题需要解决.

再者，语音合成分为前端和后端，前端是涉及自然语音处理较多的部分，即文本处理部分. 它将句子中进行分词，标音，标韵律等生成包含了分词结果、韵律、音素等的标注文件；后端则是声学处理，利用标注和语音进行语音信号处理，建模，最终产生语音的过程. 多数在语音合成中所做的努力都集中在后端. 但在合成的前端中，对文本处理这种不深入的分析正在阻碍语音合成技术取得突破. 虽然当下已经出现了一些新的、更好地表征语音的特征，然而它们本身仍然存在问题：无法判断其准确性，且无法实现自动分析[45]. Chen-Yu Yang，Zhen-Hua Ling 等人在 [46] 中提出了一种在中文普通话中采用自动的、无监督的、基于上下文独立的 HMM 模型方法来标记短语边界位置，不仅提高了标注的准确率，还省去了人工标注；William Yang Wang 和 Kallirroi Georgila[4] 采用了文本特征 TF-IDF 在单元选择中自动检测了不自然的词级段. 虽然他们迈出了第一步，但在这条路上还有很远要走. 另外，语音合成也衍生出不同侧重的研究方向，例如情感语音相关研究[47,48,50]，歌唱合成，语音和视频结合，语音和文本结合的趋势等等.

6 展望和总结

前端在后端日益完善和技术成熟的时候，渐渐被人们拾起. 一些研究人员已经发现了前端对合成的重要性[51,52]，但这种重视显然这还远远不够. 前端是语音合成中关键的一环，因为后续后端需要做的改进全部都需要前端资源的支持. 因而，可以预见语音合成技术接下来的发展主要体现在后端合成方法和前端的互溶性和互相促进，相互协调上. 语音处理技术的新方法，往往都融贯互通，因此可以用来改进语音合成. 例如前面提到的，最开始应用于语音识别的深度学习网络，在合成中也显示出其重要性.

本文介绍了语音合成主流有效的方法，对方法中固有的优缺点进行了阐述. 接下来通过介绍一些学者的相关工作来描绘合成方法的改进：根据不同的目的采取不同的合成方法，根据目的单独改进方法或是融合方法. 通过这些和语音合成有关的新的有创造性的工作，为学者今后的实验提供了思路和启迪. 一些创新想法的实现或者说某些方法的改进常常伴随着语音库设计的变动，来适应或是说支持这种实验，也就是说语音合成前端的文本处理和后端的声学处理密不可分. 给研究人员提供了语音合成简明深刻的阐述，为今后工作打下坚实基础.

References:

[1] Jing Xiao-yang，Luo Fei，Wang Ya-qi，Chinese speech synthesis technology overview [J]. Computer Science，2012，39(11A)：386-390.

[2] Youcef tabet，Mohamed boughazi. Speech synthesis techniques. a survey [C]. 7th International Workshop on Systems，Signal Processing and their Applications (WOSSPA)，2011：67-70.

[3] Sara Bahaadini, Hossein Sameti, Soheil Khorram. Implementation and evaluation of statistical parametric speech synthesis methods for the persian language [C]. IEEE International Workshop on Machine Learning for Signal Processing, 2011: 1-6.

[4] William Yang Wan, Kallirroi Georgila. Automatic detection of unnatural word-level segments in unit-selection speech synthesis [J]. IEEE Transactions on Speech and Audio Processing, 2011, 9(1): 21-29.

[5] Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, et al. Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis [J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(2): 239-250.

[6] Myeongsu Kang, Yeonwoo Hong, Formant synthesis of haegeum: a sound analysis/synthesis system using cepstral envelope [C]. International Conference on Information Science and Applications (ICISA), 2011: 1-8.

[7] Khorinphan C, Phansamdaeng S, Saiyod S, Thai speech synthesis with emotional tone: based on formant synthesis for home robot [C]. 2014 Third ICT International Student Project Conference (ICT-ISPC2014), 2014: 111-114.

[8] Lv Shi-nan, Chu Min, Xu Jie-ping, et al. Chinese speech synthesis [M]. Beijing: Science Press, 2012.

[9] Hideki Kawahara, Ikuyo Masuda Katsuse, Alain De Cheveign, Restructuring speech representations using a pitch-adaptive time ± frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds [J]. Speech Communication 27, 1999, 27(3-4): 187-207.

[10] Hideki Kawahara, STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds [J]. Acoust. Sci. & Tech. 2006, 27 (6): 349-353.

[11] Hideki Banno, Hiroaki Hata, Masanori Morise, et al. Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation [J]. Acoust. Sci. & Tech., 2007, 28(3): 140-146.

[12] Kawahara H, Morise M, Takahashi T. TANDEM-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and Aperiodicity Estimation [C]. IEEE International Conterence on Aconstics, Speed and Signal Processing, 2008: 3933-3936.

[13] HMM-based speech synthesis system (HTS) [EB/OL]. http://hts.sp.nitech.ac.jp, 2014.

[14] Zhou Zhi-hua, Chen Shi-fu. Neural network ensemble [J]. Chinese Journal of Computers, 2002, 25(1): 1-8.

[15] Xu Yong, Du Jun, Dai Li-rong, et al. An experimental study on speech enhancement based on deep neural networks [J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504-507.

[17] Zhang Zheng. Chinese speech synthesis based on the deep neural networks [D]. Beijing: Beijing Institute of Technology, 2014.

[18] Heiga Zen, Andrew Senior, Mike Schuster. Statistical parametric speech synthesis using deep neural networks [C]. IEEE International Conference on Aconstic, Speech and Signal Processing, 2013: 7962-7966.

[19] Chen Zhen-huai, Yu Kai. An investigation of implementation and performance analysis of DNN based speech synthesis system [C]. International Conference on Sociology and Psychology, 2014: 577-582.

[20] Sankar Mukherjee, Shyamal Kumar Das Mandal. F0 modeling in HMM-based speech synthesis system using deep belief network [C]. Co-ordination and Standardization of Speech Database and Assessment Techniques, 2014: 1-5.

[21] Yao Qian, Yuchen Fan, Frank K. Song. On the training aspects of deep neural network(DNN) for parametric TTS synthesis [C]. Internation Conference on Aconstic, Speech and Signal Processing, 2014: 3829-3833.

[22] Zhen-hua Ling, Li Deng, Dong Yu. Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis [C]. Internation Conference on Aconstic, Speech and Signal Processing, 2013: 7825-7829.

[23] Yang Jin-hui, Yi Zhong-hua, Straight a speech synthesis method based on focus [J]. Computer Engineering, 2005, 13(31): 46-47, 128.

[24] Wu Yi-jian, Wang Ren-hua, Research on HMM-based speech synthesis technology [D]. Hefei: University of Science and Technology of China, 2006.

[25] Deep learning [EB/OL]. http://deeplearning.net/, 2015.

[26] Ling Zhen-hua, Wang Ren-hua. Research on speech synthesis technology based on statistics acoustic modeling [D]. Hefei: University of Science and Technology of China, 2008.

[27] Dai Li-rong, Zhang Shi-liang. Deep voice signal and information processing [J]. Progress and Prospects, 2014, 29(2): 171-179.

[28] Wu Yi-jian. IFLYTEK internal data [R]. Technical Report in Trainable TTS, 2005.

[29] Udochukwu Ogbureke, Joao Cabral, Julie Berdsen. Using multilayer perception for vocing strength estimation in HMM-BASED speech synthesis [C]. The 11th International Conference on Information Sciences, Signal Processing and their Applications: Main Tracks, 2012: 683-688.

[30] Hyes Rebai, Yassine BenAyed. Arabic text to speech synthesis based on neural networks for MFCC estimation [C]. 2013 World Congress on Computer and Information Technology (WCCIT), 2013: 1-5.

[31] Kui Ling-ping, Yang Jian, He Bin, et al. An experimental study on vietnamese speech synthesis [C]. 2011 International Conference on Asian Language Processing, 2011: 232-235.

[32] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis [J]. IEEE Transactions on Speech and Audio Processing, 2011, 9(1): 21-29.

[33] Ling Zhen-hua, Xia Xian-jun, Yang Song. The USTC system for blizzard challenge [C]. Proc. of Blizzard Challenge Workshop, 2012.

[34] Andrew J Hunt, Alan W Black, Unit selection in a concatenative speech synthesis system using a large speech database [C]. Proceedings of ICASSP 96, 1996, 1: 373-376.

[35] Alan W Black, Nick Campbell. Optimizing selection of units from speech databases for concatenative synthesis [C]. In Proc. Euro speech, 1995: 5-8.

[36] Zhang Hao-jie, Yang Yong, Fundamental frequency adjustment and formant transition based emotional speech synthesis [C]. 2012 9th

International Conference on Fuzzy Systems and Knowledge Discovery,2012:1797-1801.

[37] Chen Yan-you,Kuan Ta-wen, Tsai Chun-yu,et al. Speech variability compensation for expressive speech synthesis [C]. 2013 International Conference on Orange Technologies (ICOT),2013:210-213.

[38] Tamas Gabor Csapo,Geza Nemeth. Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation [J]. IEEE Journal of Selected Topics in Signal Processing,2014,8(2):209-220.

[39] Kei Hashimoto,Junichi Yamagishi,An analysis of machine translation and speech synthesis in speech-to-speech translation system [C]. IEEE International Conference on Acoustics,Speech and Signal Processing,2011:5108-5111.

[40] Fahimeh Bahmaninezhad, Hossein Sameti, HMM-based persian speech synthesis using limited adaptation [C]. 2012 IEEE 11th International Conference on Data Signal Processing (ICSP),2012:21-25.

[41] Junichi Yamagishi,Takashi Nose,Heiga Zen,et al. Robust speaker-adaptive HMM-based text-to-speech synthesis [J]. IEEE Transactions on Audio,Speech,and Language Processing,2009,17(6):1208-1230.

[42] Martin Turi Nagy,Gregor Rozinaj,Peter Hvise. Parametrization of a slovak speech database for mobile platform speech synthesis [C]. 51st International Symposium ELMAR-2009,2009:225-228.

[43] Shu Chang,Mei Jin-shuo,Yin Jing-hua,Speech synthesis based on AMR-WB algorithm [C]. 2011 International Conference on Electronic & Mechanical Engineering and Information Technology,2011:2212-2214.

[44] Chinese home of kaidi project [EB/OL]. http://sourceforge.jp/projects/sfnet_kaldi/,2015.

[45] Zhu Wei-bin. Linguistics computational model in speech synthesis: status and prospects [J]. Contemporary Linguistics,2009,11(2):159-166.

[46] Yang Chen-yu,Ling Zhen-hua,Dai Li-rong. Unsupervised prosodic phrase boundary labeling of mandarin speech synthesis database using context-dependent HMM [C]. IEEE International Conference on Acoustics,Speech and Signal Processing,2013:6875-6879.

[47] Chen Zhen-jie,Wang Wei-hao. Study on the mechanism of emotional speech pronunciation [J]. Journal of Changchun Education Institute,2014,30(24):61-62.

[48] Han Wen-jing,Li Hai-feng,Ruan Hua-bin,et al. An Summary of emotional speech recognition [J]. Journal of Software,2014,25(1),37-50.

[49] Xiong Zi-yu. Praat speech software manuals [EB/OL]. http://202.121.96.130/Download/20101127203426_220736430186.pdf,2004.

[50] Wang Jing-hua,Liu Jian-yin,Zhang Guo-yan,et al. Research on fundamental frequency of prosodic parameters in emotional speech synthesis [J]. Journal of Chinese Computer Systems,2013,34(9):2047-2050.

[51] Zhang Sen,Liu Lei,Diao Hong-lu. Problems on large-scale speech corpus and the applications in TTS [J]. Chinese Journal of Computers,2010,33(4):687-696.

[52] Ni Chong-jia,Zhang Ai-ying,Liu Wen-ju. Mandarin stress detection using acoustic,lexical and syntactic features [J]. Chinese Journal of Computers,2011,34(9):1638-1649.

附中文参考文献:

[1] 井晓阳,罗 飞,王亚棋,汉语语音合成技术综述 [J]. 计算机科学,2012,39(11A):386-390.

[8] 吕士楠,初 敏,许洁萍,等. 汉语语音合成 [M]. 北京:科学出版社,2012.

[14] 周志华,陈世福. 神经网络集成 [J]. 计算机学报,2002,25(1):1-8.

[17] 张 征. 基于深度神经网络的汉语语音合成的研究 [D]. 北京:北京理工大学,2014.

[23] 杨金辉,易中华. 一种基于 Straight 的语音焦点合成方法 [J]. 计算机工程,2005,13(31):46-47,128.

[24] 吴义坚,王仁华. 基于隐马尔科夫模型的语音合成技术研究 [D]. 合肥:中国科学与技术大学,2006.

[26] 凌震华,王仁华. 基于统计声学建模的语音合成技术研究 [D]. 合肥:中国科学与技术大学,2008.

[27] 戴礼荣,张仕良. 深度语音信号与信息处理:研究进展与展望 [J]. 数据采集与处理,2014,29(2),171-179.

[28] 吴义坚. Trainable TTS 技术报告 [R]. 科大讯飞公司内部资料,2005.

[45] 朱维彬. 语音合成中的语言学计算模型:现状及展望 [J]. 当代语言学,2009,11(2):159-166.

[47] 陈志杰,王韦皓. 情感语音发音机理研究 [J]. 长春教育学院学报,2014,30(24):61-62.

[48] 韩文静,李海峰,阮华斌,等. 语音情感识别研究进展综述 [J]. 软件学报,2014,25(1):37-50.

[49] 熊子瑜. praat 软件参考手册 [EB/OL]. http://202.121.96.130/Download/20101127203426_220736430186.pdf,2014.

[50] 王敬华,刘建银,张国燕,等. 情感语音合成中韵律参数的基频研究 [J]. 小型微型计算机系统,2013,34(9):2047-2050.

[51] 章 森,刘 磊,刁麓弘. 大规模语音语料库及其在 TTS 中应用的几个问题 [J]. 计算机学报,2010,33(4):687-696.

[52] 倪崇嘉,张爱英,刘文举. 基于声学相关特征与词典语法相关特征的汉语重音检测 [J]. 计算机学报,2011,34(9):1638-1649.