文章编号:1001-0645(2007)05-0408-05

基于语义的语音合成
——语音合成技术的现状及展望

朱维彬[1], 吕士楠[2]
(1. 北京交通大学 信息科学研究所, 北京 100044; 2. 中国科学院 声学研究所, 北京 100080)

摘 要: 综述了语音合成技术的发展现状. 指出并分析了目前系统存在的发音质量、韵律预测、表现能力等3个方面的问题. 提出了将语义分析引入语音合成系统,使合成语音具有准确、生动的语义表现能力,并作为新一代语音合成系统的目标. 探讨了实现这一目标所涉及的理论基础、技术实现、基础资源等研究内容.

关键词: 语音合成; 语义; 韵律; 表现能力
中图分类号: TP 391 **文献标识码**: A

Semantic-Based Speech Synthesis
——Survey and Perspective on the Speech Synthesis Technology

ZHU Wei-bin[1], LÜ Shi-nan[2]
(1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China;
2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Overviews the state-of-the-art of speech synthesis technology. Problems such as speech quality, prosody prediction and expressiveness existing in current systems were analyzed. It is proposed to integrate semantic-analysis into speech synthesis systems. Thus, the goal of the next generation of systems ought to be to convey the semantic information perfectly and vividly through synthesized speech. The fundamental principles methodologies, and basic resources relevant to achieve the goals are discussed.

Key words: speech synthesis; semantics; prosody; expressiveness

语音合成系统性能可分为3个层次:表音(清晰、自然地合成出语音)、表意(准确地表达话语意图)、表情(生动地表现语意情感). 近年来,由于基于数据库的单元挑选及数据驱动建模技术的普遍采用,语音合成系统在可懂度、自然度等评价指标上有了显著提高[1-4],但在本质上仍处于表音层次. 为了使合成语音具备"表情达意"的能力,需使文本到语音的转换过程在语义层面上进行. 将语义处理机制引入语音合成,实现文本的语义分析,一方面有助于解决目前系统普遍存在的注音及韵律结构预测等方面的问题;另一方面,解决语义言语实现所涉及的语义重音、功能语调、发音方式等方面的建模与预测;在此基础之上,构建具有准确、生动表现能力的新一代语音合成系统.

在考查了语音合成技术的发展现状之后,作者基于对现有问题产生的原因进行了分析,指出了解决问题的关键在于语义分析及实现,阐述了下一阶段基于语义的语音合成技术在理论基础、技术实现

收稿日期: 2007-03-09
基金项目: 国家"八六三"计划项目(2006AA010104);北京交通大学校基金项目(2005RC014)
作者简介: 朱维彬(1966—),男,博士,讲师. E-mail:wbzhu@bjtu.edu.cn; 吕士楠(1937—),男,研究员,E-mail:lu_shinan@163.com.

以及资源建设等方面所涉及的研究内容.

1　技术现状

由于波形拼接技术的普遍采用，语音合成研究的重点已由早期的音段(segment)层级的处理转到了对整段话语(utterance)特性的建模，对合成语音质量的评价指标也由可懂度(intelligibility)转变为自然度(naturalness). 促成这种变化和进步的主要是韵律分析方法及数值建模技术两方面的突破.

用于描述英语韵律现象的符号系统 ToBI(tone and break index)的出现，是韵律研究进程中的标志性事件[5]. 尽管学术界对于 ToBI 定义的具体内容持有不同的观点[6-7]，但它所体现的离散的符号系统描述连续的韵律声学特征和对于真实样本实施标注分析的方法带来了基于标注数据库的韵律分析方法及韵律统计建模方法的盛行[8-9].

在数值建模方面普遍采用了基于数据库单元挑选的技术路线. 由于基本合成单元一般采用较大的音段单位(音素以上)，在相当程度上解决了合成语音的可懂度问题；而自然度问题的解决则是通过韵律模型约束下的最优单元挑选实现的，合成转化成了最优搜索问题. 事实上，语音识别中数据驱动建模、最优搜索等算法，在语音合成中得到了普遍采用[10].

汉语语音合成的重大突破，也是在制定汉语韵律标注符号系统 C-ToBI[11-12]及引入单元挑选的技术路线之后，普遍采用了以韵律词为基本单位的韵律层级结构作为汉语主要的韵律特征. 同时，在汉语韵律声学体现分析[13]、基于韵律标注数据库的韵律统计建模[14-15]等方面取得了实质性进展. 目前，汉语语音合成技术已经达到了实用化水平，并出现了捷通华声、科大讯飞等以语音合成系统为核心产品的专业化公司.

2　合成语音质量

虽然语音合成系统已经进入应用阶段，但依然存在诸多问题. 由美国自然科学基金支持的英语语音合成系统性能评测 Blizzard Challenge 已经连续举行了两届[16]. 参加单位基于共同的合成语音样本数据库开发各自的系统，提交合成语音样本用于评测. 这些合成语音样本设置了押韵测试(modified rhyme test, MRT)、语义不可预测句测试(semantically unpredictable sentences, SUS)，采用词错误率(word error rate, WER)反映合成样本的可懂度；设置了 5 分制的主观评价得分(mean opinion score, MOS)用以反映合成样本的自然度. 在 2005 年的测试中，由专家构成的测评组测试的最佳系统结果为：WER 为 14.7%，MOS 为 3.190 分. 2006 年的测试中，专家组给出的最高 MOS 为 3.696 分. 从测试结果看，无论是可懂度还是自然度，较自然人发音都还有明显的差距.

在国家“八六三”计划有关项目的支持下，针对汉语语音合成系统的研究，已进行了数次评测[17]. 在各次评测中，同样设置了可懂度和自然度的测试内容. 评测结果反映，在可懂度和自然度方面，汉语的语音合成质量同样还有相当大的提升空间.

无论是英语的 Blizzard Challenge 测试，还是汉语的“八六三”测试，由于采用统一的测试方案，各参加系统的性能变得可比，可以帮助各研究单位在算法层面做进一步地分析，从而推进了语音合成技术的进步. 虽然测试结果具有较高的权威性，但由于 WER 和 MOS 都是系统性能的综合性指标，不能直接反映系统各功能模块的性能和问题，为了分析影响系统性能的原因，还需进行针对性的诊断测试.

根据对数个具有代表性的汉语语音合成系统的考察，作者对观察到的合成语音缺陷进行了分类. 主要问题有：

①与可懂度直接相关的发音质量，包括音质缺陷、多音字、轻声、变调、数字串等方面的问题.

②与自然度密切相关的“分词断句”错误，反映了韵律结构预测方面的问题.

③合成语音音色单一、语调缺少变化、缺乏表现能力，直接原因是由于系统中没有轻重音、功能语调、发音风格等方面的控制.

这些问题的存在表明：目前的语音合成技术还处在“表音”层次，而且在这一层次系统性能还有提升的空间；另外，系统还不具备属于更高层次的“表情达意”的能力，还不能通过合成语音准确、生动地传递语义信息.

3　问题分析

3.1　发音质量

发音质量主要涉及音段、超音段、音色及注音方面的问题.

在音段、超音段层面，由于单元挑选策略或样本稀少的原因，被选中音段有可能与合成的语境不匹

配. 而波形拼接的过程一般为简单的平滑处理,无法完全消除语境不匹配的影响,从而影响合成语音的音质,进而会影响合成语音的可懂度、自然度. 音色层面,在构建语音数据库时,常采用朗读风格的录音,而且还需有意识地控制并保持音色、节奏、发音方式的一致性,因而基于单一音色数据库的波形拼接策略只能提供单一音色的合成语音.

由于数据库规模的限制,通过增加音色、音段变体样本的方式虽然有效果,但是改进是有限的. 根本的解决方法是研制高性能语音合成器,使之具备实施高强度的韵律、音色、音质调整的能力.

注音错误涉及了文本处理中的注音功能. 多音字、数字串的读法,目前有效的解决方案是利用基于分词及词性(part-of-speech, POS)信息的统计方法[18-19]. 但该类问题的最终解决,将依赖于正确的语义分析的结果. 例如,例句1中两个"长"字的注音,测试系统都标成了"chang2".

例句1 孩子又长了,胳膊也长了.

原因是,前后两个短句的结构及词性构成一样,而"chang2"是一个高频读音,这样,统计方法就不再有效,只有进行句子的语义类别及配价关系分析,才有可能将"长"确切地标为"zhang3". 如果系统具有上下文语义分析能力,能够判断出话题是有关"孩子试穿衣服",则可进一步将后一个"长"标为"chang2".

3.2 韵律预测

目前,汉语语音合成系统的韵律预测是指韵律层级结构的预测. 尽管为提高预测的准确性提出了多种建模方法,如决策树、有限状态机等[20],但所利用的主要信息还是待合成文本的分词及词性信息. 从原则上讲,分词断句即韵律层级划分,需要保证各韵律单元的语义完整,并准确反映单元间的语义搭配关系. 所以单纯地利用词性信息,常常会带来划分歧义. 例句2为测试系统的韵律层级结构的标注结果.

例句2 中国 农业 科学院 农业|与 农村 发展 研究所|在 当地 设立了|试验 基地 和 观察点.

这里,空格表示韵律词边界,"|"表示韵律短语边界. 对测试系统的处理过程分析表明,对例句2的分词结果没有歧义,词性标注除了"发展"、"试验"被标识为动词外,其它都不存争议. 而最终的标注结果,对于现行的韵律预测模型来说也是合理的. 显然,要解决这类歧义问题,对文本的分析需上升到语义层面.

3.3 表现能力

现有系统韵律模型,主要是在句子层面采用韵律层级结构辅助声调信息,用以刻画韵律的变化. 如此,可以得到语调流畅的合成语音. 但句子内部没有轻重音变化,句子或短语之间没有对比变化,合成语音语调只能是单调的,缺乏表现力. 虽然韵律层级结构也负载了语义结构信息,但还有很多语义信息是通过轻重、节奏、语调及语气等韵律特征传递的,为了合成生动的语音,必须对韵律特征进行更加精细、完备地描述.

重音是实现语义强调的重要手段. 作者基于感知优化的重音检测器,实现了汉语语音数据库的自动标注,并利用重音标注数据库训练得到支持重音的韵律声学预测模型,进而构成支持重音合成的语音合成系统[21]. 但如何由文本分析出重音的位置及级别,需要引入语义处理机制才能解决. 例如,例句3中需要强调的信息是"晴天"而非"阴天",这类重音,只能依据上下文的语义语境来确定.

例句3 明天是个**晴天**!

上下文语义信息的利用,意味着对文本的分析范畴将超出句子层面.

除了重音之外,功能语调如疑问语调对语义表达的完整性也非常重要,来自应用层面的需求也非常紧迫. 因此,合成语音表现能力的提高,需优先解决重音和疑问语调的实现. 此外,节奏、音色与语义、情感的表达也有密切关系,对改善合成语音的表现能力也非常重要,只是情感的语音合成将更具挑战性[22].

总之,解决系统的现有问题,提升合成语音质量,一方面需改进韵律模型,实现对韵律特征更细致的刻画,使合成语音更加生动. 更重要的是,需在文本分析阶段引入语义分析,解决多音字注音问题,克服韵律结构划分中的歧义,并利用句子及其以上层面的语义信息,完成语义焦点、语调结构类型,乃至语气的确定,从而使合成语音能够更为完整、准确地表达语义信息.

4 新一代语音合成系统展望

以清晰、自然的合成语音,准确、生动地传递语义信息是新一代语音合成系统所追求的目标. 为实现这一目标,将涉及到理论基础、技术实现及基础资源等方面的研究.

4.1 理论基础——语义的言语实现

实现既定的系统目标，依赖于有效的语义-语音计算模型，这就需要在理论层面研究语义的言语实现机制．首先，需要确定语义的形式化描述方案．到目前为止，几乎所有的汉语语义描述方案都是面向自然语言处理的，主要用于机器翻译[23-24]，但对于语音合成仍有借鉴作用．由于韵律特征与语义之间有着紧密的关系，所以可以借助韵律分析来确定面向语音合成的语义分类颗粒度、属性及关系的描述方案，同时构建语义-韵律关系理论体系．

4.2 技术实现——基于语义的语音合成

基于语义的语音合成需要实现 3 个关键性功能模块：语义分析、韵律预测、高性能语音合成器．

语义分析模块　实现由系统输入文本到形式化语义表达的分析功能．所利用的基本信息将包括待合成文本的分词、词性以及对应的词义．对词义的标注需要借助于语义词典．模块将利用数据驱动建模方法，基于语义标注数据库构建文本的语义标注模型；预期中间的计算过程将涉及或反映语句结构、新旧话题、语义焦点、句型、情态等内容，最终输出的是形式化语义．

韵律预测模块　实现由语义分析到韵律特征的预测．模块的输入除了语义分析输出，还包括一般文本分析器输出的分词、词性标注及其注音，以保持和现有系统在技术上兼容．模块输出的是有关发音的、细化的韵律描述，预期包括韵律结构、语义重音、语调类型、发音方式等特征．

高性能语音合成器　由于输出的语音要求声音清晰、语调准确、语气生动，纯粹的基于单元挑选波形拼接技术不再适用．因此，就对语音合成器性能提出了更高要求，即能够在保持声音清晰、自然的前提下，准确、生动地实现各种语调和语气，这就意味着该合成器不仅具有超音段参数调整能力，同时还需具备音色调整能力．

4.3 基础资源——语义词典与数据库构建

语义词典是进行语义分析的基本资源．现有的汉语语义词典[25]主要是面向自然语言处理任务构建的，在确定语义的形式化描述方案之后，将根据语音合成的需要进行适当地调整与改动．

数据驱动建模策略是实现系统目标的一种合理的技术选择．一定规模的带有丰富标注信息的数据库，一方面将为理论分析提供真实样本，进而可望获得对系统建模的指导意义；同时，也为基于数据驱动建模方案提供训练及测试数据．因此，数据库的构建及质量对于系统目标实现是至关重要的．在数据库构建过程中，将根据使用目的，确定收集数据的类型、内容和规模，之后是数据处理工作，其核心是各种类型标注．预期将包括以下条目．语言部分：文本、分词及词性、语义标注；语音部分：注音、音段切分、韵律标注．

5 结束语

提出了新一代语音合成系统的基本概念——基于语义分析的语音合成系统．其功能将包括：基于文本语义分析，实现与语义相关联的韵律结构、语义重音、功能语调、发音风格等细致的韵律特征预测，通过高性能合成器输出合成语音．其目标是准确、生动地传递语义信息．准确是指语义表达完整、精确；生动是指表达方式的多样性和恰如其分．其核心是语义的体现．

参考文献：

[1] Black A, Campbell N. Optimising selection of units from speech databases for concatenative synthesis [C] // Proceedings of Eurospeech 1995. Madrid, Spain: [s. n.], 1995: 581 - 584.

[2] Black A. Perfect synthesis for all of the people all of the time [C] // Proceedings of IEEE TTS Workshop 2002. Santa Monica, USA: [s. n.], 2002: 167 - 170.

[3] Chu Min, Peng Hu, Chang Eric. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer [C] // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2001. Sale Lake City, USA: [s. n.], 2001: 785 - 788.

[4] Li Wei, Lin Zhenhua, Hu Yu, et al. A statistical method for computing candidate unit cost in corpus based Chinese speech synthesis system [C] // Proceedings of International Conference on Chinese Computing 2001. Singapore City, Singapore: [s. n.], 2001.

[5] Silverman K, Beckman M, Pitrelli J, et al. ToBI: a standard for labeling English prosody [C] // Proceedings of International Conference of Spoken Language Processing 1992. Banff, Canada: [s. n.], 1992: 867 - 870.

[6] Wightman C. ToBI or not ToBI? [C] // Proceedings of International Conference on Speech Prosody 2002. Aix-en-Provence, France: [s. n.], 2002: 25 - 29.

[7] Mixdorff H. Speech technology, ToBI and making sense of prosody [C] // Proceedings of International Conference

of Speech Prosody 2002. Aix-en-Provence, France: [s. n.], 2002:31 - 37.

[8] Black A, Hunt J. Generating F_0 contours from ToBI labels using linear regression [C]// Proceedings of International Conference of Spoken Language Processing 1996. Philadelphia, USA: [s. n.], 1996: 1385 - 1388.

[9] Qian Yao, Chu Min. Prosodic word: the lowest constituent in the mandarin prosody processing [C]// Proceedings of International Conference on Speech Prosody 2002. Aix-en-Provence, France: [s. n.], 2002: 591 - 594.

[10] Ostendorf M, Bulyko I. The impact of speech recognition on speech synthesis [C] // Proceedings of IEEE TTS Workshop 2002. Santa Monica, USA: [s. n.], 2002: 99 - 106.

[11] Tseng C, Chou F. A prosodic labeling system for mandarin Chinese speech database [C]// Proceedings of International Congress of Phonetic Sciences 1999. San Francisco, USA: [s. n.], 1999: 2379 - 2382.

[12] Zhu Weibin, Zhang Wei, Shi Qin, et al. Corpus building for data-driven TTS systems [C]// Proceedings of IEEE TTS Workshop 2002. Santa Monica, USA: [s. n.], 2002: 199 - 202.

[13] Yang Yufang, Wang Bei. Acoustic correlates of hierarchical prosodic boundary in mandarin [C]// Proceedings of International Conference on Speech Prosody 2002. Aix-en-Provence, France: [s. n.], 2002: 707 - 710.

[14] Ma Xijun, Zhang Wei, Zhu Weibin, et al. Probability prosody model for unit selection [C] // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004. Montreal, Canada: [s. n.], 2004:649 - 652.

[15] Shi Qin, Ma Xijun, Zhu Weibin, et al. Statistic prosody structure prediction [C] // Proceedings of IEEE TTS Workshop 2002. Santa Monica, USA: [s. n.], 2002: 155 - 158.

[16] Black A, Tokuda K. Blizzard challenge-2500: evaluating corpus-besed speech synthesis on common datasets[C]// Proceedings of Eurospeech 2005. Lisbon, Portugal: [s. n.], 2005: 77 - 80.

[17] Liu Qun, Wang Xiangdong, Liu Hong, et al. Introduction to HTRDP evaluations on Chinese information processing and intelligent human-machine interface [J]. Frontiers of Computer Sciences in China, 2007, 1(1): 58 - 93.

[18] Zheng Min, Shi Qin, Zhang Wei, et al. Grapheme-to-phoneme conversion based on TBL algorithm in mandarin TTS system [C]// Proceedings of Eurospeech 2005. Lisbon, Portugal: [s. n.], 2005: 1897 - 1900.

[19] Zhang Zirong, Chu Min. An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese [C]// Proceedings of International Symposium on Chinese Spoken Language Processing 2002. Taipei, China: [s. n.], 2002.

[20] Shi Qin, Zhang Wei, Ma Xijun, et al. Comparisons among four statistics based methods of prosody structure prediction [C] // Proceedings of National Conference on Man-Machine Speech Communication 2003. Xiamen, China: [s. n.], 2003.

[21] Zhu Weibin. Perceptual optimization of the Chinese accent-index detector [C] // Proceedings of International Conference on Speech Prosody 2006. Dresden, Germany: [s. n.], 2006.

[22] Tao Jianhua. Emotion control of Chinese speech synthesis in natural environment [C] // Proceedings of Eurospeech 2003. Geneva, Switzerland: [s. n.], 2003: 2349 - 2352.

[23] Dong Zhendong, Dong Qiang. Resolutions to some problems in building Chinese lexical semantic resources [C]// Proceedings of 4th Chinese Lexical Semantics Workshop. Hong Kong, China: [s. n.], 2003.

[24] Wu Yunfang, Pei Yulai, Zhang Yangsen, et al. A Chinese corpus with word sense annotation [C]// Proceedings of International Conference on the Computer Processing of Oriental Languages 2006. Singapore City, Singapore: [s. n.], 2006.

[25] 王惠,詹卫东,俞士汶. 现代汉语语义词典规格说明书[J]. 汉语语言与计算学报,2003, 13(2): 159 - 176.
Wang Hui, Zhan Weidong, Yu Shiwen. The specification of the semantic knowledge-base of contemporary Chinese [J]. Journal of Chinese Language and Computing, 2003, 13(2): 159 - 176. (in Chinese)

(责任编辑:赵秀珍)