1998 年 第 38 卷　　清 华 大 学 学 报 (自 然 科 学 版)　　22/28
Journal of Tsinghua University (Sci & Tech)　　第 S1 期 第 92～95 页

文语转换系统中汉语韵律的学习和模拟*

蔡莲红,　张　维,　胡其炜

清华大学 计算机科学与技术系, 北京 100084

文　摘　在语音合成的研究中, 韵律表示和韵律学习是改善合成语音质量的关键。本文作者研制了汉语文字-语音转换系统 Sonic。在研究韵律描述和模型的基础上, 设计了韵律控制符号和韵律模拟算法; 通过对文本进行置标, 实现了重音和语调的模拟。为了进一步改善合成语音的自然度和表现力, 作者改造了 Sonic 系统, 采用神经网络学习算法, 按词语和语句两级进行韵律学习, 建立了具有韵律学习能力的文语转换(TTS)系统 Sonic-L。本文分析了汉语 TTS 技术研究的现状, 介绍了作者在韵律学习、描述、模拟方面的研究工作和实验结果。

关键词　文语转换(TTS); 汉语文字-语音转换; 韵律描述; 韵律学习

分类号　TP 391.42

计算机语音输出的研究属于人工智能的一个分支。它以信号处理的理论和方法为基础, 以自然语言理解为依托。当前计算机语音输出的研究重点是文字—语音转换(TTS)。TTS 是在语音合成技术的基础上, 增加了语言学处理、韵律模拟等处理, 其目标是输出连续自然的语声流。

与人朗读文本相比, 人具有先验知识、懂得语法、能够理解文本的内容, 在朗读时, 会将自己的意向、情感, 通过声音的形式体现出来, 声情并茂、有声有色。而文字并没有直接提供有关语音的信息, 因此, 对于一个 TTS 系统来说, 不但要解决"说什么"的问题, 而且更重要的是解决"怎样说"的问题。

本文作者已实现了一个汉语 TTS 系统 Sonic[1], 它可运行在 DOS, Windows, Unix 等操作系统下, 输出语音清晰可懂。在研究了汉语声调、重音、语调的声学特性的基础上, 设计了韵律控制符, 进行了重音和语调的模拟[2,3]。

收稿日期: 1997-09-30

第一作者: 女, 1945 年生, 教授

* 国家"八六三"高技术项目, 863-306-3-2-4

尽管 Sonic 已具有较好的清晰度, 但自然度有待提高。尽管规则较好地处理了协同发音的某些问题, 但它毕竟是从特定数据中得到的。而人类的语音是因人而异, 因境而别, 有必要使 TTS 系统具有学习能力, 不断从自然语音中学习, 优化韵律规则, 完善韵律描述, 进而实现个人发音特点的模拟。

为此作者改造了 Sonic 系统, 新系统 Sonic-L 的框图如图 1 所示。Sonic-L 系统分为韵律学习和语音合成两部分。韵律学习的输入是自然语音和经过文本分析程序处理过的文本。韵律学习的输出是韵律描述或韵律模拟之参数。

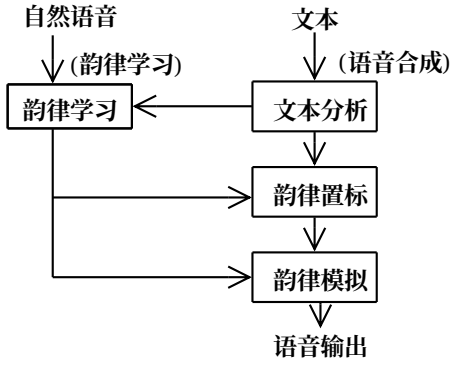

图1　Sonic-L系统框图

语音合成由三部分组成: 文本分析、韵律置标和韵律模拟。文本分析包括文本规范化、分词、数字和符号处理、多音字处理、句型分析、语法分析等。韵律置标是在文本中插入一些韵律控制符号。韵律模拟是带有参数或波形修改能力的语音合成。

1　TTS 系统中的汉语韵律描述

1.1　文字到语音的映射

语言作为社会交际的符号系统, 是人类的思维工具, 语音是语言(文字)的物质外壳, 是语义的物质载体。TTS 系统要实现文字到语音的映射, 以语音方式表达人的思维。TTS 系统需要在对文本理解

的基础上, 完成字位到音位的转换(grapheme-to-phoneme)、音位到语音的转换(phoneme-to-speech)。汉语在实现上述转换时会遇到许多困难。例如, 不能准确地决定句子中的停顿位置和长短、轻重音、语调, 不能根据人的表达需要, 给出语音感知特性的细微描述。合成语音不能完全仿真自然语音。这里除算法的缺陷外, 关键是机器缺乏知识, 没有学习说话的能力。

1.2 韵律标注

自然语言中, 语音特征变化万千, 其数据本身隐含了知识。而对这些知识, 人类可以感知, 但对其的认识、描述是远远不够的。用国际音标可对文字进行注音, 但它不能满足语音合成、识别及语音学研究的需求。如, 它不能提供精细的韵律描述。而韵律描述是提高合成语音的自然度和表现力的前提。

在韵律描述方面, ToBI[5]是当前国际上最为流行的韵律标注系统。用它标注美国英语的语调模式和一些韵律现象。韵律标注分层次进行, 每一层都设有表示韵律事件、事件相伴时刻的符号。标注分音调层(tone tier)、断引层(break index tier)、表音层(orthographic tier)和杂类层(miscellaneous tier)。ToBI能够描述自然语音的最重要的韵律特征, 易学易用。它不是一个封闭的系统, 使用者可以根据研究需要增加或调整标注项目。其它语言的研究者参考ToBI设计了各自的韵律标注系统, 如德语、瑞典语、日语等。

汉语普通话标注研究已初步展开。如吴宗济提出的"韵律标注文本"[6], 他从声学语音学考虑, 试图为普通话语音合成设计一套符号。李智强博士从音系学和语音结合的角度出发, 初步设计了普通话韵律标音系统, 并进行了初步的实验[4]。该系统分4个音层标注: 拼音层、声调语调层、重音层、韵律结构层。

1.3 韵律置标和控制

韵律标注是对已录制的自然语音标上一些符号, 以说明其发音的特点。而从语音合成来说, 首先将一些标注符号插入到文本中, 称之为置标。合成时再还原成参数, 控制合成的声音, 达到修饰或模拟的目的。

在以往的文语转换研究中, 已使用了某些韵律控制符, Microsoft公司推荐了一组控制语音符号[7], 用于控制音高、重音、时长等。如Spd——设置讲话的速度。这组符号可以插入正文间或由程序自动输入。

针对汉语的特点, 作者设计了一组韵律置标符号[7]。该组符号已用在Sonic系统中, 用于控制合成语音的特性, 还实现了重音、语调的模拟。

为了规范韵律置标, 方便计算机对音频媒体进行控制。作者初步设计了音频媒体控制语言(audio media control language)。AMCL是一种基于数据流的标注语言, 用于描述音段、旋律及复杂语音特性。AMCL允许用户对合成语音(TTS)、波形(WAVE)和音乐(MIDI)以及其它音频数据灵活组合, 用一个或多个文档对各种音效进行描述, 在AMCL播放器中播放出来。通过AMCL可以描述复杂语音模型和韵律规则, 也能够控制各个音源的空间和时间关系, 以获得复杂的、富有表现力的音效效果。

AMCL的语言特性及设计目标如下:

平台无关性: AMCL独立于计算机类型、发音设备、操作系统、TTS引擎和网络设施。

简单性: 对于人, 它是易于理解和操作的; 对于计算机, 它是易于分析和实现的。

一致性: AMCL及其扩展均不能出现导致二义性的描述。

全面性: AMCL必须足以描述所有常用的和必需的语音特性及音频模式。

可扩展性: 在固定的、统一的语法结构下, 易于增加新的功能, 并使新的功能同样具有简单性和一致性。

2 韵律的学习与模拟

语音合成技术的进步以及迫切的应用需求, 期望TTS系统具有更为优良的语音合成规则和语音模型, 以使得合成语音更平滑、更自然、更具表现力。而由于语音的复杂性和多变性, 目前, 没有也不可能有一成不变的可计算的语言和语音模型。这便要求语音合成的研究者自己解决这一矛盾, 赋予TTS系统韵律学习的能力正是基于此而提出的。

韵律学习就是借助大量的语料和已有的规则、知识, 通过机器学习的方法, 优化旧的规则和知识, 抽取新的规则和知识, 达到改善合成语音的目的。

2.1 学习方案

赵元任曾对声调和语调有过生动的比喻[8], 他说: "字调加在语调的起伏上面, 很象海浪上的微波, 结果形成的模式是两种音高运动的代数和。"他还说: "…(字调)仿佛橡皮带(音域)上的图形会随

着橡皮带的伸展而放大。”实际上, 在语音特征的许多方面, 诸如时长、幅度、停顿等, 都有类似“海浪”或“橡皮带”的性质。因而, 在学习方案的设计上, 分为词语和语句两级是比较合理的, 词语一级的学习, 主要学习调型、频谱、时长和字间停顿的调节, 而语句一级的学习, 主要学习音域、时长、幅度和词间停顿的调节。然后将两者的学习结果根据“海浪”或“橡皮带”性质综合起来, 共同控制语音的合成。

2.2 学习算法的说明

语音的特征复杂多变, 且分类数的规模难于估计, 故采取模糊分类的策略。这里学习算法采用B-P网络[9]。该算法属于有教师学习的δ学习律, 具有非线性的特点, 收敛特性比较好, 适合于语音的学习。

1) 网络设计

采用三层的B-P网络。输入层和输出层的节点数由对应网络的需求决定, 针对语音现象的分布情况复杂的特点, 适当增加中间隐层的节点数, 其代价是增加学习的时间。

约定节点输入与输出的非线性变换函数为

$$f(u)=\frac{1}{1+e^{-u}}$$

输入向量$X=(x_1, x_2, \cdots, x_I)$, 中间隐层的输出向量为$X'=(x_1', x_2', \cdots, x_m')$, 输出层向量为$Y=(y_1, y_2, \cdots, y_n)$, 输入节点$i$到中间隐层节点$j$的权值为$W_{ij}$, 中间隐层节点$j$到输出节点$k$的权值为$W'_{jk}$, 教师向量为$T=(t_1, t_2, \cdots, t_n)$。

2) 教师向量的归一化处理

由于教师向量是从自然语音数据中提取的参数, 而网络输出$y_k\in[0,1]$, 因而将从自然语音中提取的参数根据各自的上下限归一到$[0,1]$区间作为教师向量, 而上下限由统计或经验得到。

3) 误差的计算和处理

鉴于语音特性是一个模糊集合, 如时长等参数, 并不是一个确定的值。对人耳来说, 也有一个接受的范围; 而且人耳对不同的特性的敏感程度不一样, 如对声调的敏感程度大于幅度。考虑这些特点, 将B-P算法的误差计算公式作相应修改。

样本P_s的学习误差为

$$E_{P_s}=\frac{1}{2}\sum_{k=1}^{n}W_k(t_k^{P_s}-y_k^{P_s})^2$$

其中, W_k为第k个误差分量的权值, 此权值由经验决定, 下同。

对P个样本的学习, 总误差为

$$E=\frac{1}{2}\sum_{P_s=1}^{P}\sum_{k=1}^{n}W_k(t_k^{P_s}-y_k^{P_s})^2$$

算法退出条件为

$$|t_k^{P_s}-y_k^{P_s}|<W_k\epsilon\ (\forall P_s, \forall\ k)$$

实验中, 适当放宽误差的接受范围(ϵ取较大的值), 这样不但考虑到语音的特性, 而且附加的效果是加快了网络学习的速度。

2.3 学习材料的准备

1) 语料设计

初始学习语料包括音节、词语、语句; 采用同一个人的录音; 为减少主观因素带来的不良影响, 音节和词语均取自“负载句”。经验证明, 这能保障语料的质量。词语、语句的类型覆盖全面, 分布较均匀。

2) 标注

在进行学习之前, 先标注收集到的语料。如拼音、声调、词语边界、语句成分、句型、语境等。标注时, 采用人工和机器相结合的策略。以机器标注为主, 人工校对, 以防学习错误。经过标注的语料, 与其相应的标注一起存入语料库。

2.4 词语一级的学习

词语一级构造两个网络, 一个主要学习时长、停顿等音段特性的调节; 另一个主要学习基频曲线、幅度包络和共振峰等细微特性的调节。

音段特性学习网络的输入向量X由前后音节的拼音码编号、当前音节处于词语中的位置、音库单音的时长; 教师向量T由归一化的自然语音中当前音节的时长构成。

将合成音库的音节按音段特性学习网络的学习结果动态修改时长, 然后和语料库的音节分为每20 ms一帧, 帧移为10 ms, 每帧提取的参数作为细微特性学习网络的输入。

细微特性学习网络的输入向量X由前后音节的拼音码的编号、当前音节处于词语中的位置、音库单音的时长、当前帧在音节中的相对位置、当前帧的基频、当前帧共振峰信息、当前帧的幅度构成; 教师向量T由归一化的自然语音对应帧的基频、幅度、共振峰信息构成。

2.5 语句一级的学习

语句一级构造一个网络, 主要学习音域、时长、幅度等句子一级的音段特性的调节。其输入向量X由当前语句的句型编号、前后词语的拼音码的编号、当前词语在句中充当的成分构成, 教师向量T由归一化的自然语音中对应词语的音域、幅度域、时长、

与前一词语之间的停顿构成。

2.6 参数综合与韵律模型

根据"橡皮带"效应, 对两层的学习结果进行综合和模拟, 具体举例如下。

1) 字调与音域的综合: 音域决定音节的基频上下限, 而字调决定基频曲线的调型, 两者便形成了一宽窄高低都在变化的音域序列, 字调便为限于此序列上下限的曲拱, 整个的曲拱顺连成基频曲线; 由此基频曲线控制语音合成的基频调节。

2) 幅度的综合: 类似字调与音域的综合。

3) 时长的综合: 词语一级学习的时长结果决定各个音节在词语中的相对时长, 语句一级学习的时长结果决定词语总的时长, 两者的综合便得到每一音节在句中的时长。

4) 停顿的综合: 停顿的综合较简单, 由词语中字与字之间的停顿及词语之间的停顿分别控制语音的节奏。

5) 频谱的修改主要集中于词语中音节相接部分的修改。

3 结束语

韵律学习和韵律模拟已成为提高合成语音质量的关键, 也是 *TTS* 系统的研究方向。*Sonic-L* 中的学习算法分为短语和语句两级, 而短语一级又分为两级。网络学习速度快, 且具有很大的灵活性。

具有学习能力的 *TTS* 系统也可扩展为说话人模拟系统, 主要把学习算法中做为教师的语料改为需要模拟人的语料, 同时提供再多一些的特征参数。

参 考 文 献

1 蔡莲红, 刘 灏, 周俏峰. Windows下汉语文-语转换系统的设计与实现. 微型计算机, 1995, 15(3): 10～12

2 Zhou Qiaofeng. Simulation of stress in Chinese TTS system. In: Key-sun Choi eds. NLP'95. Seoul Korea. 1995. 532～537

3 蔡莲红, 周俏峰. 基于PSOLA的汉语TTS韵律修改算法. 软件学报, 1996, (10): 33～38

4 李智强. 普通话韵律标音系统的初步设计. 见: 吴泉源, 钱跃良编. 智能计算机接口与应用进展. 北京: 电子工业出版社, 1997. 169～173

5 Silverman K. ToBI: a sandard for labeling English prosody. ICSLP, 1992. 867～870

6 吴宗济. 用于普通话语音合成的韵律标记文本的设计. 见: 第三届全国语音研讨会论文集. 北京: 中国社会科学研究院, 1991. 27～29

7 蔡莲红, 罗 恒. 文语转换系统韵律置标方法的研究. 软件学报, 1996, 17(增刊): 514～518

8 赵元任. 中国现代语言的开拓与发展. 北京: 清华大学出版社, 1992

9 张立明. 人工神经网络的模型及其应用. 上海: 上海复旦大学出版社, 1993

Prosody learning and simulation for Chinese text to speech system

CAI Lianhong, ZHANG Wei, HU Qiwei

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract In the research of speech synthesis, prosody expression and study are the kernels of improving performance of speech synthesis. The authors established a Chinese text to speech (TTS) system, Sonic. The researchers have designed prosody symbols and algorithm, basing on studying prosody descriptions and models, then marked up the text with prosody symbols, and implemented simulation of stress and intonation. To get the better naturalness and expressivity of the output speech, the system have been improved and a prosody learning TTS system, Sonic-L, have been developed. In sonic-L, learning with neural network is splitted two levels: based on words and based on sentences. The article introduces the actuality of technology of text to speech, describe the author's research and result on prosody studying, description and stimulation.

Key words text to speech (TTS); Chinese text to speech; prosody description; prosody learning