www.ccw.com.cn
周报全文
首页 | 时讯与评论 | 知识中心 | 管理与职场 | e海航标 | I社区 | 计
2003年2月19日 综合新闻 / 产品与技术 / 应用与方案 / 网络与通信 / 市场与

神经网络与汉语TTS韵律模型

陶建华 蔡莲红

韵律模型

每个人说话的语音中都有一个基本频率，被称做基频，它体现了说话人声音的高低。在汉语文语转换系统（TTS）中，对基频、语音单元的长度、说话停顿、能量等韵律信息进行预测的模块一般称做韵律模块。

众所周知，汉语是一个有调的语言，这是它与其他西方语系最大的不同之处。汉语的每一个字（儿化音除外），通常都被认为是一个有调的音节。每一个声调都有一些固定的调型（基频形状），但我们通常所说的话往往是由多个字组成的连续语句，这些声调的调型受相邻其他字或词的影响，常常会产生变换，甚至失去原有的调型，这就是汉语中常说的协同发音现象。这也就是为什么人说话时会有连续感，而不是一个字一个字地发音。同时，连续语句发音的中间还会有短暂的停顿，这些又体现了人说话的节奏感。汉语TTS韵律模型的主要任务就是根据文字中的信息，通过对基频、音长、停顿等参数的预测，达到控制TTS系统发音方式的目的，使发音自然、好听。

采用神经网络模型的背景

随着语音学和计算机技术的发展，TTS系统的研究目前已获得了重大进展，并成功地应用在许多不同的场合。但是，以往语音合成的结果与人自然流畅的发音仍相去甚远，其中的关键就在于语音韵律模型还不很完善。另外，人有思想、会思考，语音合成系统不仅应该发音清晰、自然，还应该能像人一样具有自我学习的功能，具有个人特色，甚至具有模拟特定人发音的能力。

近几年来，随着计算机处理的进一步深入，从大量语料中提取连续语句的韵律特征已逐渐成为可能。鉴于神经网络具有良好的自动学习和参数映射的特点，可以使系统具有不断自我学习和输出优化功能，因此，将神经网络用于语音合成系统的研究越来越受到重视。研究结果表明，对比传统的规则语音合成方法，运用神经网络技术合成的语音的自然度均得到了相当程度的提高。

清华大学计算机系在国内最早进行了神经网络用于汉语TTS系统的研究，目前已经取得了非常成功的结果。所提出的带特殊加权因子的神经网络韵律模型，无论在提高TTS系统自然度方面，还是在执行效率上，相比较其他已有的模型，均获得了较大的提高。

清华大学计算机系对人机语音交互的研究始于1979年，并长期致力于语音合成的声学模型、韵律模型、文本分析、韵律描述语言、语音数字编码、多媒体等相关技术的研究和开发。下面介绍由清华大学计算机系人机交互与媒体集成研究所提出的汉语TTS系统神经网络韵律模型。

神经网络韵律模型的输入和输出

构筑神经网络韵律模型必须首先解决模型的输入和输出问题。对TTS系统来说，系统的输入就是

从计算机屏幕或文件中得到的文字，输出则是连续语音。因此，神经网络韵律模型的输入必须是与文字相关的信息，通常称其为语境信息，而输出则是与语音相关的韵律信息。

正如前面所述，当汉语中多个字组成词或词组而连续发音时，它们之间会相互影响，形成较独立、完整的韵律块，这些韵律块的韵律特征对语音的自然度起着非常重要的作用，而不同的韵律块组合在一起，往往可以形成不同的语调，使人的发音具有不同的语气。根据这样的思路，可以将汉语的文字信息沿着语句（sentence）→短语（phrase）→音节（syllable）的思路划分，共分为五组：音节（字）信息、相邻音节（字）信息、短语信息、语句信息及重音信息。有17个参数能对汉语韵律产生重要的影响，这些参数就是神经网络韵律模型的输入。当然，这些参数都能够从文字中得到，但必须辅以另外的文本分析模块。

神经网络的输出就是汉语韵律控制参数。在基频方面，使用SPiS模型，如图1所示。

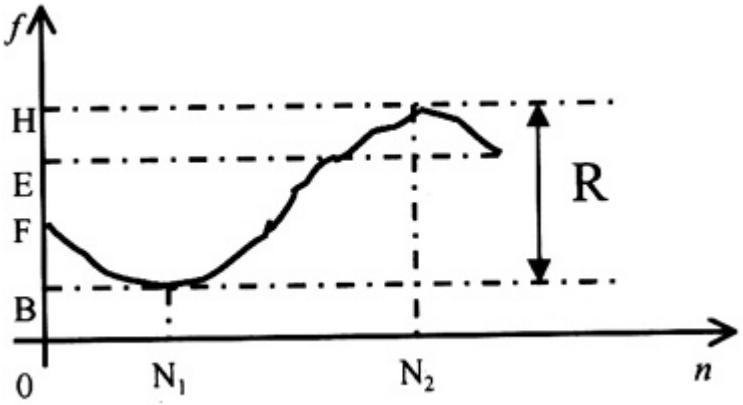

图1 汉语音节基频规格化模型－－SPiS

神经网络的结构

神经网络的拓扑结构如图2所示，基本可以分为三层，即输入层（语境标注矢量层）、输出层（韵律控制矢量层）和中间隐层。

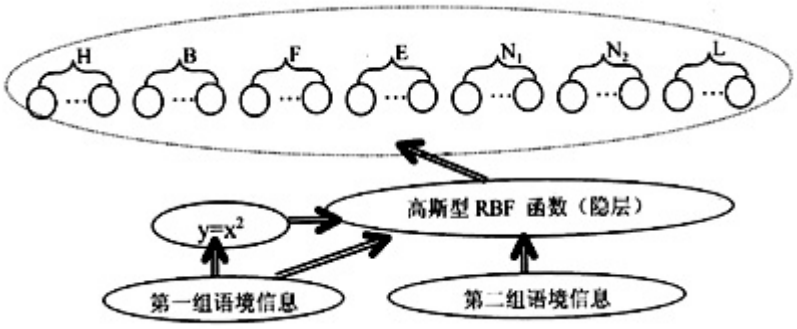

图2 韵律神经网络模拟

语音学的研究表明，汉语较其他语言更强调文字发音的轻重和语气的走势。前面所述的模型输入参数（语境参数）被分为两组，同时在其中一组上加入一个特殊的加权隐层，以突出改组的权重，该隐层的神经元函数为:y=x2。

测试结果证明，加权隐层的引入使网络结构进一步体现了汉语独特的韵律特点，使网络的收敛速度在原有的基础上提高了约18%，从而较大地改善了网络的收敛性。同时，在模型的建立中，还利

用概率分布的原理，采用输出离散化并取其质心的方法，对神经网络的输出进行优化，使网络的输出精度进一步提高了约7%，从而增强了网络输出值的稳定性，最大限度地减少了因输入和输出参数的随机特性而导致的输出误差。

结果分析

1．可训练汉语TTS系统

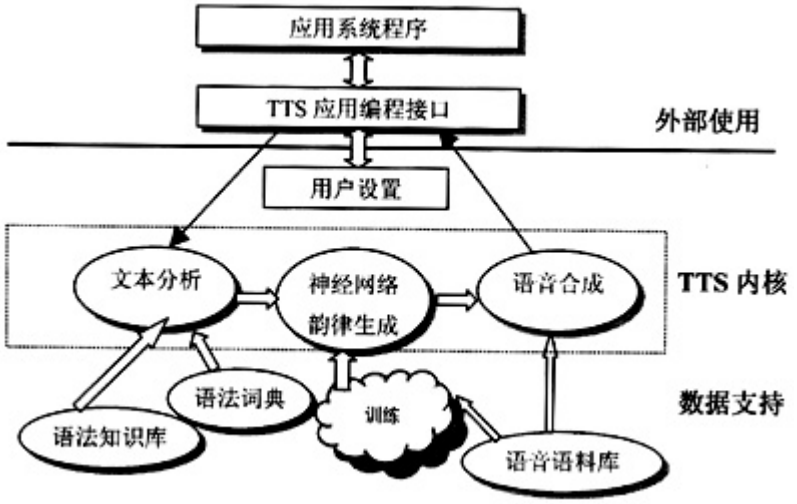

图3 可训练汉语TTS系统结构

图3给出了完整的可训练汉语TTS系统结构。

系统的构成分为用户编程接口和TTS内核两大部分。其中，内核部分又可按照系统运作的不同过程分为多个子模块，包含了训练模块、文本分析模块、神经网络韵律生成模块、语音合成模块以及与语料库之间的通信协议等。同时，系统还考虑了不同类型用户对TTS系统功能的需要，提供了丰富的编程接口。

系统使用了2270个句子分别对模型进行了训练和测试。语句内容涵盖了汉语中常见的句型、汉语中所有的读音、文字上下文的特性、声调、重音等信息。语音的采样频率为16kHz。其中，75%的语料用来进行训练，而25%的语料则用来测试。

2．基频控制参数(SPiS参数)的测试结果

图4 陈述句基频曲线的测试结果

韵律模型的基频输出基本反应了汉语语句的韵律特征。由图4可以看出，其基频参数的测试结果与真实的基频参数比较接近，基频变化过程基本保持了陈述语气的下倾趋势，同时它还反映出了发音过程的韵律块特性。如图中陈述句“他总标榜自己是一个老手”，受发音停顿的影响，“是”作为一个韵律短语的开头，其基频和音域变得相对较高。另外，神经网络韵律模型还能很好地反映上声变调的现象。如“老手”中的“老”字，受后音的影响，由上声变为了阳平。

3．连续语句中音长参数的测试结果

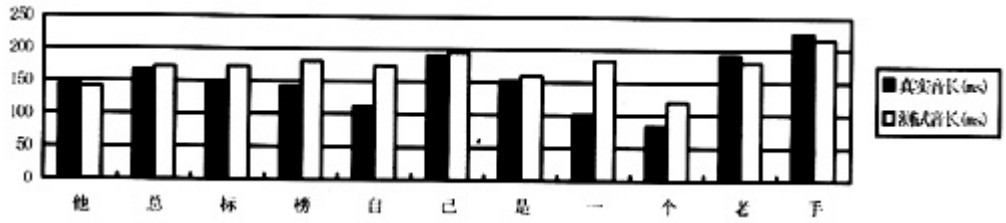

图5 陈述句音节音长参数的测试结果

神经网络韵律模型同样输出了较好的音节音长参数，图5很好地反映出了语句音长的变换趋势。由于在自然语句中，音节音长参数对控制音节发音的节奏和轻重起着非常重要的作用。我们对所有测试结果进行的统计表明，81%的音节输出误差在0～50ms，约14%的音节输出误差在50～120ms，而只有约5%的音节输出误差会超过120ms。从音长改变的百分比上看：89.8%的音节，其音长输出误差占目标音长的百分比在0～20%之间；另外，9%的音节输出误差百分比在20%～50%之间，而只有1.2%的音节输出误差百分比会超过50%。因此，该模型的音长参数输出结果基本上满足了较高质量韵律控制参数的要求。

将神经网络模型与已有的TTS系统相结合，改变了传统的TTS系统的构筑方式。新系统合成语音的自然度得到了提高，同时也使语音合成系统中的韵律模型具有更强的适应性和可训练性。新系统经过学习和训练，合成的语音能体现不同的韵律特征，增加了系统的灵活性和风格的多样性。大量测试表明，汉语神经网络韵律模型及其输出参数的优化方法，能适于汉语韵律特征的处理。目前，这一模型已集成在清华大学计算机系研制的语音合成系统中，输出了较为满意的合成语音，其输出的语音自然度在相当程度上几乎可以和自然语音相比，整体水平上达到了国际先进水平并获得专家和用户的一致好评。