第 28 卷 第 3 期
2002 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 3
May, 2002

研究简报

# 汉语文语转换系统中基于小波神经网络的韵律信息合成

吴 震 万 千 陈小平
(中国科学技术大学计算机科学系 合肥 230027)
(E-mail: ddragon@mail.ustc.edu.cn)

关键词 小波神经网络, 韵律信息, 韵律模型, 汉语文语转换
中图分类号 TP391

# SYNTHESIS OF PROSODIC INFORMATION FOR CHINESE MANDARIN TEXT-TO-SPEECH USING A WAVELET NEURAL NETWORK

WU Zhen WAN Qian CHEN Xiao-Ping
(*Department of Computer Science, University of Science and Technology of China, Hefei* 230027)
(E-mail: ddragon@mail.ustc.edu.cn)

**Key words** WNN, prosodic information, prosodic model, chinese TTS

## 1 引言

文语转换系统(text-to-speech, TTS)的研究是一个涉及多个学科领域的课题, 它包括文本处理和语音合成两个组成部分. 文本处理部分将输入的任意文本进行分析后, 产生表征发音信息的系统内部代码; 语音合成部分则根据这些内部代码, 按照一定的韵律模型(也称控制规则), 产生语音数据.

要使文语转换系统能够产生接近自然语言的语音效果, 建立完备的韵律模型是其最关键所在[1, 2]. 与常规的韵律产生方法不同, 我们提出一种基于小波变换的神经网络, 来产生语音合成需要的韵律信息. 网络分为输入层、隐层和输出层, 以小波函数作为神经元的基本函数, BP 算法作为基本训练方法. 输入参数是待处理文本的语言特征, 它包括音节信息和关于音节在文中的环境信息. 这样, 将二者同时考虑, 输出符合自然语言特点的韵律参数. 本文提出了将各种韵律信息的获取问题同步解决的方案, 它通过一个简单的输入、输出网络将所有的韵律参数同时得到. 通过该网络, 汉语语言的韵律模型可以自动建立, 韵律特征则反映在网络的权值上. 本网络的输入参数由先行的文本分析给出, 且相对较为简单, 避免了词性判

收稿日期 1999-08-02 收修改稿日期 2000-01-20

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

定等繁杂的语法分析

Chen S H 等在 1998 年利用一个四层循环的神经网络（RNN）来得到合成语音所需的韵律参数[4], 也取得了较好的结果 与 Chen 的工作相比, 本文在传统的神经网络中引入了功能强大的小波函数, 充分利用了它的时频局部化性质, 使得网络收敛速度加快, 预期值与实测值具有更好的一致性 本文将音节信息和音节上下文信息一起作为网络输入、直接输出所需的韵律参数 这充分应用了系统文本处理阶段的结果, 简化了网络结构, 改善了网络的收敛性质 Chen 的韵律模型只考虑了词层次上的语言特征 而我们以语音词、呼吸群和节奏等重要的语言特征为基础, 可望达到更好的语音合成效果 将小波神经网络运用到文语转换这一领域, 国内尚未有文献报道

## 2　小波神经网络简介

小波变换定义为 $w_f(x,y)=\int_{-\infty}^{\infty} f(t)\,h_{x,y}(t)\,\mathrm{d}t$, 其中 $h(x,y,t)=\frac{1}{\sqrt{|x|}}h_{\mathrm{basic}}\left(\frac{t-y}{x}\right)$ 称为小波, 而 $h_{\mathrm{basic}}(t)$ 称为母小波 由于小波的特殊性质, 使小波变换在时域和频域具有良好的局部性, 克服了 Fourier 分析的诸多弱点 小波神经网是一种基于BP 网和小波分析的新型前馈网络, 其神经元函数是小波函数 小波神经网利用小波函数良好的时频局部化性质和神经网络的学习功能, 将小波空间作为模式识别的特征空间, 将小波基与信号向量的内积进行加权, 从而实现信号的特征提取 小波神经网具有较强的容错能力[5].

## 3　算法和实现

本文所采用的网络模型如图 1 所示, 它是一个包括输入层、隐层和输出层的三层BP 网络 网络的神经元函数是 $h(t)=\cos(1.75t)\exp\left(-\frac{t^2}{2}\right)$, 称为Morlet 母小波 网络输入参数 $x_n(i)$ 的数据类型如表 1 所示 我们选取的呼吸群、词和节奏等语言特征已经可以基本反映独立音节所处的语言环境, 为后续的语音合成工作打下良好的基础 当然, 韵律特征的获取是由文本分析完成的 本网络的输出参数是音节起长和起点基频, 它们都是表征语音韵律特征的最重要参数

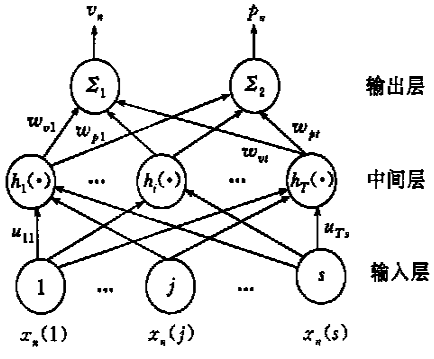


图 1　小波神经网络结构

表 1　输入参数数据类型

| 编号 | 参数含义 |
|---|---|
| 1 | 本音节声母类别 |
| 2 | 本音节韵母类别 |
| 3 | 本音节调型类别 |
| 4 | 后音节声母类别 |
| 5 | 后音节韵母类别 |
| 6 | 后音节调型类别 |
| 7 | 前音节声母类别 |
| 8 | 前音节韵母类别 |
| 9 | 前音节调型类别 |
| 10 | 呼吸群在句中位置 |
| 11 | 词在呼吸群中位置 |
| 12 | 呼吸群内词数 |
| 13 | 节奏在词中位置 |
| 14 | 语音在词内音节数 |
| 15 | 语音在词内节奏数 |
| 16 | 音节在节奏中位置 |
| 17 | 节奏长度 |

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

具体算法如下:

1) 将伸缩因子 $a_t$、平移因子 $b_t$、网络连接权重 $u_{ti}, w_{vt}, w_{pt}$ 赋予随机初始值;

2) 输入学习样本 $x_n(i)$ 及相应的期望值输出 $V_n^T, P_n^T$;

3) 利用当前网络参数计算出网络的输出

$$v_n = \sum_{t=1}^{T} w_{vt} h\left(\frac{\sum_{i=1}^{S} u_{ti} x_n(i) - b_t}{a_t}\right), \quad p_n = \sum_{t=1}^{T} w_{pt} h\left(\frac{\sum_{i=1}^{S} u_{ti} x_n(i) - b_t}{a_t}\right);$$

4) 计算瞬时梯度向量

$$\delta_{w_{vt}} = \frac{\partial e}{\partial w_{vt}} = -\sum_{n=1}^{N} (V_n^T - V_n) h\left(\frac{\sum_{i=1}^{S} u_{ti} x_n(i) - b_t}{a_t}\right),$$

$$\delta_{a_t} = \frac{\partial e}{\partial a_t} = -\sum_{n=1}^{N} (V_n^T - V_n) w_{vt} \frac{\partial h}{\partial a_t} - \sum_{n=1}^{N} (P_n^T - P_n) w_{pt} \frac{\partial h}{\partial a_t},$$

$$\delta_{b_t} = \frac{\partial e}{\partial b_t} = -\sum_{n=1}^{N} (V_n^T - V_n) w_{vt} \frac{\partial h}{\partial b_t} - \sum_{n=1}^{N} (P_n^T - P_n) w_{pt} \frac{\partial h}{\partial b_t},$$

$$\delta_{u_{ti}} = \frac{\partial e}{\partial u_{ti}} = -\sum_{n=1}^{N} (V_n^T - V_n) w_{vt} \frac{\partial h}{\partial x_n} x_n(i) - \sum_{n=1}^{N} (P_n^T - P_n) w_{pt} \frac{\partial h}{\partial x_n} x_n(i),$$

其中 $x_n' = \sum_{i=1}^{S} u_{ti} x_n(i)$, $t_n' = \frac{x_n' - b_t}{a_t}$, 误差函数为 $e = \frac{1}{2}\left[\sum_{n=1}^{N} (V_n^T - V_n)^2 + \sum_{n=1}^{N} (P_n^T - P_n)^2\right]$;

5) 误差的反向传播

$$\Delta w_{vt}^{new} = -\eta \frac{\partial e}{\partial w_{vt}^{old}} + a\Delta w_{vt}^{old}, \quad \Delta w_{pt}^{new} = -\eta \frac{\partial e}{\partial w_{pt}^{old}} + a\Delta w_{pt}^{old},$$

$$\Delta u_{ti}^{new} = -\eta \frac{\partial e}{\partial u_{ti}^{old}} + a\Delta u_{ti}^{old}, \quad \Delta a_t^{new} = -\eta \frac{\partial e}{\partial a_t^{old}} + a\Delta a_t^{old},$$

$$\Delta b_t^{new} = -\eta \frac{\partial e}{\partial b_t^{old}} + a\Delta b_t^{old},$$

修改网络的参数 $a_t$, $b_t$, $w_{vt}$, $w_{pt}$, $u_{ti}$;

6) 当误差函数值小于事先设定的某个值时, 停止学习, 否则返回步骤 2).

## 4 实验结果

由实验结果可以看出, 相关参数的小波神经网络模拟值和实测值吻合得很好, 如图 2 所

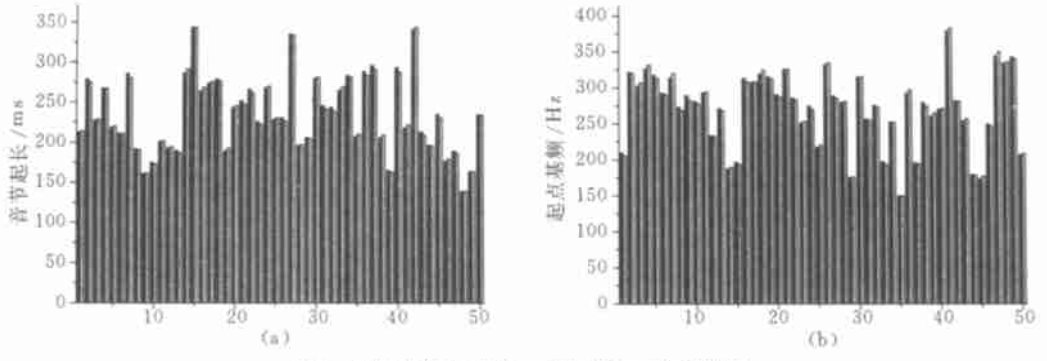


图 2 实验结果示意图(实测值;模拟值)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

示. 按照 3.8% 这种误差水平, 语音合成的质量可以很接近自然语言. 当然, 鉴于小波神经网络目前尚处于实验阶段, 我们只选择了反映韵律特征的几个最重要参数进行模拟, 因而没有给出语音合成的最后结果和实际测量的比较. 但是, 从良好的实验结果可以预见, 小波神经网络在汉语文语转换、语音分析等领域有着很好的应用前景.

**致谢**　感谢中国科学技术大学人机语音通讯实验室, 尤其是王仁华教授及马钟柯同学给予的热情帮助.

## 参 考 文 献

1 Ostendorf M, Veilleux N. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computat. Linguist.*, 1994, **20**: 27～54

2 Mixdorff H, Fujisaki H. A scheme for a model-based synthesis by rule of F0 contours of German utterances. In: Proc. EUROSPEECH, 1995. 1823～1826

3 Hwang S H, Chen S H. Neural network synthesizer of pause duration for Mandarin text-to-speech. *Electron. Lett.*, 1992, **28**: 720～721

4 Chen S H, Hwang S H, Wang Y R. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *IEEE Trans. Speech Audio Processing*, 1998, **6**(3): 226～239

5 Zhang Q, Benvenisite A. Wavelet network. *IEEE Trans. Neural Network*, 1992, **3**(6): 889～898

**吴　震**　1995 年考入中国科技大学少年班, 主要从事语义分析、神经网络及逻辑学方面的研究工作.

**万　千**　1995 年考入中国科技大学少年班, 主要从事神经网络、自动控制等领域的研究.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net